ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

PSYCHOLOGY

Ciclo 38

Settore Concorsuale: 11/E1 - PSICOLOGIA GENERALE, PSICOBIOLOGIA E PSICOMETRIA

Settore Scientifico Disciplinare: M-PSI/03 – PSICOMETRIA

PSYCASSIST: DIGITAL INNOVATION IN ADAPTIVE AND PERSONALIZED NEUROPSYCHOLOGICAL ASSESSMENT OF EXECUTIVE FUNCTIONS AND FLUID INTELLIGENCE

Presentata da: Matilde Spinoso

Coordinatore Dottorato

Mariagrazia Benassi

Supervisore

Sara Giovagnoli

Co-supervisore

Mariagrazia Benassi

Esame finale anno 2026

Acknowledgement

Writing these acknowledgments means looking back and retracing a journey that has been at once incredibly long and fleeting. Three years that at first had seemed like an eternity and somehow flew by in the blink of an eye. Years full of explorations and challenges, sleepless nights, and mornings brimming with excitement over new ideas. A journey that has profoundly enriched me, teaching me not only the rigor of scientific research but also patience, resilience, and the value of collaboration.

Today, as this journey comes to an end, faster than I could have ever imagined, I feel the need to pause and thank all the people who made it possible, who shared it with their presence, and who supported me when the path seemed too steep.

Firstly, this journey was possible thanks to two exceptional mentors: Prof. Sara Giovagnoli and Prof. Mariagrazia Benassi, my supervisors. You were not only scientific guides but true teachers of academic life. You taught me to look beyond the numbers, to seek the meaning behind every data point, and to persevere in the face of methodological challenges that at times seemed insurmountable. Your ability to combine a passion for research with scientific rigor and profound human sensitivity has been a constant source of inspiration. Thank you for believing in me even when I struggled to believe in myself, for your steady support, and for encouraging me to take risks and embrace challenges despite my doubts. You have helped me grow not only as a researcher but also as a person capable of facing obstacles with determination, curiosity, and humility.

Special thanks also to Prof. Giulia Balboni, who, although recently joining our department and research group, has been a dedicated presence and support from the very beginning of this project. I am also grateful to the professors from the University of Padua, Luca Stefanutti, Debora De Chiusole, and Pasquale Anselmi, as well as to my colleagues from

the University of Perugia, Irene Pierluigi and Alice Bacherini. Collaborating with all of you, across cities and institutions, has been among the most enriching experiences of this journey. Your contributions were essential not only for the scientific advancement of this project but also for demonstrating the true value of collaboration in research. Thank you for your guidance, expertise, and the team spirit you consistently exemplified, which made it possible for us to achieve our goals together.

I also wish to express my deep gratitude to all my colleagues at the Psychometric and Neuropsychology Laboratory of the Department of Psychology in Cesena. Within this laboratory, under your guidance, I found much more than a workplace: a lively, ironic, and welcoming scientific community where discussion was always constructive, ideas could circulate freely, and every challenge became an opportunity for shared learning. Thank you for welcoming and supporting me every day, for making collaboration and mutual growth not just a value but a daily practice. Thank you for making the lab not just a research space but a refuge where work, friendship, and organized chaos coexist in perfect balance. For being home, team, and family often all at once on the same day. And, of course, thank you for watering my plants while I was away, a simple gesture that perfectly embodies your kindness and care, and one that, I'm sure, saved more than one plant (and perhaps a little bit of me, too).

I warmly thank all the thesis students and interns who collaborated with me over these years. Your enthusiasm, curiosity, and commitment made every phase of the research more stimulating and rewarding. At the end of the dissertations and internships, you were always the ones thanking me, but perhaps you never realized how much you enriched me in return. Today, I want to thank you for reminding me of the importance of growing up with mentors by your side, of engaging in discussions without fear of feeling more or less competent, but with a genuine desire to learn and grow together. I hope this experience has enriched you as much as it has enriched me.

A profound thanks goes to all the participants in the project: the schools, the children, adolescents, and their families who took part in these studies. You opened your doors and your time to this research. Without your trust and willingness, this work would have remained just an idea. Each of you has contributed to advancing our understanding, and for that, I am infinitely grateful. Thank you for believing in the importance of science and for being an essential part of this journey.

Heartfelt thanks to all my friends, who in different ways have walked alongside me throughout these years. Some of you participated directly in this journey, helping me with data collection—thank you for enduring Sunday afternoons between Raven and the Tower of London tasks, for your endless patience, and for putting up with my outbursts when I had to rush to catch the train. Others supported me with a word, a laugh, or a quiet but constant presence. You were my anchor in moments of fatigue and confusion, my lightness when I took everything too seriously, and my energy when my strength seemed to falter. You reminded me to take things “step by step”, made me laugh when I needed to lighten the tension, and dragged me to endless dinners and weekend escapes that restored my breath. In moments of discouragement, you gave me back my energy and motivation; in moments of success, you celebrated as if it were your own achievement. So, if you are reading these lines, it is because you were present, and still are, during this significant chapter of my life. You are part of this milestone, each in your own way, and I am infinitely grateful for your presence, whether silent or loud, constant and sincere. You have made this journey lighter, simpler, and more human.

To Niccolò, my life partner, my safe harbor: there are no words enough to thank you. You shared every phase of this journey with me, or perhaps I should say, I dragged you through every phase. You listened to my endless monologues on statistical models and long speeches for conference presentations, gracefully pretending to understand everything. You corrected my English pronunciation and endured my simultaneous frustration. You tolerated my mood

swings between “I can do this” and “I want to throw everything out the window.” You believed in me with unwavering faith, even when I doubted myself. You reminded me to breathe, take breaks, laugh, and sometimes sleep by turning off my alarm. You celebrated every small achievement as if it were yours (perhaps also because it meant an evening without Excel). You supported, encouraged, and comforted me. This PhD is also yours because you lived it alongside me, with unconditional love, humor, and infinite patience. For all of this and so much more, I will always be grateful.

To my brother Carlo Alberto: thank you for being a lifelong accomplice, someone who has always known and understood me without the need for many words. Thank you for reminding me, when needed, not to take myself too seriously, for keeping me grounded with your humor, and for always being my support. We share a deep bond, filled with experiences, secrets, dreams, and even quarrels and pillow fights that have made us stronger and closer. You truly know me, have seen me at my best and my worst, and have never stopped believing in me. The bond we share is one of the most beautiful certainties in my life.

And finally, the greatest thanks, the one that encompasses all the others: to my parents, and again to my brother. If I am here today, at the end of this journey, it is only thanks to you. You gave me roots to stay grounded and wings to fly. You taught me the value of study, dedication, and perseverance. You made silent sacrifices so that I could pursue my goals; you supported me in every choice, and you believed in me with unwavering faith even when the path was uncertain. Every page of this thesis carries the imprint of your love, your example, and your teachings. This achievement is not just mine: it is ours. Without you, none of this would have been possible. Thank you, from the bottom of my heart, for everything you are and everything you have given me.

To all of you, thank you

Abstract

In recent years, neuropsychological assessment has undergone a profound methodological and technological transformation, shifting from traditional paper-and-pencil tests to a digital, adaptive, and personalized approach. This doctoral dissertation is situated within this transition, contributing to the development of innovative tools for computerized and adaptive evaluation of executive functions and fluid intelligence. Classical instruments such as the Tower of London and Raven's Progressive Matrices, while offering methodological robustness and clinical validity, present significant limitations: lengthy and inflexible administration, data collection restricted to single time points, and uniform item presentation regardless of individual ability levels. These constraints are particularly problematic for clinical populations, where cognitive vulnerabilities, attentional difficulties, and fatigue render testing challenging and less reliable.

This work is part of the broader research project PRIN 2020 (Prot. 20209WKCLL) entitled "Computerized, Adaptive and Personalized Assessment of Executive Functions and Fluid Intelligence." The main objectives were: (1) to build a new generation of tests for adaptive neuropsychological assessment of planning abilities and fluid intelligence; (2) to develop an intelligent web-based system capable of administering tests, analyzing data, and providing personalized reports; (3) to apply innovative methodological approaches, specifically Knowledge Space Theory (KST) and Procedural Knowledge Space Theory (PKST), which enable efficient evaluation and effective personalization through dynamic item selection adapted to individual performance. The project resulted in the development of PsycAssist, an artificial intelligence-based web platform for comprehensive adaptive neuropsychological assessment with automated scoring and clinical report generation. To date, two primary instruments have been implemented: Adap-ToL, an adaptive version of the Tower of London task for measuring planning abilities, and MatriKS, a computerized adaptive test of the Raven's Matrices task for assessing fluid intelligence.

The dissertation is organized into two parts. Part I outlines the theoretical and methodological foundations across four main sections. The first section traces the evolution of neuropsychological assessment, from traditional paper-and-pencil tests to contemporary digital and adaptive approaches, highlighting their limitations and potential, and addressing crucial aspects such as usability, acceptability, and the ethical challenges associated with the technological innovation context. The second section focuses on executive functions as higher-order cognitive abilities, with particular emphasis on planning and the Tower of London task. The third section explores fluid intelligence, reviewed through major theoretical frameworks and traditional assessment instruments, with specific reference to Raven's Matrices. Finally, the fourth section provides an integrated overview of executive functions and fluid intelligence in clinical populations, with a particular focus on Specific Learning Disorders (SLD).

Part II presents five empirical studies: (1) PsycAssist's technological architecture, adaptive algorithm integration, and results of a pilot study on usability and acceptability; (2) development and validation of the Usability and Attitude Scale (UAS) for populations of children and adolescents; (3) multi-method validation of MatriKS for developmental age through Classical Test Theory, Rasch modeling, and KST; (4) process-oriented modeling of planning performance through Markov models (5) error pattern analysis in MatriKS examining quantitative and qualitative differences between typically development and Specific Learning Disorders children and adolescents.

Findings demonstrate that computerized adaptive assessment substantially enhances precision, efficiency, and ecological validity compared to traditional testing, reducing administration time and cognitive load while enabling process-based analysis. The UAS represents a novel contribution for evaluating the usability and acceptability of digital tools in children and adolescents. Overall, this work supports a reconceptualization of neuropsychological assessment as a dynamic, adaptive, and user-centered process, integrating psychometric innovation and digital technology to enhance both clinical practice and scientific research.

Table of contents

General Introduction

In recent decades, neuropsychological assessment has undergone a profound transformation, shifting from traditional paper-and-pencil methods to digital, adaptive, and personalized approaches. Historically, neuropsychological evaluation has relied on standardized tests administered in person, with instruments such as the Tower of London (Shallice, 1982) and Raven's Progressive Matrices (Raven, 1938), which represent foundational tools for analyzing executive functions and fluid intelligence. Although methodologically robust and clinically valid, these instruments have significant limitations: administration is often lengthy and inflexible, data collection is restricted to single time points, and all participants must respond to the same set of items, regardless of their ability level or personal characteristics. These constraints are particularly problematic when assessing clinical populations, where cognitive vulnerabilities, attentional difficulties, and fatigue render the testing experience challenging, sometimes frustrating, and consequently less reliable (Bauer et al., 2018; Parsons, 2016).

The advent of digital technologies has profoundly reshaped neuropsychological practice through the integration of computerized tests that optimize cognitive assessment. These tools offer numerous advantages: standardized administration, accurate recording of response times, and automated scoring, enabling not only greater measurement precision but also the collection of more detailed and continuous data, useful for faster, more precise assessments and targeted interventions (Wild et al., 2008; Parsey & Schmitter-Edgecombe, 2013).

In parallel with digital tests, advanced and innovative psychometric models have emerged, such as Knowledge Space Theory (KST), a mathematical framework that enables adaptive assessment. These approaches dynamically tailor item selection to individual performance, personalizing the assessment path, reducing the number of questions and effort required, and simultaneously improving data reliability and feedback precision (Doignon & Falmagne, 2011; Embretson & Reise, 2013). A further

innovation is represented by Procedural Knowledge Space Theory (PKST; Brancaccio et al., 2022), which analyzes the entire solution process rather than limiting itself to correct/incorrect outcomes, ensuring greater accuracy and efficiency through the exploitation of richer information (Stefanutti, 2019).

Beyond psychometric innovation, the effective adoption of digital tools critically depends on their usability and acceptability, that is, the perception of ease of use, appropriateness, and satisfaction by end users (Davis, 1989; ISO, 1998). These aspects are particularly relevant in pediatric populations, such as children and adolescents, who require tools designed with reduced cognitive and linguistic complexity, visual support, and engaging interfaces (Markopoulos et al., 2008; Hourcade, 2007). However, specific tools for evaluating usability and acceptability in pediatric populations remain scarce, representing a significant gap in the literature (Lewis, 2018; Read & MacFarlane, 2006), which highlights the urgency of developing valid instruments to assess these crucial aspects already during the design phase of digital tools.

This doctoral dissertation is situated within this landscape of methodological and technological renewal and constitutes a contribution within a broader project funded by the Ministry of University and Research (PRIN 2020, Prot. 20209WKCLL), entitled: "Computerized, Adaptive and Personalized Assessment of Executive Functions and Fluid Intelligence."

The main objectives of the project were:

1) To build a new generation of tests for adaptive and efficient neuropsychological assessment of cognitive functions, such as planning and intelligence.
2) To develop the prototype of an intelligent web-based system capable of administering tests, collecting and analyzing data, and providing personalized reports with results of multidimensional assessments and recommendations for rehabilitation.
3) To apply innovative methodological framework for constructing assessment instruments that

enable efficient evaluation and effective personalization of interventions, represented in this case by Knowledge Space Theory (KST; Doignon & Falmagne, 2011) and its recent generalization, Procedural Knowledge Space Theory (PKST; Stefanutti, 2019).

The project materialized in the development of PsycAssist, an artificial intelligence-based web platform designed as a comprehensive environment for adaptive neuropsychological assessment, with automated scoring and clinical report generation capabilities. Within PsycAssist, two primary instruments were developed: 1) Adap-ToL, an adaptive version of the Tower of London task for measuring planning abilities. 2) MatriKS, a computerized adaptive test of fluid intelligence based on automatic, rule-based generation of Raven-like matrices.

The present dissertation is organized into two main macro-sections: a theoretical part (Chapter I) and an empirical part (Chapter II).

Part I, devoted to theoretical background, provides a comprehensive overview establishing the conceptual and methodological foundations of the research and it is divided into four sections: Section 1. traces the evolution of neuropsychological assessment from traditional paper-and-pencil instruments to contemporary digital and adaptive approaches, with particular emphasis on Knowledge Space Theory (KST) and Procedural Knowledge Space Theory (PKST) as innovative frameworks, while addressing usability, acceptability, and ethical challenges of digital assessment tools. Section 2. examines executive functions as higher-order cognitive abilities, exploring their theoretical models, neuroanatomical foundations, and developmental trajectories, with a specific focus on planning and the Tower of London task, establishing the rationale for Adap-ToL. Section 3. analyzes fluid intelligence through major theoretical frameworks, discussing developmental trajectories and traditional assessment instruments, such as Raven's Progressive Matrices providing the conceptual foundation for MatriKS. Section 4. presents a comprehensive analysis of executive functions and fluid intelligence in Specific Learning Disorders, reviewing empirical evidence on planning and

reasoning abilities in this population and highlighting the need for sensitive assessment tools.

Part II focused on empirical studies presents five original contributions related to the development, validation, and application of the PsycAssist platform: Study 1) Introduction to PsycAssist, description of technological architecture, integration of adaptive algorithms, development of Adap-ToL and MatriKS, and results of a pilot usability study; Study 2) Development and psychometric validation of the Usability and Attitude Scale (UAS), an instrument adapted to evaluate digital tools in pediatric and adolescent populations; Study 3) Multi-method validation of MatriKS through Classical Test Theory, Rasch modeling, and KST; Study 4) Markov solution process models for assessing planning in the Tower of London, highlighting how analyzing the entire solution process provides richer information compared to evaluating only final outcomes; 5) Analysis of error patterns in MatriKS in children and adolescents with SLD, investigating quantitative and qualitative differences compared to typically developing peers, highlighting strengths and weaknesses in logical-nonverbal reasoning and problem-solving strategies.

This dissertation offers a concrete contribution to the evolution of contemporary neuropsychological assessment, promoting a vision of cognitive evaluation as a dynamic, adaptive, and person-centered process. The integration of advanced psychometric models (KST, PKST) and digital tools does not replace clinical expertise but enhances it, providing new instruments to better understand cognitive functioning. The results show that computerized adaptive assessment overcomes many limitations of traditional instruments, offering shorter administration times, reduced fatigue, greater precision, and process-oriented information. The development of the UAS fills a critical gap in evaluating digital technologies in pediatric populations, while the multi-method validation of MatriKS and Adap-ToL confirms their psychometric soundness and clinical utility.

This study seeks to contribute to the development of a new generation of assessment instruments that are scientifically rigorous, clinically informative, and accessible to all populations, including those

with special needs. By integrating established theoretical models with emerging technologies while maintaining a focus on user experience, this work demonstrates how neuropsychological assessment can evolve to meet the challenges of contemporary clinical practice and research, providing innovative, precise, and accessible tools without compromising empirical, methodological, or ethical rigor.

Chapter 1: Theoretical Background

1. The evolution of neuropsychological assessment

1.1. From paper-and-pencil tests to digital transition assessment

Every thought, decision, and action we perform is orchestrated by the brain; however, understanding how neural processes manifest in observable behavior requires more than intuition—it requires systematic and rigorous assessment (Lee & Suhr, 2023). In this context, neuropsychological assessment serves as a bridge between brain and behavior, providing a rigorous, empirical approach to evaluate cognitive, emotional, and behavioral functioning across the lifespan (Lezak, 2004; Lezak et al., 2012; Carone, 2024). It embodies a multidimensional methodology that combines standardized instruments, systematic clinical observations, and comprehensive interviews, thereby enabling the detection of specific cognitive deficits, supporting differential diagnosis, and informing targeted rehabilitative strategies (Hebben & Milberg, 2009).

Neuropsychological evaluations can address a wide range of clinical questions. As outlined by Lezak and colleagues (2012), the most common referrals typically fall into one or more of the following six categories:

1) Diagnostic decision-making, including the identification of signs and symptoms of brain dysfunction and differentiation among various neurological, psychiatric, or medical conditions. In this respect, neuropsychological diagnostic conclusions can guide or complement neurological or neuroimaging investigations to clarify the presence or absence of pathology.
2) Characterization of strengths and weaknesses in cognitive, behavioral, and psychological functioning.
3) Guidance for treatment and rehabilitation planning, assisting patients in managing areas of impairment.

4) Prediction of treatment effects across clinical populations, including monitoring changes in functioning or recovery rates over time. In addition, follow-up assessments are useful for evaluating the effectiveness of interventions, rehabilitation programs, or medical procedures, as well as for tracking decline and its functional implications in neurodegenerative conditions.
5) Research support, providing tools to better understand brain-behavior relationships.
6) Forensic applications, such as in legal proceedings to establish whether an event caused known or suspected brain injury and to assess its consequences on functioning, including decision-making and judgment capacity.

Historically, neuropsychological practice has relied heavily on paper-and-pencil instruments, which have long provided the empirical foundation of the discipline. Classic tests such as the Tower of London (Shallice, 1982), the Trail Making Test (Tombaugh, 2004), the Stroop Color and Word Test (Golden & Freshwater, 2002), the Rey Complex Figure Test (Meyers & Meyers, 1995), and the Raven Progressive Matrices (Raven, 1938; Raven et al., 1989) have offered detailed insights into executive function, attention, memory, problem-solving, and general cognitive ability. These instruments are supported by decades of normative research, allowing clinicians to compare individual performance with populations stratified by age, gender, and education (Spinnler & Tognoni, 1987).

The advantages of traditional assessments are manifold. Direct administration allows for observation of non-verbal behaviors, compensatory strategies, and emotional responses, often providing critical diagnostic information not captured by numerical scores alone (Hebben & Milberg, 2009). Moreover, these tests are widely accessible, cost-effective, and offer a high degree of flexibility in both clinical and research settings. Nevertheless, significant limitations remain. The low frequency of data collection represents one of the major limitations of traditional neuropsychological tests, as they typically capture only a single snapshot of a patient's cognitive functioning (Harris, 2024). Even when follow-up assessments are conducted, they often occur months or years apart, limiting the ability to

detect short-term changes, such as daily or weekly fluctuations, as well as subtle cognitive variations that are particularly relevant in the early stages of neurodegenerative diseases, which may therefore go unnoticed (Bauer et al., 2018; Matar et al., 2020). Moreover, neuropsychological testing is generally conducted in highly controlled, artificial settings (i.e., one-on-one in a quiet room), which constrains ecological validity, a limitations frequently noted by clinical neuropsychologists (Rabin et al., 2016). Traditional tests are also time-consuming, often requiring long hours of administration, which can lead to fatigue, decreased motivation, and variability in performance (Parsons, 2016). The long time required for the evaluations also limits the number of patients who can be seen in a timely manner and results in long wait lists (Loumidis & Shropshire, 1997). Although neuropsychological instruments follow standardized administration protocols, minor inconsistencies among different human examiners are unavoidable (Overton et al., 2016). Another significant limitation arises from linguistic and cultural biases, as only a small number of non-English instruments have been rigorously validated with representative normative data (Bender et al., 2010). Even tests that are primarily nonverbal and impose minimal linguistic demands can still be affected by bias, since many of these measures were originally developed on highly homogeneous samples (Rosselli & Ardila, 2003). Although direct observation in paper-and-pencil tests provides an advantage in capturing qualitative behaviors, it does not allow for automated or continuous recording of responses. To address these limitations, the field has gradually transitioned to digital assessments over the past two decades, thereby improving their quality. The adoption of computer- and tablet-based assessments has reshaped the administration, scoring, and interpretation of neuropsychological tests (Parsey & Schmitter-Edgecombe, 2013). Digital platforms allow standardized stimulus presentation, millisecond-level precision in response recording, automated scoring, and immediate report generation, enhancing both efficiency and reliability (Wild et al., 2008). Moreover, since data are already stored digitally, they can be easily aggregated into large normative databases (e.g., the

National Neuropsychology Network; Loring et al., 2022), enabling more comprehensive analyses and cross-sample comparisons. In parallel, several studies showed that touchscreen interfaces and other digital tools provide a more direct and intuitive interaction with tests, supporting patients with motor or cognitive impairments and improving ecological validity (Shneiderman, 2000; Deguchi et al., 2013). Another important development is the remote administration via telehealth platforms, which has expanded accessibility to patients with mobility limitations or those living in remote areas, a process further accelerated by the COVID-19 pandemic (Kruse et al., 2018; Bilder et al., 2020). Overall, digital assessments not only address many limitations of traditional tests but also create new opportunities for continuous, precise, and contextually adaptive data collection, better reflecting patients' real-world functioning. Moreover, computerized assessments allow continuous data recording, which enables analyses that are not possible with traditional methods (Witt et al., 2013). This feature is particularly useful in neuroimaging contexts and for adaptive testing, where the selection of items can be adjusted in real-time based on individual performance (Mead & Drasgow, 1993). Unlike fixed physical materials, computerized tests allow for the creation of an almost unlimited number of alternative forms (Zygouris & Tsolaki, 2015) by randomly selecting items from a large stimulus database, thereby reducing practice effects. Finally, thanks to automation, test instructions can be translated into multiple languages (Zygouris & Tsolaki, 2015), enabling the assessment of non-English speakers without the need for multilingual examiners. Despite these advancements, critical challenges remain. Digital tools require rigorous validation against traditional methods to ensure diagnostic equivalence, and attention must be given to patient usability, data security, and clinician training. Importantly, despite automatization, the neuropsychologist's role remains crucial during test administration and interpretation (Bauer et al., 2018; Canini et al., 2014). It is also critical to distinguish between fully computerized tests, which automate both stimulus presentation and scoring, and those supported by computers, in which scoring depends on the

examiner and may therefore influence the interpretation of results (Witt et al., 2013). In conclusion, although the shift from traditional to digital neuropsychological assessment offers clear advantages in efficiency, precision, and data richness, these benefits must be balanced with careful consideration of the clinical context, evaluation goals, and individual patient needs. Ultimately, the choice between traditional and computerized tools should be guided by the assessment's purpose, ensuring that technological innovations complement rather than replace existing methods, preserving the accuracy and meaningfulness of neuropsychological evaluation.

1.2. Adaptive and computerized neuropsychological assessment

Beyond facilitating digital administration, the widespread use of computers and related technologies over the past decades has also contributed to the development of adaptive tests, which are characterized by a process in which the items they comprise are selected based on the examinee's responses to previously presented items. In a typical adaptive test, a participant administers an item and, based on their response, is assigned a score; subsequently, the examinee's performance level and precision are estimated, and a new item is selected. The combination of item response theory (IRT), adaptive testing, and interactive administration is known as Computer Adaptive Testing (CAT). A CAT procedure consists of the following components: an item response model, an item pool, a starting level, item selection rules, a scoring method, and a termination criterion (Weiss & Kingsbury, 1984).

Item response theory (IRT) adopts an approach based on the characteristics of individual items, estimating their difficulty and discrimination level and relating them to the examinee's performance level through a specific statistical model. It assumes that the examinee's response to a given item reflects a certain ability level, which corresponds to a proportional probability of correctly answering that specific item (Schmidt & Embretson, 2003). The item pool is one of the most important aspects of CAT and refers to a collection of items that cover the entire range of ability levels of a given trait

in a specific population. Starting levels must be established so that the test can begin with an item appropriate to the examinee's ability. To estimate the participant's starting level, it is sufficient to outline their initial profile and calculate the average ability level of individuals with similar profiles. Regarding the step of determining rules for selecting items after each response, there are various models, such as the maximum likelihood model and Bayesian models, whose purpose is to select items that maximize the information gathered about each examinee's performance level. Scoring models vary depending on the type of response expected for each item: dichotomous IRT for correct/incorrect responses and polytomous IRT for items allowing more than two responses. The termination rule is determined based on the purpose of the assessment: the acquisition of a particular process, the administration of a pre-determined number of items, or reaching the time limit set for that specific test (Navarro & Mourgues-Codern, 2018).

In assessments, choosing a CAT procedure offers several advantages over other approaches. First, it allows shorter administration times through instant recording of performance, which can be used to adapt task difficulty to the individual's performance level, thereby maximizing the use of time, unlike fixed-length tests that force participants to spend time on tasks that are too easy or too difficult for their performance level. This feature is particularly useful for participants at the lower and upper extremes of the distribution. Additionally, items are selected to determine the examinee's maximum performance level. CAT advantages are also observed in scoring analysis through automated calculation procedures, reduction of examiner bias, and the creation of a database containing extensive information, with a wide range of levels, to produce parallel or increasingly difficult test versions (Navarro & Mourgues-Codern, 2018). CATs also allow for long-term cost reduction, immediate feedback on individual performance, particularly useful in clinical settings, and easy import of data into statistical software, useful for research purposes. Computerized tests are also ideal for repeated testing, for example, to monitor changes and intervention effects in clinical populations.

One potential application of CAT is in the educational context, particularly related to reading difficulties. Most traditional tests evaluate only a small portion of the processes involved in this skill, providing information on current performance but not on learning potential or optimal conditions to promote it. This limitation is particularly relevant in assessing students with special educational needs, such as reading difficulties, intellectual disabilities, or socio-cultural disadvantages. Effective learning requires knowledge of the most appropriate teaching strategies for each individual. Dynamic assessments are interactive assessment techniques that include an intervention component aimed at exploring specific cognitive processes and intervention types to ensure effectiveness. They exploit various learning opportunities, including feedback, modified instructions, metacognitive guidance, and graduated prompts, to provide information on students' current and potential performance (Navarro & Mourgues-Codern, 2018).

Dynamic assessments are based on Vygotsky's concept of the zone of proximal development (ZPD) and evaluate both current and potential learning development: the former corresponds to performance achieved without help, while the latter corresponds to performance achieved with guided assistance. To produce a change in performance, effective prompts (i.e., aids) should be within the ZPD. Integrating CAT with elements of dynamic assessment allows combining the strengths of both: CAT records students' responses online and progressively adapt the sequence of items based on performance level, while dynamic assessments contribute through graduated prompts, feedback, pre- and post-test items, and cognitive/metacognitive scaffolding, aiming to evaluate the student's learning potential.

Some learning domains addressed by dynamic assessments include phonological awareness, working memory, reading comprehension in a second language, morphosyntactic awareness, and reading disabilities. Caffrey and colleagues (2008) reviewed several studies and found that dynamic assessments were more efficient than traditional assessments in predicting students' future

performance when considering the necessary aids to achieve optimal performance, suggesting that their use in schools could help connect assessment and teaching. Examples of dynamic assessment applications include computerized algorithms to collect and analyze information during the assessment process. These include the Computer Assisted Modifiability Enhancement Techniques (CAMET, Jensen, 2000), a dynamic and interactive program used in learning and assessment contexts, built with three difficulty levels to gradually adapt to specific educational needs. This tool supports the examiner, without replacing them, in collecting information and analyzing student responses. Another example is the Adaptive Computer Assisted Intelligence Learning Test Battery (ACIL, Guthke et al., 1995), which gathers information to reconstruct the student's performance profile, including time spent on each task, number of completed tasks, number of prompts provided, latency between task presentation and response, and more. ACIL tests use an adaptive approach, meaning tasks are adjusted to the student's performance level. Another example is STAR-EL (Shapiro, 2012), which includes a series of computerized adaptive assessments in reading, mathematics, and communication.

Another computerized battery for assessing reading processes is the EDPL-BAI, implemented on a web platform called "Siette" by the AI research and application team at the University of Málaga, Spain. EDPL-BAI consists of 38 tests grouped into six process blocks: visual coordination and monitoring; grapheme-phoneme association; lexical-morphological processes; morphosyntactic processes; text integration, prior knowledge, and metacognition; and socio-personal adaptation. Each test has an item bank and a system of graduated prompts. Prompt criteria include an initial phase where the student attempts the item without help, a second phase providing a prompt for error types if difficulties occur, and if solved, subsequent item selection depends on how accurately each item reflects the construct being assessed. The intensity, type, and number of prompts required determine the student's specific educational needs. EDPL-BAI records initial and final difficulty levels,

successes by attempts, number and type of errors and graduated prompts needed, and execution time (Navarro & Mourgues-Codern, 2018).

Overall, recent research demonstrates that computerized adaptive tests can reduce administration times by 50–80% while maintaining or even improving diagnostic accuracy relative to traditional assessments, marking a significant methodological advancement in contemporary neuropsychological practice (Kimura, 2017).

1.3. An alternative approach to adaptive assessment: Knowledge space theory (KST) and procedural KST (PKST)

An alternative and innovative approach to adaptive assessment is provided by Knowledge Space Theory (KST; Doignon & Falmagne, 2011), a formal mathematical framework for representing and evaluating an individual's knowledge structure. Unlike classical psychometric methods, which typically attempt to quantify ability through linear scores, KST focuses on describing what an individual knows and what they are ready to learn, thereby identifying the outer fringe of their knowledge state. The set of all knowledge states in a given population forms the knowledge structure, representing the organization of knowledge within the domain and the (quasi-)order relations among problems.

KST-based adaptive assessment allows for personalized testing, dynamically selecting items based on previous responses. By doing so, each administered item provides maximally informative data about the individual's state of knowledge, reducing fatigue and effort while improving accuracy and discrimination between examinees with different levels of competence. This approach is particularly useful in neuropsychological contexts, where attention or cognitive processing deficits may limit performance on standard tests (D'Antuono et al., 2016; Willner et al., 2010).

Recent extensions of KST have broadened its applicability beyond knowledge assessment in educational contexts. Procedural Knowledge Space Theory (PKST; Stefanutti, 2019; Stefanutti, de

Chiusole & Brancaccio, 2021) integrates KST with Problem Space Theory (PST; Newell & Simon, 1972), providing deterministic and probabilistic models for evaluating problem-solving and planning skills. Deterministic models account for both response accuracy and the sequence of actions (observable solution process), while probabilistic models, such as Markov Solution Process Models (MSPM), incorporate latent knowledge states with transition probabilities that consider careless errors and lucky guesses. This allows adaptive assessments to exploit the full observable solution path, rather than reducing responses to simply correct or incorrect.

By integrating the individual's full problem-solving behavior, KST and PKST enhance the precision and efficiency of the assessment, minimizing the number of items needed while maximizing information gain. Deterministic and probabilistic modeling enables fine-grained profiling of specific planning and problem-solving skills at the individual level, offering richer insight than classical scoring methods. Moreover, assessment results obtained through these approaches have greater predictive validity, better forecasting real-life adaptive behavior and informing tailored rehabilitation interventions. Adaptive item selection also limits exposure to redundant problems, reducing learning effects during testing and preserving the reliability of results.

The application of KST and PKST is particularly relevant for executive function (EF) and fluid intelligence (FI) assessment, where traditional scoring methods may fail to capture the multidimensionality of these cognitive processes (Berg et al., 2010; Shallice, 1982). For instance, in the Tower of London (ToL) test, KST-based adaptive procedures can predict an individual's performance across the full domain of 1,260 possible problems by administering only a small subset of representative items, thus increasing both predictive power and reliability (McKinlay, 2008).

KST and PKST have already been successfully implemented in web-based and computerized platforms, including intelligent tutoring and adaptive learning systems, demonstrating the feasibility of scalable, personalized, and precise assessment (Doignon & Falmagne, 2011; Stefanutti, 2019).

Integrating these frameworks into neuropsychological evaluation can overcome limitations of traditional and standard computerized tests, offering an efficient, adaptive, and clinically informative method for assessing EF and FI, while providing a foundation for individualized rehabilitation planning.

1.4. Evaluating Usability and Acceptability of new digital tools

The rapid advancement of digital technologies has profoundly impacted clinical and research settings in neuropsychology, offering innovative tools for assessment, rehabilitation, and therapeutic interventions. Computerized tests, interactive applications, and online platforms offer advantages such as standardized procedures, automated data collection, and the possibility of tailoring tasks to individual characteristics (Falmagne & Doignon, 2010; Heller & Stefanutti, 2024). However, theoretical or clinical validity alone is not sufficient: the effectiveness of these tools also depends on the extent to which they are perceived as easy to use (usability) and appropriate or enjoyable (acceptability) by the end users (Davis, 1989; Venkatesh & Bala, 2008). The relevance of these aspects must always be adapted to the specific target population, considering that adults, children, and clinical groups present different cognitive, linguistic, and motivational profiles. For instance, children require tools characterized by simple language and visual support (Markopoulos et al., 2008; Hourcade, 2007), whereas individuals with neuropsychological deficits may need simplified interfaces and guided instructions to ensure accessibility and valid data collection (Diamond, 2000).

Nevertheless, specific tools designed for developmental populations are still lacking; most existing questionnaires were originally developed for adults and often prove too complex or cognitively demanding for younger users (Lewis, 2018). This makes it necessary to adapt existing instruments or design new ones that meet the needs of children and adolescents, reducing cognitive load while promoting engagement and comprehension (Read & MacFarlane, 2006). As will be illustrated in Study 2 of this dissertation, part of the present work contributes to this perspective by proposing and

validating an innovative scale specifically developed to assess usability and acceptability in younger populations.

Usability refers to the degree to which a system allows users to achieve specified goals with effectiveness, efficiency, and satisfaction in a defined context of use (ISO, 1998). In clinical and neuropsychological contexts, usability assessment is crucial since poorly designed or overly complex tools may hinder engagement, increase errors, and compromise data quality (Nielsen, 1994; Rubin & Chisnell, 2008). Among the most widely used models, the System Usability Scale (SUS; Brooke, 1996) is considered a gold standard and has been validated across several application fields (Bangor, Kortum, & Miller, 2008; Lewis, 2018). Usability metrics generally reflect three dimensions: effectiveness (the ability to complete tasks successfully), efficiency (the resources required to do so), and satisfaction (subjective comfort and acceptance during use; ISO, 1998). Usability can be measured through qualitative methods, such as think-aloud protocols (Nielsen, 1994; van den Haak, de Jong, & Schellens, 2003) or heuristic evaluations (Nielsen & Molich, 1990; Jeffries & Desurvire, 1992), as well as through standardized questionnaires. For adults, SUS and its derivatives (e.g., UMUX-Lite; Lewis, Utesch, & Maher, 2015) are among the most employed measures. For children and adolescents, however, adaptations are required to reduce linguistic and cognitive complexity (Markopoulos et al., 2008; Hourcade, 2008). Simplified questionnaires incorporating visual supports, such as icons or emoticons, enhance comprehension and engagement (Read & MacFarlane, 2006).

On the other hand, acceptability refers to the extent to which a digital tool is perceived as useful, appropriate, and satisfactory by its users (Davis, 1989). This construct is rooted in the Technology Acceptance Model (TAM; Davis, 1989), which identifies two primary determinants of technology adoption: perceived usefulness and perceived ease of use (Davis, Bagozzi, & Warshaw, 1989). Later extensions, such as TAM2 (Venkatesh & Davis, 2000) and TAM3 (Venkatesh & Bala, 2008), further incorporated factors including social influence, computer self-efficacy, perceived control, and

enjoyment, making the framework highly applicable in clinical and educational contexts (Lee, Kozar, & Larsen, 2003; Teo, Lee, & Chai, 2008). In neuropsychological settings, acceptability is critical because it directly influences motivation and engagement (Druin, 2002). For children, acceptability is not only shaped by usefulness and ease of use, but also by factors such as perceived enjoyment, sense of control, and willingness to repeat the experience (Markopoulos et al., 2008). While traditional assessment methods rely on interviews, focus groups, or TAM-based questionnaires, developmental populations require adaptations with simplified language and playful, visual elements (Hourcade, 2007). In this respect, instruments such as the Usability and Attitude Scale, presented in this dissertation, enable the simultaneous assessment of usability and acceptability, providing an integrated understanding of younger users' experiences with digital technologies.

Systematically evaluating usability and acceptability provides several advantages. On one hand, it allows researchers and developers to optimize the design of digital tools, making them more intuitive, accessible, and engaging (Lewis, 2018). On the other hand, it ensures ecological validity and compliance, which are fundamental in clinical and neuropsychological research (Nikolaus et al., 2014). Moreover, such evaluations offer valuable feedback for developing truly user-centered technologies, thereby enhancing the effectiveness of both educational and therapeutic interventions (Bangor et al., 2008; Druin, 2002).

Nonetheless, some methodological considerations must be taken into account: (1) selecting psychometrically validated and culturally adapted instruments (Lewis, 2018); (2) accounting for the cognitive and linguistic abilities of children, with examiner support when needed (Read & MacFarlane, 2006); (3) avoiding negatively worded or overly complex items that increase cognitive load and reduce reliability (Stewart & Frye, 2004; Kortum, Acemyan, & Oswald, 2021); and (4) employing a multi-method approach, combining quantitative tools (standardized questionnaires) with qualitative methods (observations, interviews) to provide a comprehensive evaluation (Kuniavsky,

2003). In conclusion, accurate assessment of usability and acceptability ensures the development of digital technologies that are not only scientifically valid but also truly usable and engaging for their intended users, whether adults, children, or clinical populations with specific cognitive needs.

1.5. Ethical and methodological issues

The technological evolution of neuropsychological assessment carries significant ethical and methodological implications, which require careful consideration (Bush et al., 2005; Bauer et al., 2012; Harris, 2024). From an ethical perspective, informed consent represents a central issue, as patients must fully understand the procedures, data collection methods, storage protocols, and intended uses of their information, especially in digital contexts characterized by complex transmission and processing systems (Luxton et al., 2014). Such procedures should clearly describe technological aspects, potential privacy risks, protective measures, and safeguards against unauthorized access (Nebeker et al., 2019). The increasing adoption of artificial intelligence algorithms introduces additional challenges related to transparency, as automated decision-making processes can sometimes be opaque even to professionals (Gerke et al., 2020).

Another critical issue concerns equitable access to digital tools to prevent the widening of pre-existing disparities in healthcare. Socioeconomic, cultural, geographic, and digital literacy factors can hinder the use of computerized assessments (Luxton, 2016). Rural populations, older adults, and individuals from disadvantaged backgrounds are at greater risk of encountering difficulties (Maheu et al., 2018), as are generational differences in digital competence, which can influence performance and introduce biases that confound cognitive ability with technological familiarity (Tun & Lachman, 2010). These issues are particularly relevant when assessing cultural and linguistical diverse samples, where technological factors may interact with contextual variables and compromise the validity of measurements. Furthermore, given the highly sensitive nature of neuropsychological information, data security and privacy protection require rigorous attention. Digital storage and transmission create

vulnerabilities to unauthorized access, data breaches, and cyberattacks (Luxton et al., 2014). Compliance with data protection regulations, such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States, is therefore mandatory and requires the implementation of robust encryption systems, secure transmission protocols, and clear policies regarding data retention and destruction (Nebeker et al., 2019). From a methodological standpoint, the main challenges concern the validity and reliability of digital assessments. Continuous updating of normative datasets is essential, as norms developed for paper-and-pencil tests are not always applicable to computerized versions, which are exposed to potential mode effects, i.e., systematic differences related to the test administration format (Parsey & Schmitter-Edgecombe, 2013; Germine et al., 2019; Axelrod, 2017; Cernich et al., 2007). In this regard, the validation of digital tools requires equivalence studies with traditional tests, sensitivity and specificity analyses, and reliability checks, especially considering the possibility of repeated administrations without significant practice effects (Parsons, 2016). At the same time, although automation improves procedural efficiency, it cannot replace clinical judgment, which remains crucial in complex cases and in the integration of quantitative data with qualitative and contextual observations (Bilder, 2011; Sweet et al., 2015; Rabin et al., 2016). Professional training must therefore evolve to integrate technological and psychometric competencies while maintaining solid theoretical and clinical foundations (Parsons et al., 2016). Finally, the standardization of digital assessment presents new challenges, as differences between hardware devices, connectivity issues, and environmental variables during administration can introduce variability in results, reducing their comparability and reliability (Germine et al., 2019; Brearly et al., 2017). In conclusion, integrating digital tools into neuropsychological practice requires a balance between technological innovation, methodological rigor, privacy protection, and clinical expertise in order to ensure assessments that are valid, reliable, and accessible to all patients.

Assessment of Executive Functions

1.6. Definition and theoretical models of executive functions

Executive functions (EFs; also called executive control or cognitive control) refer to a family of top-down mental processes needed when you have to concentrate and pay attention, when going on automatic or relying on instinct or intuition would be ill-advised, insufficient, or impossible (Burgess & Simons 2005; Espy 2004; Miller & Cohen 2001). Using EFs is effortful; it is easier to continue doing what you have been doing than to change, it is easier to give in to temptation than to resist it, and it is easier to go on "automatic pilot" than to consider what to do next. There is ongoing debate about whether executive functions (EFs) should be conceptualized as a general domain or as a domain composed of multiple subprocesses (Burgess & Simons, 2005). Theories of the first type argue that damage to a single process or system may explain the different dysexecutive symptoms (Cohen, Braver, & O'Reilly, 1998; Grafman, 2002). Other theories, on the contrary, posit that EFs are based on multiple components that generally work together but can be examined in a relatively independent way at the experimental level (Poletti, 2014). With the advancement of EF research, several models have been proposed. Cristofori et al. (2019) report some of these:

Luria (1973) was the first to propose a model of EF, conceptualizing problem-solving behavior as dependent on a set of overarching skills, or EFs, which rely on the proper functioning of the frontal lobes. He identified the primary components of the executive system as anticipation (establishing realistic expectations and understanding consequences), planning (organization), execution (flexibility and maintenance of set), and self-monitoring (emotional regulation and error detection). Subsequently, Lezak (1995), inspired by Luria's model, defined EFs as the mental capacities necessary to formulate a goal, plan, and implement actions to achieve it. These functions are involved in performing complex tasks and operate when behavioral control processes are required. These functions are engaged when behavioral control is necessary and involve three core cognitive

processes: shifting between tasks or mental sets, inhibiting irrelevant automatic responses, and updating working memory representations (Miyake et al., 2000; Van der Linden et al., 2000). The concept of self-regulation refers to the ability to modulate behavior according to internal goals and constraints (Goldberg & Podell, 2000), which is supported by maintaining mental representations to inhibit inappropriate responses (Levine et al., 1998), thereby enabling ongoing behavioral adjustments while considering potential consequences.

Stuss and Benson (1986) proposed another EF model, related to Lezak's model and inspired by Luria's work. However, Stuss (1991) placed EFs within a hierarchical structure, in which they receive input from lower-level processes (e.g., attention, memory, language, perception) and from higher-level metacognitive processes. Sohlberg and Mateer (1987; 2001) conceptualized a clinical model of EF involving six components: (1) initiation and drive (activation of cognitive processes), (2) response inhibition (suppressing automatic or prepotent responses), (3) task persistence (maintaining behavior until task completion), (4) organization (structuring and sequencing information), (5) generative thinking (producing multiple solutions and demonstrating cognitive flexibility), and (6) awareness (monitoring and adjusting one's own behavior). This model serves as a practical guide for assessment, observation, and intervention planning.

Grafman (2002) introduced the "Structured Event Complex" (SEC), a goal-oriented sequence of events representing thematic knowledge, morality, abstractions, concepts, social rules, event features, event boundaries, and grammar, suggesting that the prefrontal cortex stores unique forms of hierarchical knowledge that, when activated, appear as EFs. Two additional models rely heavily on functional neuroimaging evidence. Badre and D'Esposito (2007) emphasized the PFC's role in supporting flexible and organized action, providing evidence for a rostro caudal hierarchical organization of control within the PFC. They demonstrated systematic posterior-to-anterior gradient corresponding to the level of representation, and fMRI data across four experiments confirmed that

PFC subregion activation aligned with sub- and superordinate relationships in an abstract representational hierarchy. These findings provide empirical constraints on theoretical accounts of PFC control hierarchies. Similarly, Koechlin and Summerfield (2007) proposed a model grounded in information theory, suggesting that action selection is guided by hierarchically organized control signals processed within a network of lateral PFC regions along the anterior–posterior axis. This framework clarifies how executive control can operate as an integrated function while coordinating information across multiple, functionally specialized prefrontal regions.

There is general agreement that there are three core EFs (Diamond, 2013, Lehto et al. 2003, Miyake et al. 2000): inhibition (inhibitory control, including self-control (behavioral inhibition) and interference control (selective attention and cognitive inhibition)), working memory (WM), and cognitive flexibility (also called set shifting, mental flexibility, or mental shifting and closely linked to creativity). Specifically, inhibitory control, which allows the regulation of attention, behavior, thoughts, and/or emotions, is essential for inhibiting automatic or inappropriate behavior; without it, we would be at the mercy of our impulses. Moreover, working memory, which enables holding information in mind to mentally manipulate it, is necessary for performing mental calculations, translating information into action plans, incorporating new information, and organizing sets of items. It is also crucial for our ability to see connections between seemingly unrelated things and to separate elements from an integrated whole; reasoning would not be possible without working memory. Finally, cognitive flexibility, understood as the opposite of rigidity, relies on the first two systems and develops subsequently, is fundamental for perspective shifts, both spatial and interpersonal, and for flexible thinking; it implies the ability to adapt to changing demands or situations, acknowledge mistakes, and take advantage of sudden, unexpected opportunities.

Inhibitory control and working memory support each other in an interdependent relation. To relate different facts and ideas, one must be able to maintain focus on a single item and resist old patterns

of thought. Inhibitory control can help working memory by keeping the mental workspace "organized," suppressing intrusive thoughts, and removing no-longer-relevant information from this limited workspace. To change perspective, we must inhibit (or deactivate) our previous perspective and "load" (or activate) a different perspective in working memory. In this sense, cognitive flexibility requires and relies on inhibitory control and working memory.

As previously mentioned, Diamond (2013) describes three main systems underlying "higher order" executive functions (see Figure 1). From these three systems arise "higher order" executive functions, also known as fluid intelligence (Diamond, 2013). Fluid intelligence is the ability to reason, solve problems in novel situations, and to see patterns or relations among items (Ferrer et al. 2009). It includes both inductive and deductive logical reasoning. It involves being able to figure out the abstract relations underlying analogies. It is synonymous with the reasoning and problem-solving subcomponents of EFs (see Figure 1). It is not surprising then that measures of fluid intelligence (e.g., Raven's Matrices (Raven, 1938; 1989)) are highly correlated with independent measures of EFs.

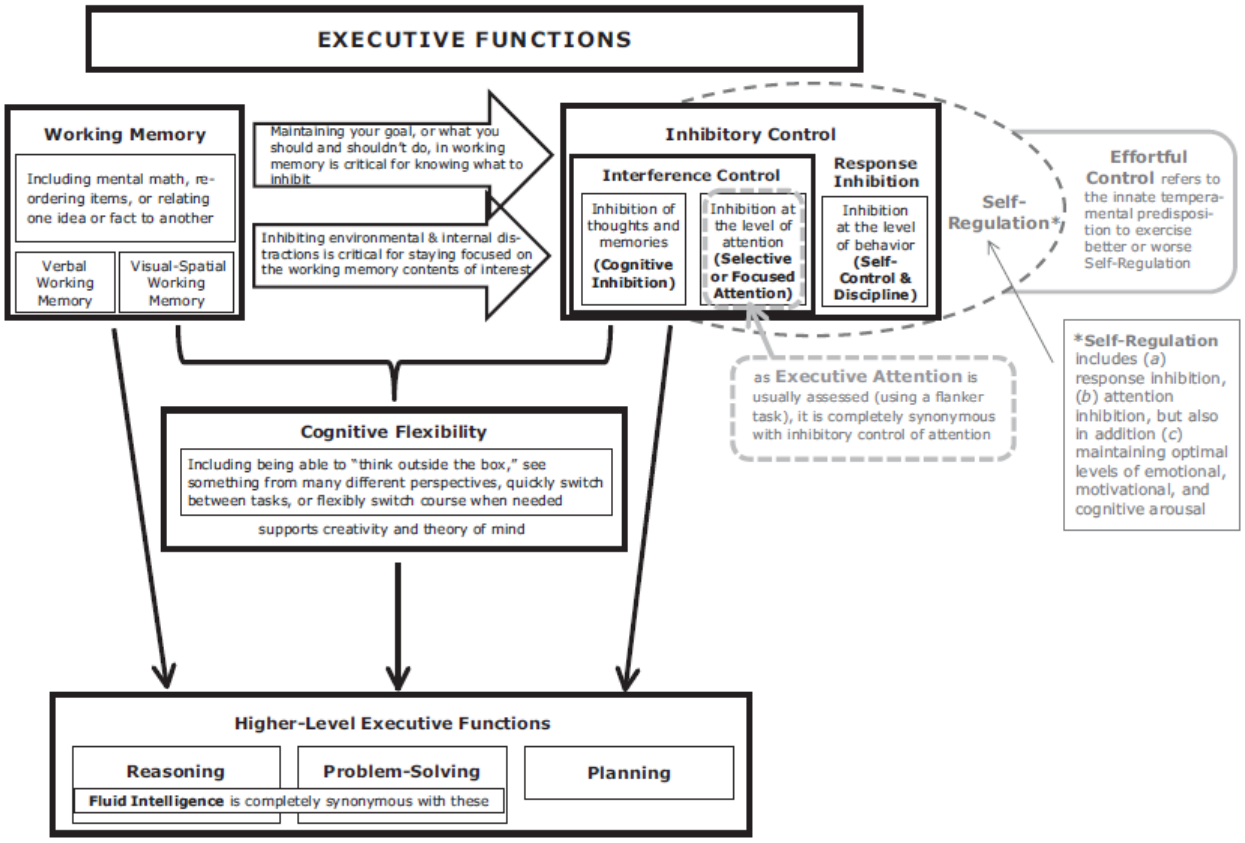


Figure 1. The executive functions model (Diamond, 2013).

From these, higher-order EFs are built, such as reasoning, problem solving, and planning (Collins & Koechlin 2012, Lunt et al., 2012). EFs facilitate the voluntary control and regulation of lower-order cognitive processes and allow individuals to comprehend complex or abstract concepts, solve novel problems, plan future events, and manage interpersonal relationships (Cristofori et al., 2019). This occurs through continuous interaction, optimizing the allocation of cognitive and energetic resources to achieve specific goals (Gilbert & Burgess, 2008). In this sense, EF act as a "controller" that regulates thoughts and behaviors (Miyake & Friedman, 2012).

Moreover, it's known that EFs are skills essential for mental and physical health; success in school and in life; and cognitive, social, and psychological development (Diamond,2013, Hanks et al., 1999; Busch et al., 2005). Consequently, deficits in EF can have severe impacts on daily life, including inappropriate behavior and impaired social functioning which can lead to difficulties in maintaining

an independent and socially productive life.

Executive functions (EFs) can therefore be defined as a multidimensional construct encompassing a wide range of cognitive skills that enable the efficient execution of goal-directed behavior (Lezak, 1995; Shallice, 1990; Stuss, 1992). Specifically, these skills allow us to identify the problem, anticipate the steps needed to solve it, generate and monitor a plan, organize time and space, recall completed steps, and exercise flexibility when changes occur. Although the models described above provide different perspectives on the organization of executive functions, they all agree that EFs involve multiple subcomponents, primarily associated with the prefrontal cortex (Cristofori et al., 2019). Studies using lesion analyses and neuroimaging are particularly valuable for elucidating the neural underpinnings of EFs and for evaluating which theoretical framework best predicts executive performance. In other words, a theoretical understanding of executive functions cannot be separated from knowledge of the brain structures that support them. The prefrontal cortex, together with its connected cortical and subcortical networks, provides the essential biological foundation for the execution of these complex cognitive processes. This link between theory and neuroanatomy sets the stage for the next section, which will explore the structure and function of the neural networks involved in executive functions.

1.7. Neuroanatomy of executive functions

The prefrontal cortex (PFC) is the brain region most closely linked to executive functions (EFs). The PFC constitutes over 30% of the total cortical neurons and represents the most recently evolved area of the brain (Semendeferi et al., 2002). It is also the portion of the cortex that is more extensively developed in humans than in other primates (Hoffmann, 2013). The PFC can be broadly divided into three main regions: the medial region, the dorsolateral region, and the orbitofrontal area (Ongür et al., 2003; see Fig. 2). The PFC receives afferent projections from other neocortical areas, including parietal and temporal cortices. It also obtains input from the hippocampus, cingulate cortex, substantia

nigra, and thalamus, primarily via the medial dorsal nuclei. In turn, the PFC projects back to the medial dorsal nuclei as well as to the amygdala, septal nuclei, basal ganglia, and hypothalamus (McGarry & Carter, 2017), resulting in a dense network of connections with both cortical and subcortical structures.

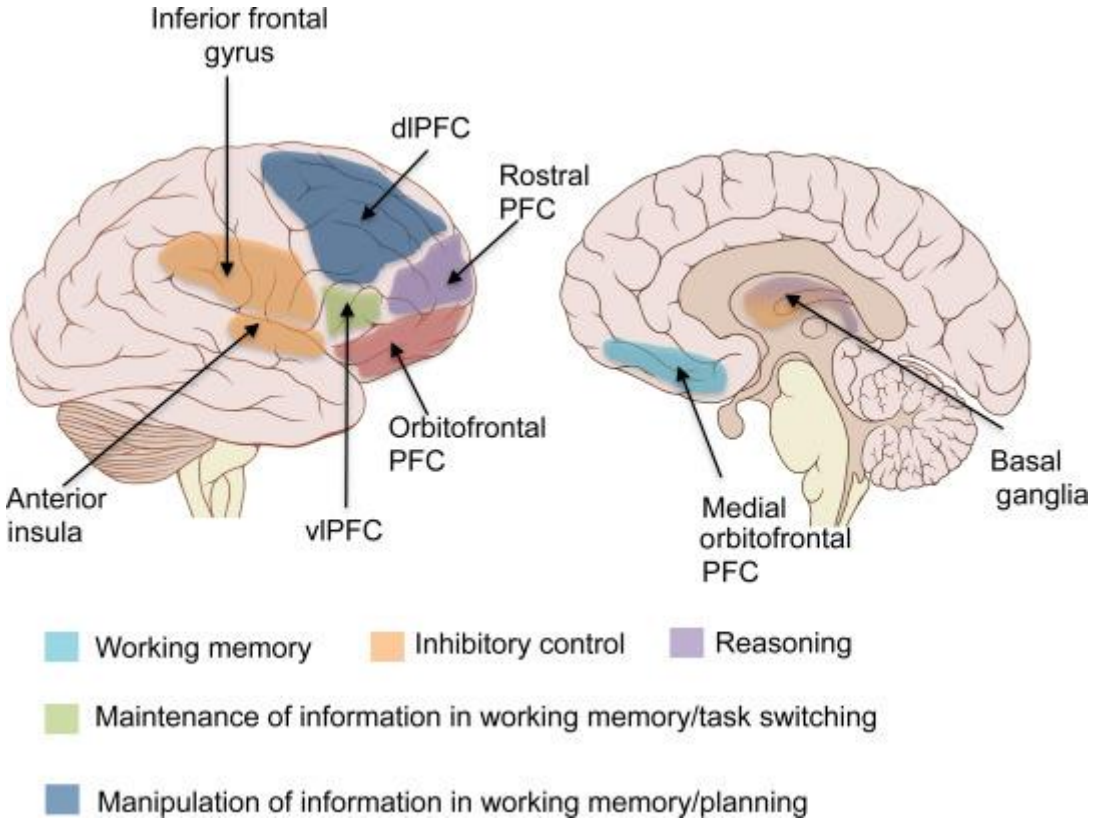


Figure 2. Brain areas associated with different EFs based on lesion and/or neuroimaging studies (Cristofori et al., 2019).

As previously mentioned, the neural substrates of executive functioning were originally assumed to be located in the frontal lobes, since patients with lesions in the anterior part of the brain frequently demonstrated impaired performance on a wide range of tasks assessing executive functioning such as planning abilities, inhibition processes and rules detection (e.g., Burgess & Shallice, 1996a,b; Owen et al., 1990; Shallice, 1982), although a normal executive performance was also reported in some cases. For example, studies using tasks that require similar cognitive processes (such as dual-task coordination or verbal fluency) have demonstrated either preserved (e.g., Ahola et al., 1996; Baddeley

et al., 1997) or impaired performance in frontal-lobe patients (Cowey & Green, 1996; Perret, 1974). Moreover, executive deficits are found more frequently following diffuse (Cowey & Green, 1996; Simkins-Bullock et al., 1994) than focalised frontal lesions (Andrès & Van der Linden, 2002; Andrès & Van der Linden, 1998; Vilkki et al., 1996). On the other hand, there is also evidence suggesting that patients with non-frontal lesions can exhibit executive deficits similar to those of frontal patients (e.g., Andrès & Van der Linden, 2000; Mountain & Snow-Williams, 1993). Taken as a whole, these data seem to indicate that the presence of frontal lesions does not necessarily involve executive dysfunction and that executive processes are not exclusively based upon a network of prefrontal regions.

Studies exploring the neural substrates of executive functioning using cognitive subtraction designs have demonstrated that many cerebral areas are associated with different executive processes. Heterogeneous regions were found to be activated, not only by the four executive processes reviewed here but also by a single process, depending on the exact requirements of the tasks administered. Taken as a whole, these studies clearly demonstrate that the different executive functions depend upon the intervention of both prefrontal and posterior (mainly parietal) regions during the performance of various executive tasks. This is in accordance with the hypothesis that executive functioning is based on a network of anterior and posterior cerebral areas and is not localised only within the frontal lobes (D'Esposito & Grossman, 1996; Fuster, 1993; Morris, 1994; Weinberger, 1993). More specifically, Duncan and Owen (2000) proposed that a specific frontal lobe network including the mid-dorsolateral, mid-ventrolateral and dorsal anterior cingulate cortex is consistently associated with a broad range of tasks requiring, among other processes, response selection, working memory maintenance and stimulus retrieval, while much of the remainder of the frontal cortex, including most of the medial and orbital regions, is largely insensitive to these task demands. In support of this hypothesis, Collette and Van der Linden (2002) showed that some prefrontal areas

(BA 9/46, BA 10 and anterior cingulate gyrus) are systematically activated by a wide range of executive tasks, suggesting that they are involved in general executive processes. However, other frontal areas (BA 6, 8, 44, 45, 47) and parietal regions (BA 7 and BA 40) are also activated during the performance of executive tasks. Since these regions are involved less systematically in the different executive processes explored in that review, it was hypothesised that they have more specific functions. Finally, in a recent meta-analysis, Wager and Smith (2003) showed that different executive processes (continuous updating, memory for temporal order, manipulation of information in working memory and selective attention) are associated with specific cerebral areas. For example, manipulation of information (including dual-task requirements or mental operations of switching and inhibition) most frequently activates the right inferior prefrontal cortex (BA10 and 47). The superior frontal cortex (BA 6, 8 and 9) responds most when working memory must be continuously updated and when memory for temporal order must be maintained. Selective attention to features of a stimulus to be stored in working memory activates the medial prefrontal cortex (BA 32) in storage tasks. The posterior parietal cortex (BA 7) is involved in all three of these executive processes and is also associated with basic control over the focus of attention. These results show that executive functions may be fractionated into different component processes and that these components are associated with specific cerebral areas. In conclusion we can say that EFs are mainly regulated by the PFC, but also other posterior and subcortical regions play an important role, especially in the integration of information and emotion (Cristofori et al., 2019). Further research is still needed to better understand how the subcortical regions are involved in the regulation of EFs.

Evidence from neurodevelopmental disorders further informs our understanding of executive function networks. Conditions such as Attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), and specific learning disorders (SLD) are frequently associated with executive deficits, often linked to atypical development of prefrontal and frontoparietal networks

(Willcutt et al., 2005; Hill, 2004; Crisci et al., 2021). For instance, individuals with ADHD commonly exhibit difficulties in inhibitory control and working memory, functions supported by dorsolateral and ventrolateral prefrontal regions (Castellanos & Proal, 2012). Similarly, ASD has been associated with impairments in cognitive flexibility and planning, potentially reflecting altered connectivity between prefrontal and parietal areas (Geurts et al., 2014). These observations suggest that disruptions in the maturation or integration of executive networks during development may underline the cognitive and behavioral challenges characteristic of these conditions. Thus, neurodevelopmental research not only emphasizes the functional importance of prefrontal and parietal circuits but also reinforces the conceptualization of executive functions as emergent properties of distributed neural systems rather than the output of a single brain region.

A huge amount of studies revealed the presence of inhibitory processes impairments in children with ADHD (Sonuga-Barke et al., 2003; Willcutt et al., 2005; Toll et al., 2011; Shimoni et al., 2012; Crosbie et al., 2013; Rajendran et al., 2013; Schreiber et al., 2014). Martinussen and Tannock (2006) also indicated WM as having an essential role in ADHD deficits. According to Alderson et al. (2010), major deficits can be seen in the central executive system, followed by visuospatial WM, and then verbal WM. Anomalies in visuospatial WM are thought to be among the most important deficits in the neuropsychological profile of children with ADHD (Prins et al., 2011). Finally, a few studies focused on shifting abilities in children with ADHD, with mixed results. Some studies found no shifting deficit (Biederman et al., 2007); some reported impaired shifting functions in terms of both accuracy and response times (O'Brien et al., 2010); and some only identified a lower accuracy (Holmes et al., 2010) or slower response times (Oades and Christiansen, 2008). These conflicting findings are probably due to the tasks chosen, which usually involve other EFs (Irwin et al., 2019).

The structural and functional characteristics of the prefrontal cortex and associated networks evolve over time, underpinning the gradual development of executive functions from infancy to adulthood

1.8. Development of executive functions across the lifespan

The development of cognitive functions progresses in parallel with the neural maturation of the central nervous system, underscoring the close relationship between the growth of executive functions and the structural as well as functional development of the prefrontal cortex. Empirical evidence indicates that, in early childhood, executive functions are not yet clearly differentiated (Brydges et al., 2012; Hughes et al., 2009; Nelson et al., 2017; Wiebe et al., 2011; Willoughby et al., 2012). However, they become increasingly refined and specialized during adolescence as a result of the ongoing neural maturation of the prefrontal cortex (Baggetta & Alessandro, 2016; Best & Miller, 2010; Blair & Ursache, 2011; Miyake, 2000).

The prefrontal cortex (PFC) is recognized for its slow rate of maturation, with certain areas continuing to develop through adolescence and into adulthood (Gilbert & Burgess, 2008). Similarly, executive functions (EFs) progress gradually and are among the last cognitive abilities to reach full maturity. This protracted development of EFs is thought to account for many of the observed differences across the stages of cognitive development throughout life. On average, both children and older adults demonstrate lower EF performance compared to young adults.

Through childhood and adolescence, EFs steadily improve (Jurado & Rosselli, 2007). Infants initially respond mainly to immediate stimuli and events in their surroundings. Nevertheless, even within the first year, infants can display early capacities for extracting basic rules as well as for attention and memory (Rose et al., 2016). These fundamental information-processing abilities are believed to form the basis for the subsequent development of executive functioning. By the end of the first year, children typically acquire the ability to inhibit overlearned behaviors, which enhances attentional control over their environment. Abilities such as planning and mental set-shifting start to emerge around age three, with notable improvements observed after age seven. The capacity to suppress irrelevant information tends to peak slightly later, between ages six and ten.

There are considerable individual differences in the rate and extent of EF development during childhood, influenced by both genetic and environmental factors. For instance, one study found that parental socioeconomic status predicted children's working memory at 4.5 years of age and planning skills by first grade (Hackman et al., 2015). These performance disparities largely persisted through middle and late childhood, suggesting that early environmental conditions play a significant role in shaping EF development. While the development of EFs has been extensively studied in early and late childhood, fewer investigations have focused on changes during adolescence. Evidence indicates that abilities such as inhibitory control, processing speed, working memory, and decision-making continue to mature throughout this period. For example, improvements have been noted in tasks assessing selective attention, working memory, and problem-solving. Research also points to a non-linear pattern of development across adolescence, with some declines in EF performance during late adolescence relative to early or middle adolescence, possibly due to neural reorganization (Blakemore & Choudhury, 2006). A recent longitudinal study tracking individuals aged 17–19 over 12–16 months observed performance declines on a sorting task within this age range (Taylor et al., 2015). Overall, most EFs either improve or stabilize during adolescence, although some may show minor decreases. Just as the maturation of the PFC has been linked to EF development, age-related declines in EFs at the other end of the lifespan are associated with structural changes in the brain during normal aging. The human brain undergoes gradual volume reduction beginning in middle adulthood, with regions such as the hippocampus and prefrontal cortex particularly vulnerable (Raz et al., 2005). Although increasing age generally correlates with decreased EF performance, it is often challenging to separate this decline from general cognitive slowing (Crawford et al., 2000). A well-documented example involves age-related changes in executive attention and inhibition, which manifest as heightened susceptibility to interference and reduced working memory capacity (Hasher & Zacks, 1988). Studies focused on planning ability have shown that younger adults (mean age 22.7) outperform older adults

(68.1 and 78.75 years) on the Tower of Hanoi task (Sorel & Pennequin, 2008). Similarly, meta-analytic findings from the Wisconsin Card Sorting Test (WCST) reveal that older adults complete fewer categories and make more perseverative errors (Rhodes, 2004), demonstrating pronounced age differences in standard EF assessments. Although some theories of cognitive aging emphasize the impact of executive decline throughout adulthood, others propose that executive control plays only a modest role in explaining age-related cognitive changes. Importantly, executive control—defined as resistance to interference and flexibility in task shifting—does not account for additional age-related variation in complex cognition beyond that explained by slower processing speed (Verhaeghen, 2011). Given the dynamic development of executive functions, reliable and valid assessment tools are essential for evaluating these abilities at different stages of life. Neuropsychological tests allow researchers and clinicians to quantify and monitor EF performance over time

1.9. Traditional tools and assessment methods for executive functions

Given the multidimensional nature of executive functions and the variety of theoretical models proposed, the assessment of EFs presents significant challenges. Some approaches are rooted in the unitary perspective, employing global measures of executive control, while others are based on the multicomponent view, targeting specific subprocesses such as inhibition, working memory, or cognitive flexibility. Consequently, the choice of assessment tools must consider not only the theoretical framework adopted but also the ecological validity and sensitivity of the measures employed. This complexity has led to the development of a wide range of standardized neuropsychological tests and experimental paradigms, each designed to capture different aspects of executive functioning.

This section describes the main neuropsychological tools used to assess executive functions (EFs). Among these are card sorting tasks and the Trail Making Tests (Reynolds & Horton, 2014). The most well-known card sorting test is the Wisconsin Card Sorting Test (WCST; Berg, 1948; Johnco et al.,

2014; Tchanturia et al., 2012), a set-shifting task designed to evaluate abstract thinking and the ability to demonstrate flexibility in response to change. The Controlled Oral Word Association Test (COWAT; Benton & Hamsher, 1989), on the other hand, is a verbal fluency test that measures the spontaneous production of words belonging to the same category or starting with a specific letter. The Trail Making Test, version B (TMT-B; Reitan, 1958) assesses visual and spatial exploration, switching, divided attention, and psychomotor speed during a visuo-motor task (Reynolds & Horton, 2014; Cristofori et al., 2019).

In the literature, verbal inhibition and interference tasks are also commonly reported, such as the Stroop Color Word Interference Test (Stroop, 1935), problem-solving tasks like the Tower of London (Shallice, 1982) and Tower of Hanoi (Humes et al., 1997), and cognitive estimation tests such as the Cognitive Estimation Test (CET; Shallice & Evans, 1978), in which participants are asked questions that require estimating quantities in the absence of precise knowledge. Other paradigms include Go/No-Go tasks (which assess the tendency to respond to a stimulus) and tasks like the Iowa Gambling Task (IGT; Bechara et al., 2000), aimed at evaluating the inability to delay gratification, typical of individuals with orbital and/or medial prefrontal lesions.

In addition, several test batteries provide a more comprehensive assessment of multiple aspects of executive functions, such as the Frontal Assessment Battery (FAB), The Delis–Kaplan Executive Function System (D-KEFS) and the Behavioral Assessment of Dysexecutive Syndrome (BADS) (Antonucci et al., 2014)

The FAB (Dubois et al., 2000), for example, is a brief battery consisting of six subtests that investigate classification, cognitive flexibility, motor programming, sensitivity to interference, and inhibitory control. The FAB (Appollonio et al., 2005) has also been translated, adapted, and validated in Italy, serving as a short cognitive and behavioral screening tool for bedside assessment of individuals with global executive dysfunction. Its widespread use is due to its ease of administration and sensitivity.

A common feature of these tasks is that they require cognitive processing that extends beyond well-learned stimulus–response associations.

Other tools used in Italy include the Phrase-Figure Association Test (Papagno et al., 2007),the Italian version of the CET (Della Sala et al., 2003), and the Phonemic, Semantic, and Alternating Fluency Tests (Costa et al., 2014).

Figure 2 offers an overview of the main neuropsychological assessment tools (Cristofori et al., 2019).

Test	EF component
Clinical rating/bedside assessment	
Cambridge Neurological Inventory (Chen et al., 1995)	Motor initiation, sequencing, and inhibition
Lab-based/constraint environment	
WCST (Heaton et al., 1993)	Switching, perseveration
Verbal Fluency Test (Lezak, 1995)	Verbal production
Design Fluency Test (Jones-Gotman and Milner, 1977)	Nonverbal switching
Stroop Test (Stroop, 1935)	Verbal inhibition
Hayling Sentence Completion Test (Burgess and Shallice, 1996)	Verbal inhibition
Brixton Spatial Anticipation Test (Burgess and Shallice, 1996)	Rule detection and impulsivity
Tower of London (Shallice, 1982) **and Tower of Hanoi** (Humes et al., 1997)	Planning
Sustained Attention to Response Task (Robertson et al., 1997)	Sustained attention and inhibition
Six Elements Test (Wilson et al., 1996)	Planning, strategy allocation
Greenwich Test (Burgess et al., 2000)	Executive memory, planning, and intentionality
Naturalistic environment	
Multiple Errands Test (Shallice and Burgess, 1991)	Strategy allocation, planning
Iowa Gambling Task (Bechara et al., 1994)	Emotion and decision-making
Assessment of Motor and Process Skills (Fisher, 1993)	Motor and process skills
Frontal Systems Behavior Scale (Chiaravalloti and DeLuca, 2003)	Initiation and disinhibition

Among these traditional tools (Figure 2), planning tasks stand out due to their complexity and

centrality in everyday executive functioning, leading to the development of specialized tests such as the Tower of London.

1.10. Focus on planning

Among the various executive functions discussed so far, planning represents one of the most critical components and occupies a central role. It refers to the ability to anticipate future events, formulate goals, generate strategies, and organize actions across time and space to achieve complex objectives (Shallice & Burgess, 1991; Lezak, 1995). The ability to plan plays a central role in organizing behavior, allowing individuals to structure their personal and social lives, to adapt and interact effectively with the environment, and to face and resolve complex and changing situations (Das, Kar, & Parrila, 1996). Planning requires the integration of core processes—working memory to hold and update information, inhibition to suppress irrelevant responses, and cognitive flexibility to adjust strategies as circumstances change. Impairments in planning are often associated with difficulties in daily life, including managing and organizing routines, problem-solving in novel contexts, and achieving long-term goals (Lezak et al., 2012). Lezak (1983) conceptualized planning as the process through which an individual evaluates and organizes the actions necessary to progress from one step to another, identifying a sequence of intermediate stages essential for problem-solving and goal attainment. Planning is defined as a set of cognitive skills that regulate behavior, encompassing environmental representation, anticipation of problem solutions, and monitoring of strategies aimed at achieving objectives (Lezak, 1983). Accordingly, planning involves multiple cognitive abilities, including sustained attention, the capacity to generate alternatives, evaluate and select the most appropriate choices for goal attainment, and develop a conceptual framework to guide actions until task completion. This ability, thus, requires anticipating consequences and mentally organizing actions in sequence before their execution (Lezak, 1982). Owen and colleagues (1997) define planning as the capacity to organize behavior in time and space. Lazeron and colleagues (2000)

describe it as a necessary ability in situations where a goal must be achieved through a series of intermediate steps, each of which alone does not directly lead to the final objective. Similarly, Anderson (2002) defines planning as the capacity to formulate an efficient action strategy to achieve a goal. Across decades, definitions converge in describing planning as the ability to pre-organize a sequence of actions necessary to complete a task and attain a specific objective (Shallice, 1982; Owen, Downes, Sahakian, Polkey & Robbins, 1990; Unterrainer, Rahm, Halsband & Kaller, 2005). Planning is particularly important because it facilitates the design and organization of behavior, enabling individuals to adapt and successfully tackle complex tasks in dynamic environments. In other words, it can be conceived as a cognitive process guiding individual action, involving environmental representation, problem identification, selection of appropriate responses, and monitoring of these responses to evaluate their effectiveness in problem-solving (Scholnick & Friedman, 1987; Fancello, Vio & Cianchetti, 2006). Even in routine daily activities, the ability to plan ahead is essential for effective performance. Since planning involves mental modeling and the anticipation of action consequences prior to execution in the real world (Goel & Grafman, 1995; Ward & Morris, 2005), it is one of the most complex cognitive functions and is closely associated with prefrontal cortex functioning (Shallice, 1982; Unterrainer & Owen, 2006). Cazzato and colleagues (2010) highlighted that planning involves the integration of multiple cognitive processes, including strategy formulation, coordination of mental functions, recognition of goal attainment, and memory representation. They also argue that planning relies on the concept of "cognitive saving," whereby individuals adopt strategies to achieve goals through simplified schemes, called cognitive heuristics. These heuristics are defined as behavioral patterns that, in order to achieve a goal, lead the individual to use a reduced amount of cognitive resources; this results in the generation of a strategy, that is, a set of actions that serve as a guide to reach the solution of the problem (Duncan, 1985). In other words, planning is the ability to organize a series of steps aimed at achieving a goal. Within the scope of planning, it is

essential to consider that implementing future-oriented projects involves a range of higher-order control processes, including focusing attention on target stimuli, inhibiting distractors, suppressing impulsive responses, and maintaining goal representations in working memory (Unterrainer & Owen, 2006). According to Klenberg and colleagues (2001), working memory (WM) is an essential prerequisite for planning and action selection, as these activities rely on information processing within WM. Additional executive functions, such as inhibitory control and cognitive flexibility, also contribute to planning, providing a foundation for predicting successful cognitive performance (Atance & Meltzoff, 2006). Furthermore, Steinberg et al. (2008) and Luciana et al. (2009) identify another crucial aspect for adequate planning: sustained cognitive control of behavior, also referred to as strategic self-monitoring, which ensures task-focused attention until the goal is achieved. Across the numerous definitions proposed by experts over the years, the complexity of planning becomes evident. Planning requires the coordination and modulation of multiple mental functions, including attention, working memory, cognitive flexibility, reasoning, and abstraction. It becomes particularly critical when addressing non-routine problems that require innovative solutions. In summary, planning represents a cornerstone of executive functioning, integrating multiple cognitive processes to enable individuals to anticipate, organize, and adapt their behavior effectively. Its critical role in navigating both routine and novel situations underscores its relevance for adaptive functioning, goal-directed behavior, and problem-solving in complex and dynamic environments.

To assess planning abilities in both research and clinical contexts, one of the most widely used tasks is the Tower of London, which operationalizes planning as a measurable sequence of cognitive steps

1.11. Tower of London task

The Tower of London test (TOL), developed by Shallice in 1982, is a modification to the Tower of Hanoi task (Klahr, 1978; Simon, 1975), later refined by Norman and Shallice in 1986. The Tol is widely used in both clinical and research settings to analyze and assess an individual's planning and

monitoring abilities during task performance. The Tower of London is a planning and problem-solving task that involves several processes, such as task organization, initiation of the plan, the ability to maintain the plan in memory while executing it, the ability to inhibit potential distractors, and the flexibility to modify the strategy when necessary (Injoque Ricle et al., 2014). This tool requires participants to carry out a series of sequential operations only after carefully reflecting on the sequence of actions needed to complete them, allowing them to modify the sequence in case of errors. The TOL involves visuospatial and problem-solving skills, requiring participants to move beads from an initial configuration to a final configuration while adhering to a set of predefined rules. The Tower of London consists of a structure with a base on which three vertical pegs of different heights are located approximately 7.5 cm apart. There are also three colored beads—green, red, and blue—each 4 cm in diameter and centrally perforated to fit onto the pegs. The participant's goal is to reproduce the presented visual model using the number of moves indicated by the examiner (Figure 3). In its original form, the test comprises 12 items distributed by complexity: two items require two moves, two items require three moves, four items require four moves, and four items require five moves. In problems requiring fewer moves, less planning is needed, and a direct movement strategy toward the goal may be effective. For moderately difficult problems, intermediate analysis or error correction is required, while more complex problems demand more elaborate planning strategies (Fancello, 2021). The task also includes route constraints, or rules that participants must consider while solving the problem: (1) a bead can only be moved if no other bead is on top of it; (2) a maximum of three beads can be placed on the tallest peg, two on the middle peg, and one on the shortest peg; (3) two beads cannot be moved simultaneously with both hands, nor can one be temporarily placed on the table. To achieve satisfactory performance on the Tower of London task, participants must demonstrate efficient planning abilities to solve the problem using the fewest possible moves within a defined time frame.

Two primary planning measures emerge from this task: the score and planning time. The score represents the number of problems solved and is commonly used as an indicator of planning ability. Planning time, also known as latency or pre-planning time, refers to the period between the presentation of the target model and the participant's first move, during which an initial plan for solving the problem is generated. Problems requiring more moves are presumed to require longer planning times than those requiring fewer moves. An alternative measure is the level achieved, where each level represents the minimum number of moves required to solve the items, though this measure is less commonly used. Another less frequent measure is the total number of moves made, which does not account for problem complexity but reflects the overall number of movements executed. Excess moves, i.e., additional moves beyond the minimum required, can also be quantified. Other temporal measures derived from this tool include execution time, defined as the time between the first and last move, and total time, which encompasses both planning and execution time. Total time is useful for evaluating whether the participant completed the problem within the allotted time and contributes to assessing solution accuracy. Execution time can provide insights into subjective variables such as distractibility or task engagement but does not directly reflect the ability to generate and organize a sequence of steps to solve the problem (Injoque Ricle et al., 2014).

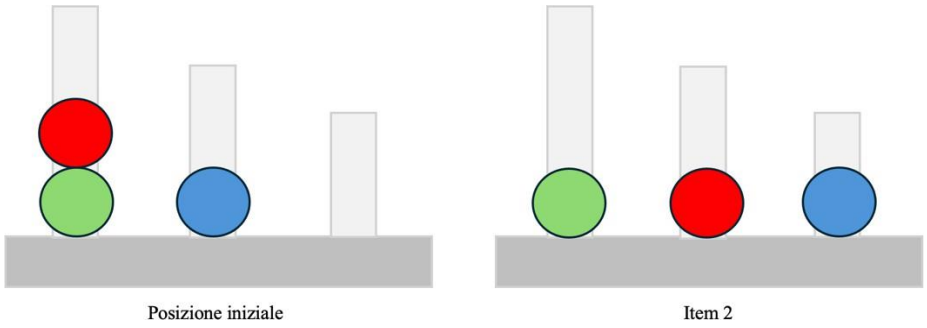


Figure 3 Tower of London: on the right, the initial configuration given to the participant; on the left, the model the participant must reproduce for the second item.

The Tower of London task requires "forward-thinking," commonly referred to as planning, because

an early incorrect move can make the problem virtually unsolvable, as all previous steps would need to be retraced and reversed to correct the inappropriate move (Owen, 1997). According to Owen (1997), successful performance on the Tower of London (TOL) task generally involves several cognitive stages, outlined in the following sequence: (1) assessing the overall situation by comparing the initial state with the goal state; (2) defining a series of sub-goals; (3) mentally generating a sequence of moves to achieve these sub-goals; (4) reviewing the sequence through mental rehearsal; and (5) executing the correct sequence.

Ward and Allport (1997) conducted a series of experiments to investigate the nature of sub-goals in the context of planning within the TOL task. They concluded that a sub-goal represents a "segment of indirect moves," meaning a sequence of movements that does not directly move a disk to its final position but is essential for problem resolution. Tasks requiring such indirect moves generate conflicts between the primary goal (placing disks in the final positions) and the secondary goal (moving a disk away from its final position). These goal–sub-goal conflicts pose particular difficulties for patients with frontal lobe lesions (Goel & Grafman, 1995; Grafman, 1995; Morris, Feigenbaum, Bullock, & Polkey, 1997; Shallice, 1982) and are associated with prolonged planning time on the TOL (Ward & Allport, 1997).

Furthermore, Goel and Grafman (1995) distinguish between the plan formulation phase and its subsequent execution, defining planning as the ability to mentally project into the future, separate from the act of implementing the plan itself. Ward and Allport (1997), in turn, differentiate between preliminary mental planning, which occurs before the task execution begins, and "online" planning, which occurs during the task execution phase. In one TOL study, a time limit of 2.5 seconds per move was set, and participants were required to solve the problems using the minimal number of moves possible. According to Ward and Allport, these conditions ensure that the entire planning process in the TOL is pre-execution mental planning. This approach aligns with Owen's (1997) performance

model, in which all sub-goals are planned in advance before execution.

Owen and colleagues (1995) examined a TOL variant in which participants were asked to plan a minimal sequence of moves to solve a problem, followed by reporting the number of moves needed for the solution. This manipulation resulted in significantly longer planning times compared to conventional TOL methodologies. However, Phillips and colleagues (2001) found that TOL performance is primarily influenced by planning during task execution (online) rather than by preliminary mental planning. In both experiments, the time spent on preliminary planning did not translate into greater efficiency in solving the TOL in terms of accuracy or speed. Participants rarely adopt a fully detailed planning strategy unless explicitly allowed to do so based on task parameters. Investing time in creating detailed mental plans for TOL solutions does not appear to be an efficient approach. Conditions in which participants were "forced" or explicitly instructed to develop complete preliminary plans did not result in faster execution times or fewer moves compared to conditions discouraging or preventing preliminary planning.

Moreover, limitations exist in individuals' capacity to plan an entire sequence of actions in advance during the TOL (Phillips et al., 2001). The requirement to plan the full sequence contrasts with naturalistic planning, where individuals plan only the beginning of a sequence and alternate between execution and planning, following an "opportunistic" planning model (Hayes-Roth, 1979). This type of planning places less demand on memory than complete preliminary planning and, as tasks grow more complex, planning all moves in advance may become extremely challenging due to memory overload (Phillips et al., 2001).

In a preliminary study using the Tower of London (TOL) developed by Shallice in 1982, 12 TOL problems characterized by sequences of 3, 4, and 5 moves were presented to a group of patients with frontal and posterior brain lesions. Patients with left frontal lesions showed a significant reduction in problem-solving within a 60-second time limit, as well as a longer delay in initiating the first correct

move, compared to patients with right anterior or bilateral posterior lesions. Performance on the TOL was not correlated with intelligence or spatial problem-solving ability. Further studies have confirmed the usefulness of the TOL as a tool for assessing higher-order problem-solving in both clinical and neurotypical populations (Lange et al., 1992; Mazzocco, Hagerman, Cronister-Silverman, & Pennington, 1992; Morris et al., 1988; Owen, Downes, Sahakian, Polkey, & Robbins, 1990; Owen et al., 1995).

Concerning the number of problems solved as an indicator of planning, several studies have shown that efficiency in performing the Tower of London task gradually improves from ages 3 to 14, eventually reaching adult levels (Krikorian, Bartok, & Gay, 1994b; Lipina, Martelli, Vuelta, Injoque Ricle, & Colombo, 2004; Mahone et al., 2002; Malloy-Diniz et al., 2008). Fewer investigations have focused on planning or latency time as measures of planning, particularly in typical populations. In research with young adults aged 18 to 25, Phillips and colleagues (1999) reported that planning time increases in accordance with the number of moves required to solve a problem. Conversely, Huizinga, Dolan, and van der Molen (2006) studied groups aged 7, 11, 15, and 21, finding that planning time decreases between ages 7 and 15, with no significant difference between 15- and 21-year-olds. The total number of moves across all problems is rarely used as a measure of planning.

In a developmental study using the Tower of Hanoi in children aged 6 to 12, Díaz et al. (2012) found that planning skills progressively improve with age and identified three distinct stages: first-grade children performed significantly worse than the rest of the sample; second- to fourth-grade children performed similarly, with no significant differences among them; and fifth- and sixth-grade children performed significantly better than the other groups, again with no significant differences between these two grades.

Understanding the typical development of executive functions (EF) is highly important, as their proper functioning is essential for carrying out a wide range of everyday tasks, including those critical

for academic success. Furthermore, EF is often disrupted in developmental disorders such as ADHD (Happé, Booth, Charlton, & Hughes, 2006; Shimoni, Engel-Yeger, & Tirosh, 2012; Willcutt et al., 2005) and autism spectrum disorders (Gilbert, Bird, Brindley, Frith, & Burgess, 2008; Happé et al., 2006; Hill, 2004; Robinson, Goddard, Dritschel, Wisley, & Howlin, 2009). Therefore, knowledge of typical EF development enables professionals to identify potential impairments.

1.12. Digital adaptation and alternative tools for planning assessment

While the Tower of London (TOL) remains a prominent tool for evaluating planning abilities, several traditional neuropsychological tests have been used to assess this cognitive function. For instance, the Porteus Maze Test (PMT), created by Stanley Porteus in 1924, requires individuals to navigate increasingly complex mazes by drawing continuous lines from a starting point to a designated endpoint. This test aims to measure planning capacity and has been shown to be particularly sensitive to frontal lobe lesions. Similarly, the Rey-Osterrieth Complex Figure (ROCF), introduced by Rey in 1941, asks participants to reproduce a complex two-dimensional figure and later redraw it from memory. Inconsistent or fragmented responses can indicate planning deficits. The Bender-Gestalt test, developed by Bender in 1938, involves copying a series of nine drawings, allowing examiners to observe how well participants plan the layout and execution of their drawings, providing insights into their organizational strategies.

In 1979, Klosowska developed the Action Program Test to directly evaluate planning in goal-directed actions. Participants were presented with a set of objects and instructed to achieve a specific goal, such as capping a bottle, by combining a sequence of discrete steps into a coherent action plan. Studies on patients with unilateral or bilateral frontal lesions revealed significant deficits in completing these tasks, reflecting difficulties in planning and organizing everyday activities. While these traditional tests clearly require cognitive planning, they also place substantial demands on visuospatial processing, which may independently contribute to the observed deficits.

To address these limitations, Shallice (1982) developed the Tower of London, which, unlike the Tower of Hanoi, provides participants with information about the number of moves needed to solve each problem, allows for additional attempts in case of errors, and tracks both errors and the time taken to make decisions. These features make the TOL a more refined tool for assessing planning, monitoring, and decision-making in a structured and measurable way, providing a reliable method to evaluate higher-order executive functions.

Building upon the TOL, several computerized versions have been developed to adapt the test to different needs and specific experimental or clinical contexts (Kaller et al., 2012; McKinlay & McLellan, 2011; Mohai et al., 2022; Owen et al., 1995; Robbins et al., 1994; Ruocco et al., 2014; Ward & Allport, 1997). Some of these versions remain very close to the original task, while others diverge to suit particular populations, modifying the number of items, the variables assessed, or the presentation of the apparatus.

Among the most well-known adaptations is the Stockings of Cambridge (SOC), part of the Cambridge Neuropsychological Test Automated Battery (CANTAB; Robbins et al., 1994). The SOC replaces the traditional pegs with three "pockets," maintaining the same bead capacity as the original pegs: the first can hold three beads, the second two, and the third one. Unlike the original 3D physical apparatus, the SOC uses a screen displaying two images: a static upper image showing the final configuration, and a lower image that participants can manipulate. The initial phase of the SOC requires the computer to make a few moves on the upper image, which participants then replicate on the lower image. This phase provides a baseline for latency and execution times, which are then used to calculate planning and execution times in the second part of the test (Luciana & Nelson, 1998). The objective remains the same as in the original TOL: to move the beads to reach the final configuration within a predefined number of moves.

Owen and colleagues (1995) developed a "One Touch" version, similar to the SOC but simplified

with a touchscreen, useful for investigating planning deficits in patients with Parkinson's disease.

A more innovative approach was proposed by Mohai et al. (2022), in which items are tailored to participants' abilities, selecting problems based on previous responses. This represents an attempt at computerized adaptive testing, structured in three phases: a familiarization phase with animated exercises and two practice trials with feedback, an assessment phase with up to four items of varying difficulty, and finally a testing phase in which problems are generated based on the participant's prior performance.

Ward and Allport (1997) proposed a digital version closely mirroring the original, with minor modifications: the pegs are of equal height, and the beads are replaced with four colored discs. Participants must plan the sequence of moves in advance and press the "Ready" button to begin the task. Another faithful digital adaptation was developed by McKinlay and McLellan (2011).

Finally, the Scarborough Tower of London (S-TOL), developed by Ruocco et al. (2014), was designed for use in neuroimaging protocols. Participants interact with the computer using specialized mice and must observe the problem for seven seconds before responding to a specific question, such as whether it can be solved in two moves. This version allows for a more precise and controlled assessment of planning and problem-solving abilities. The variety of TOL versions reflects the need to adapt the test to different clinical populations and testing contexts (Owen et al., 1995).

Taken together, these traditional and digital adaptations illustrate the evolution of planning assessment, highlighting both the challenges of measuring complex executive processes and the advantages offered by modern computerized tools.

2. Assessment of Fluid Intelligence

2.1. Definition and theoretical models of Intelligence

The term intelligence is used in many contexts, even outside the scientific field, in everyday life. This contributes to making it an elusive concept, defined in different ways depending on cultural and situational contexts. In general, intelligence is described as a mental ability useful for reasoning, problem-solving, and learning. As described in the study by Colom and colleagues (2010), intelligence integrates several functions, including perception, attention, memory, language, and planning. The APA recognizes intelligence by emphasizing individual differences: "individuals differ from one another in their ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought" (Colom, Karama, Jung, & Haier, 2010, p. 490).

To describe intelligence, one can also refer to the two definitions proposed by Cornoldi (2007): general definitions, which consider intelligence as mental activity or higher-order mental operations, linking it to the set of cognitive functions with adaptive value for the individual; and differential definitions, which instead view it as the element that differentiates individuals in their ability to deal with cognitive tasks (Cornoldi, 2007, p. 17).

From a historical perspective, the first researcher to address intelligence and its measurement was Sir Francis Galton, who conceived intelligence as an inheritable construct—an overarching ability that could be measured by considering the speed of mental processes. In 1869, he published Hereditary Genius, where he argued that intelligence was entirely hereditary. Galton's perspective was strongly influenced by an evolutionary approach, likely due in part to his cousin Darwin, and he extended this framework to intelligence. His aim was to use quantitative tests to distinguish more intelligent individuals—those with faster response times and, therefore, more rapid nervous processes—from less intelligent ones.

Since then, the construct of intelligence has evolved and expanded over the years, giving rise to the development of different theories.

There are several approaches to the study of intelligence, which differ according to the conception of the construct itself that, as noted above, lacks a single, universally accepted definition. In general, the approaches that can be described, as indicated by Sternberg, Jarvin, and Grigorenko (2011), are the following:

- Psychometric approach: the assumption underlying this approach is to render the concept of intelligence scientific, measurable through tests that are easy for professionals to administer. Intelligence is conceived as hereditary and stable; therefore, the environment and culture to which individuals belong are not regarded as fundamental aspects in its development, but rather as secondary factors that need to be considered and controlled.
- Systemic approach: most prominently represented by Gardner and Sternberg, who developed theories of intelligence considering it as a system composed of multiple aspects and characteristics.
- Cognitive and biological approaches: the cognitive perspective examines how individuals mentally represent, and process information learned from the environment, while the biological perspective focuses on identifying physiological indicators of intelligence, considering the brain as the organ responsible for these processes.

Theories of intelligence can also be categorized as unitary, multiple, or hierarchical. Spearman's (1904) theory aimed to study intelligence by identifying a general factor (*g*). Spearman was the first to apply a statistical method—factor analysis—to the study of intelligence and its components. Factor analysis "consists first in calculating the correlations among all tests taken in pairs and then verifying whether there are groups of tests that correlate more strongly with each other than with others. It is then supposed that there is a common factor of variation for this group of tests" (Huteau & Lautrey,

2000). By applying this method to intelligence, Spearman theorized the existence of a general factor common to all tasks, together with specific factors for each task that contribute to intelligence. The *g* factor, hierarchically superior, represents the general intellectual level, considered immutable to the effects of environment or schooling, whereas the *s* factors can be enhanced through education, learning, and social, academic, or familial contexts. Thus, an individual's performance on a task depends on both the general *g* factor and the influence of the task-specific factor. Even today, Spearman's theory remains a reference model for intelligence measurement. One of the main tests designed to assess the *g* factor is Raven's Progressive Matrices(Raven, 1989), which will be described later as a key tool employed in the present dissertation research.

Using factor analysis in a fundamentally different way, Thurstone proposed the Primary Mental Abilities Theory, which viewed intelligence as composed of several primary abilities: verbal comprehension (vocabulary tests), verbal fluency (word production tasks), numerical ability (mathematical problem-solving tasks), spatial ability (tasks requiring mental rotation of objects and figures), memory (recall tasks), inductive reasoning (analogies and series completion), and perceptual speed (tasks requiring figure recognition). This represents a multiple, psychometric theory of intelligence.

Within the multiple theories tradition, Guilford (1967) developed the Structure of Intellect (SOI; see Figure 4) theory, in direct opposition to Spearman's unitary conception of intelligence. He hypothesized that intelligence was composed of 150 independent factors. According to Guilford, intelligence can be understood as the intersection of three dimensions: operations, content, and products. Operations refer to basic mental processes such as perception, memory, or production; content refers to the material of the problem (visual, auditory, symbolic, semantic, or behavioral); and products refer to the responses provided to problems, including units (single words, numbers, or figures), classes, inferences or implications.

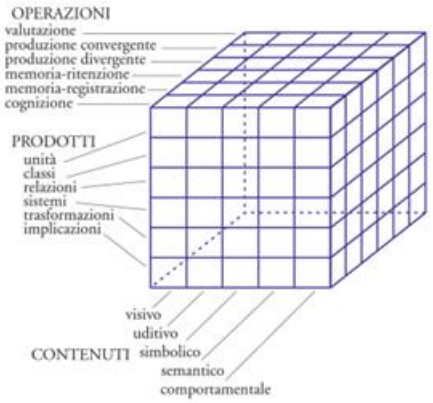


Figure 4 – The Structure of Intellect (Guilford, 1967)

Several theories expanded on Spearman's *g* factor. Horn and Cattell (1966) proposed a hierarchical theory that incorporated *g* within two additional factors: fluid intelligence (*gf*) and crystallized intelligence (*gc*). Fluid intelligence, influenced by *g*, tends to decline with age and is considered central, while crystallized intelligence is shaped by social and cultural environment and is not age-dependent. More specifically, fluid ability is an innate phenomenon with a hereditary basis that is influenced by age and the aging process, increasing until adolescence and then declining across the lifespan. It corresponds to the capacity for judgment and reasoning, as well as the ability to solve novel problems, and is not determined by culture or experience. Crystallized intelligence, on the other hand, refers to knowledge acquired about the nature of the world. It is strongly influenced by the amount of instruction, experience, and knowledge to which one is exposed and is therefore considered a product of learning. Crystallized intelligence may take the form of semantic knowledge and may involve the ability to understand communicated messages, as well as the capacity for judgment and reasoning in everyday situations. The content of crystallized intelligence is culture-specific and changes over time in parallel with cultural transformations; it does not decline across the lifespan but improves until approximately the age of 60, due to the cumulative effect of experience. Fluid intelligence enables the acquisition of crystallized intelligence.

Later, Vernon (1971) proposed a theory dividing Spearman's *g* into two main abilities: verbal-educational ability and mechanical-spatial ability. The former is influenced by schooling and thus learned, while the latter is not. This model was considered an intermediate position between Spearman's unitary approach and Thurstone's multifactorial one, though it is no longer widely used today due to its demonstrated limitations.

Carroll (1993) also proposed a hierarchical model of intelligence, expanding on Spearman's *g*. His model took the form of a three-stratum pyramid (Figure 5):

1. Stratum I: specific abilities influenced by experience and education:
2. Stratum II: broad abilities, stable across individuals, that influence behavior (e.g., crystallized intelligence, fluid intelligence, learning and memory, visual perception, verbal fluency, perceptual speed). These are considered group factors and are influenced by *g*;
3. Stratum III: the general cognitive ability (*g* factor), underlying all intellectual activities.

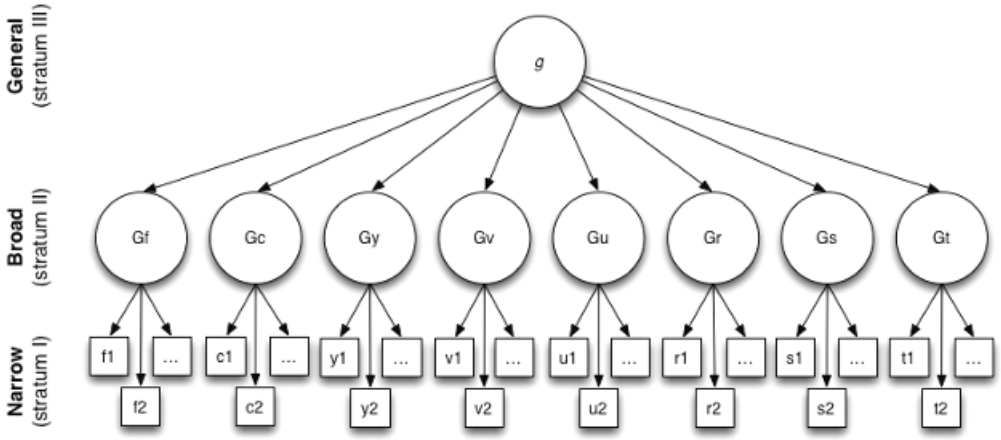


Figure 5. The Three-Stratum Theory by Carroll (1993).

Among multiple theories, one of the most well-known is Gardner's (1983) Theory of Multiple Intelligences. According to this framework, intelligence cannot be conceived as a unitary construct with a single general component; instead, there are at least eight independent types of intelligence.

Thus, a person excelling in just one of these is considered intelligent to the same degree as others. The intelligences identified are:

1. Visual-spatial intelligence: ability to create mental images and solve non-visual problems;
2. Linguistic-verbal intelligence: mastery of language, including rhetorical and poetic expression;
3. Logical-mathematical intelligence: deductive reasoning, logical thinking, and problem-solving;
4. Bodily-kinesthetic intelligence: effective coordination of body movements through mental control;
5. Musical intelligence: ability to recognize and compose music, with a strong sense of rhythm and tone;
6. Interpersonal intelligence: ability to understand others;
7. Intrapersonal intelligence: ability to understand one's own feelings and motivations;
8. Naturalistic intelligence: ability to discriminate among living beings and other features of the natural world, often linked to exploration and learning from nature.

Equally important is Sternberg's Triarchic Theory of Intelligence (1985), which defines intelligence as comprising three main abilities: creative, analytical, and practical. In a later revision, renamed the Theory of Successful Intelligence, a fourth component—wisdom-based ability—was added. In this model, individuals are considered intelligent if they are able to achieve their goals within their given environmental context, recognize their strengths and weaknesses, adapt or shape environments accordingly, and integrate creative, analytical, practical, and wisdom-based skills. This theory has also been applied to giftedness, defining a gifted individual as one who demonstrates the four characteristics of wisdom, analytical intelligence, practical intelligence, and creativity. For Sternberg, intelligence encompasses three relations: with the individual's internal

world, with experience, and with the external world. Regarding the internal world, intelligence involves information processing through three components: metacognitive components responsible for executive functions, performance components, and knowledge acquisition components. Experience interacts with and influences these components: when faced with a novel task, greater intelligence is required to solve it. Finally, the four abilities (creative, analytical, practical, wisdom-based) enable three external functions: adapting to existing environments, shaping them to create new ones, and selecting alternative environments. These aspects of intelligence can be assessed through specific tests, providing predictions of an individual's performance in both academic and non-academic domains.

In 1997, Kevin S. McGrew emphasized the lack of a clear link between theoretical and empirical research on the factors of intelligence and the development of instruments for assessing cognitive abilities and academic performance. His aim was to provide clinicians with the possibility of selecting the most appropriate intelligence scale depending on the specific abilities to be measured. To this end, McGrew proposed creating a taxonomy of cognitive abilities that would coincide with the formulation of a new model of intelligence. Based on this model, he reclassified the subtests of different scales. By comparing the Horn-Cattell and Carroll (1993) models, McGrew (1997) noted both strong similarities and important discrepancies—such as the presence of a general factor *g* in Carroll's model, its absence in the Cattell-Horn framework, and the different conceptualizations of broad versus narrow abilities. In July 1997, in a personal communication, Horn and Carroll agreed that their two models could be unified into the Cattell-Horn-Carroll (CHC) Theory of Cognitive Abilities. This framework proposed both broad and narrow abilities, with the broad factors including: Crystallized Intelligence (Gc), Visual Processing (Gv), Quantitative Knowledge (Gq), Reading and Writing Ability (Grw), Short-Term Memory (Gsm), Fluid Intelligence (Gf), Processing Speed (Gs), Long-Term Storage and Retrieval (Glr), Auditory Processing (Ga), and

Reaction Time (Gt) (Rivolta et al., 2010).

Through the CHC model, McGrew classified the subtests of the most widely used intelligence scales, with the goal of selecting those subtests that most precisely measure specific abilities. The contributions of Carroll (1993) and McGrew (1997) transformed both the construction criteria and the interpretative parameters of next-generation instruments. In earlier assessments, particularly those developed before the mid-1980s, the underlying theoretical model played only a secondary role, meaning that many instruments were not based on the Gf-Gc framework.

The shift from viewing intelligence as a unitary g factor to conceptualizing it as a set of multiple abilities led to a reduction in the emphasis on the Full Scale IQ (FSIQ), an increase in the number of composite scores to be calculated, and a greater level of specificity in their interpretation. For instance, in the Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV; Wechsler, 2003), the traditional Verbal IQ and Performance IQ indices were eliminated, the importance of the FSIQ was diminished, and greater emphasis was placed on the indices of Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed. This restructuring highlighted the multidimensional nature of intelligence.

2.2. Development of fluid intelligence across the lifespan

The earliest theories of cognitive development across the lifespan (Lifespan Theories of Cognitive Development, LTCD) can be traced back to Johann Nicolaus Tetens (1736–1807), a philosopher and psychologist, who observed that well-trained abilities are less likely to decline with advancing adult age than basic abilities. Most LTCD approaches rely on two central assumptions identified by Baltes (1987; Baltes et al., 1999): cognitive development reflects the interplay of two interwoven components, one biological and the other cultural. The biological component is typically interpreted as the expression of the neurophysiological architecture of the mind. Age-related changes are expected to be strongly influenced by genetic factors and the organizational properties of the central

nervous system, leading to a gradual decline with increasing maturity. In contrast, the cultural component refers to knowledge resources available through culture and experience; developmental changes here are expected to mirror the accumulation of culturally transmitted knowledge, increasing throughout adulthood, remaining relatively stable into old age, and declining only at advanced stages. The most influential LTCD model is the Cattell-Horn theory of fluid and crystallized intelligence that recognizes that age-related changes result from the joint functioning of these two systems, as the outcome of complex and dynamic interactions between biological factors (e.g., brain maturation and aging) and environmental influences (Lindenberger et al., 2001).

As previously described, fluid reasoning (Gf) is the ability to think logically and solve problems in novel situations, independently of previously acquired knowledge (R. B. Cattell & Cattell, 1987). It is an essential component of cognitive development (Goswami, 1992), serving as a scaffold for the acquisition of other abilities (Blair, 2006; Cattell, 1971, 1987) and predicting academic, professional, and life outcomes (Gottfredson, 1997). Gf is thought to emerge within the first two to three years of life, following the development of general, perceptual, attentional, and motor capacities (Cattell, 1987). It develops rapidly during early and middle childhood, continues to increase more slowly into early adolescence, reaches an asymptote in mid- to late adolescence, and stabilizes around age 25, before starting to decline gradually by the mid-30s, with a more marked decline from the mid-40s onwards (Li et al., 2004; McArdle et al., 2002; Lifshitz et al., 2021). By contrast, crystallized intelligence follows a slower trajectory, peaking around age 40, remaining relatively stable through much of late adulthood, and declining only in advanced age (Li et al., 2004; Ackerman, 2008).

As shown in Figure 6 (Cattell, 1987; McArdle et al., 2002), both Gf and Gc increase throughout adolescence into early adulthood. Gf then declines more rapidly, while Gc continues to rise until the ages of 60–70.

.

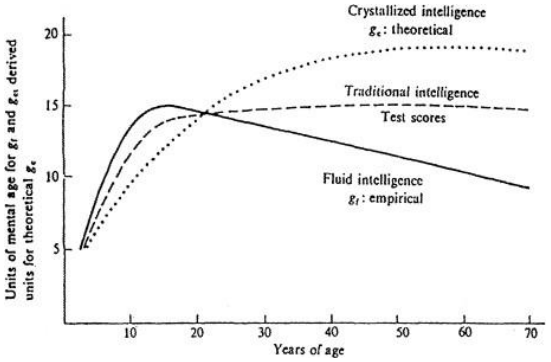


Figure 6. Theoretical description of intellectual ability curves across the lifespan. Source: Intelligence: Its Structure, Growth and Action (p. 206), R. B. Cattell, 1987, Amsterdam: North-Holland. Copyright by Elsevier Science Publishers. (McArdle et al., 2002).

The combined curve of these functions, traditional intelligence, obscures important distinctions in the developmental trajectories of the two constructs. Jones and Conrad (1933) described the main characteristics of the Gf curve as linear growth until about age 16, a negative deceleration with a peak between 18 and 21 years, followed by a gradual decline, such that by age 55, abilities return to levels comparable to those observed in mid-adolescence. Even in the first years of life, traces of goal-directed planning and problem-solving behavior are evident. Tool use, for example, reveals emerging sequential planning abilities: by eight to nine months, infants can remove a cloth or pull a string to retrieve a hidden object, or repeatedly drop toys to prompt parental responses (Jones & Conrad, 1933). Dumontheil (2014) reviewed evidence showing that preschoolers already display iconic analogical reasoning and the ability to use prior knowledge to solve novel but similar problems. By the age of one, children can select strategies based on both problem difficulty and their knowledge of strategy effectiveness. With development, they progress from rigid, outdated strategies toward more flexible and adaptive approaches. During adolescence, deductive and inductive reasoning skills emerge, enabling more abstract and logical thought. Around 11 or 12 years of age, children refine conditional reasoning, and as Gf continues to develop, they become capable of solving increasingly complex tasks.

Play also assumes a central role in fostering problem-solving skills and fluid reasoning. Prince (2017) emphasized the role of natural curiosity in motivating infants and children to explore their environment, thereby developing problem-solving and fluid reasoning abilities through repeated experience. Hanscom (2016) argued that unstructured outdoor play promotes the development of both procedural and declarative knowledge, leading to neural changes—particularly in the frontal lobes—in ways distinct from structured games governed by explicit rules and fixed goals. This is consistent with the view that the fronto-parietal network, described as an "executive control circuit," supports the development of fluid intelligence and decision-making by integrating both internal and environmental experiences (Prince, 2017).

2.3. Traditional tools and assessment methods for fluid intelligence

Among the traditional tools for assessing fluid intelligence, the most widely used are primarily nonverbal tests designed to minimize the influence of cultural and linguistic factors, providing a more direct measure of reasoning abilities independently of prior knowledge. R.B. Cattell (1940), in his studies on fluid and crystallized intelligence, argued that even across different cultural groups there exists a common core of knowledge that allows reasoning operations. In line with this theoretical perspective, the Culture Fair Intelligence Test (CFT) was developed to reduce the effects of cultural learning and social environment, without compromising the predictive value for practical behaviors. The CFT is based on nonverbal stimuli, ensuring accurate cross-cultural comparisons and minimizing cultural bias, and it differs from traditional verbal and numerical tests by being specifically aimed at measuring fluid intelligence (R. Cattell & Cattell, 1981). The test is designed to assess general intelligence from early childhood through adulthood, also helping to identify individual potential for professional tasks requiring specific cognitive abilities, making it useful for educational and occupational guidance. The original 1949 version, which remains the basis of the test today, is organized into three scales differentiated by difficulty, each comprising

four sub-tests preceded by examples: the Series sub-test involves completing an incomplete progressive series of four figures; the Classifications sub-test requires identifying figures that differ from the others; the Matrices sub-test evaluates the ability to correctly complete a matrix; and the Conditions sub-test involves selecting the alternative that duplicates the conditions shown (R. B. Cattell, 1949; R. Cattell & Cattell, 1981). Scale 1 is intended for children aged 4 to 8 years and for adults with cognitive impairments, comprising eight sub-tests and requiring the ability to understand and respond to verbal instructions. Scale 2, with two levels (A and B), is suitable for subjects aged 8 and above, as well as adolescents and adults with average cognitive abilities. Scale 3, also with two levels, contains more difficult items and is intended to provide more precise estimates of higher intelligence ranges. The untimed administration allows assessment of additional variables such as cultural background, personality traits, motivation, and attitudes toward time management (R. Cattell & Cattell, 1981). The updated version, CFT-20-R, published in 2003 and made available in Italian in 2019, preserves the original structure in two parts: Part 1, with 56 items, allows a brief administration, and Part 2, with 45 items, together with Part 1 constitutes the full test. Administration of the complete test takes about one hour and can be conducted individually or in groups. The CFT-20-R is highly versatile and suitable for clinical and psychodiagnostics evaluation, including assessment of children and adolescents with dyslexia, non-native speakers, or individuals with low educational levels, as well as in personnel selection and neuropsychological research on fluid intelligence (Weiss, 2019; Saggino & Tommasi, 2019). Another widely used nonverbal measure of cognitive abilities and fluid intelligence is the Leiter Performance Scale, originally developed by Russell Leiter in 1929 to assess children with difficulties in verbal response (Bradley-Johnson, 1998). The test was designed for children and adolescents from multicultural backgrounds, using a unique “block and frame” response method that requires participants to move wooden blocks within a wooden frame to complete puzzles, figures, numerical sequences, visual

matching, and geometric or pictorial object sequences (Roid & Koch, 2017). The original administration did not require verbal input from the examiner or participant; instructions were pantomimic, and responses involved manipulation of the blocks. The Leiter-R and the more recent Leiter-3 (2013) revised the standardization to include nationally representative normative samples and expanded the age range, making the test suitable for individuals with communication disorders, hearing or motor impairments, traumatic brain injury, attention deficits, or learning difficulties (Roid & Miller, 1997; Roid & Koch, 2017; Salvia & Ysseldyke, 1991). The test conceptualizes nonverbal cognitive abilities in terms of memory, attention, reasoning, and visualization. Administration relies on images, figures, and symbols with pantomimic or other nonverbal instructions, ensuring no verbal interaction is required during testing, although examiners may interact verbally between sub-tests if participants are capable of hearing and speaking. The 20 sub-tests of the Leiter-R are divided into two main batteries: Visualization and Reasoning (VR) – including Figure Ground, Design Analogies, Form Completion, Matching, Sequential Order, Repeated Patterns, Picture Context, Classification, Paper Folding, and Figure Rotation – and Attention and Memory (AM) – including Associated Pairs, Immediate Recognition, Forward Memory, Attention Sustained, Reverse Memory, Visual Coding, Spatial Memory, Delayed Pairs, Delayed Recognition, and Attention Divided (Bradley-Johnson, 1998). Administration time for each battery is approximately 40–45 minutes. The Leiter-3, designed for children, adolescents, and adults aged 3 to 75+, includes two series of sub-tests: the first series assesses cognitive abilities, with four sub-tests providing an estimate of nonverbal IQ and the fifth focusing on attention or memory; the second series supplement the assessment of cognitive processes associated with disorders. Sub-tests include Figure Ground (FG), Form Completion (FC), Classification/Analogies (CA), Sequential Order (SO), Visual Patterns (VP), Attention Sustained (AS), Forward Memory (FM), Reverse Memory (RM), Attention Divided (AD), and the Nonverbal Stroop (NS), a

nonverbal version of the Stroop interference task (Roid & Koch, 2017). Administration times range from 20–40 minutes depending on the battery. In summary, tests such as the CFT and the Leiter constitute established traditional methods for assessing fluid intelligence, aiming to evaluate reasoning independently of prior knowledge and cultural context.

Finally, Raven's Progressive Matrices constitute one of the primary references in the assessment of fluid intelligence, due to their ability to estimate abstract reasoning through nonverbal visual stimuli that are culturally neutral. The high reliability and validity of these instruments make them particularly suitable for both clinical practice and psychological research. The psychometric characteristics and applications of the Raven tests will be discussed in detail in the following paragraphs.

2.4. Raven's Progressive Matrices

In 1938, one of the most widely used tests for assessing fluid intelligence was developed: the first form of Raven's Progressive Matrices. John Carlyle Raven created this test in response to a need identified during a study on the biological and environmental origins of intellectual deficits, conducted with geneticist Lionel Penrose, to assess individuals who were unable to follow written instructions. Raven aimed to capture the essence of intelligence, independent of culturally acquired knowledge from formal education or personal experience (Picone et al., 2018).

Raven specifically intended to design tests that were easy to administer and interpret, while remaining theoretically meaningful. In doing so, he sought to make the two main components of general intelligence directly measurable through robust procedures. These two components, as identified by Spearman (1923), are deductive ability, defined as the capacity to derive meaning from confusion and generate high-level, typically nonverbal, patterns to manage complexity, and reproductive ability, the skill to absorb, retain, and reproduce information explicitly communicated by others.

Raven's Progressive Matrices (1938) were designed to assess the full range of mental abilities and to be generally applicable across age, culture, nationality, and physical condition. The stimuli in the matrices originated from Raven's reinterpretation of experiments by Spearman, who had asked participants to verbally describe the relationships among parts of visual stimuli. Raven retained these stimuli but removed the verbal component, requiring participants to select the figure that completed the matrix, with correct responses depending on the comprehension of inter-part relationships.

The first test form consisted of five sets (A–E), called "Series," each with 12 items of progressively increasing difficulty. This initial version, known as Progressive Matrices (PM 38), revealed limitations as a "one-size-fits-all" tool for children and adults, being too difficult for children and insufficiently discriminating for intellectually gifted adults. This led to the development of two revised versions: the Advanced Progressive Matrices (APM) for higher-ability individuals and the Colored Progressive Matrices (CPM) for those with lower abilities.

The APM includes Series I, which covers all cognitive processes of the original matrices and can be used as a brief screening tool for adults or as an introduction to Series II, which examines "all analytical and integrative operations involved in higher-order thinking" (Raven, 1938), allowing discrimination among individuals with superior intellectual abilities. The CPM, a simplified colored version, is intended for children or adults suspected of intellectual weakness. It retains Series A and B of the original form, while adding 12 new items in Series AB, of intermediate difficulty. Parallel versions have also been developed for CPM and SPM to minimize practice effects while maintaining equivalent difficulty levels.

The Standard Progressive Matrices (SPM) measure a participant's ability to understand abstract figures. The task involves observing the figures, identifying patterns, and selecting the figure that completes the matrix using nonverbal logical reasoning. The SPM consists of 60 items divided into

five series (A–E), each with 12 items. Within each series, the first item is usually straightforward, while subsequent items increase in difficulty, allowing precise evaluation of analogical reasoning. Completion of the test in the standard order enables both individual and group administration, and the final score reflects intellectual ability regardless of education or background. Each series addresses distinct cognitive themes: continuous patterns (Series A), analogical relationships (Series B), progressive pattern transformations (Series C), figure permutations (Series D), and decomposition of figures into constituent parts (Series E) (Burke, 1958). Angelini (1969) analyzed item characteristics, noting that Series A presents simple perceptual units structured according to Gestalt principles, solvable through a perceptual approach. Series B focuses on analogical reasoning, present in children from around age six. Series C emphasizes abstract reasoning and the concept of magnitude, generally acquired around age three. Series D increases the number of varying parameters, requiring identification of changing elements and their direction, while the last items introduce distractors, with related reasoning skills developing around age 13. Series E involves abstract reasoning over component elements and their combinations, typically mastered after age 14 (J. C. Raven et al., 1989).

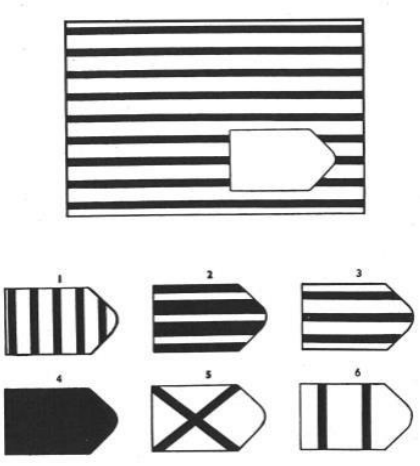


Figure 7. Example of SPM, item 6 from Series A (J. C. Raven et al., 1989).

The Colored Progressive Matrices (CPM) were designed for clinical or research contexts, including children, older adults, and individuals with linguistic or verbal reasoning difficulties, such as

immigrants, hearing-impaired individuals, or those with intellectual weakness. CPM comprises 36 matrices grouped into three series (A, AB, B) of increasing difficulty.

Two formats exist: a manipulative/pegboard version, with boards and movable pieces, and a booklet version. The manipulative format, used with children aged 3–6 and adults with cognitive impairments, allows observation of problem-solving processes, including intermediate attempts. The booklet version, suitable from age 3, can be administered individually or in groups, with participants selecting the missing piece from six alternatives, typically from age 8½ onward.

CPM engages all perceptual reasoning processes, including identification of relationships within a meaningful perceptual context, suitable for children aged 3–11. According to Raven et al. (1998), CPM assesses the ability to compare reason analogically and organize spatial perceptions into systematically interconnected sets. Das's (1989) cognitive theory suggests problem-solving requires attention, processing, and planning. Stimuli are processed holistically as a Gestalt, integrated with the original matrix, and combined sequentially. Planning involves setting goals, implementing strategies, and monitoring performance.

Series A focuses on identification based on shape, color, size, quantity, direction, orientation, figure/ground, and density. Series AB requires recognition of symmetry through Gestalt configurations. Series B involves analogical and conceptual reasoning, recognizing abstract relationships and maintaining them in working memory.

CPM is primarily used to measure cognitive development in typically developing children aged 3–11, diagnose developmental delays in children and adults, and assess cognitive decline in older adults with or without neurological conditions such as aphasia or dementia (Belacchi et al., 2008). Scores are converted to percentiles, with the 25th–75th percentile indicating average intelligence, ≤5th percentile indicating severe deficits, and ≥95th percentile indicating giftedness. Qualitative error analysis further allows identification of specific cognitive weaknesses, with errors categorized

by Raven et al. (1998) as differences, inadequate identification, figure repetition, or incomplete relations. These categories do not fully encompass all errors, as mistakes may overlap and CPM was not originally intended to assess inferential processes.

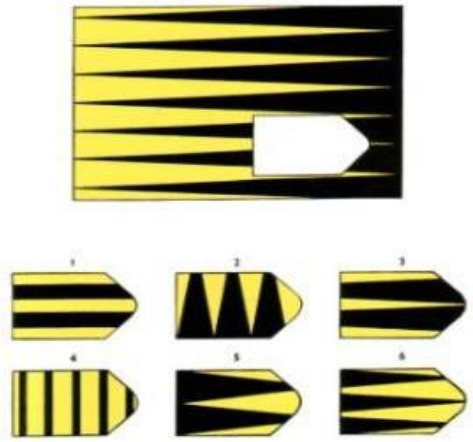


Figure 8. Example of CPM, item 10 from Serie A (J. C. Raven, 1947).

Alongside strictly psychometric use, the Progressive Matrices can also provide relevant qualitative data: through the analysis of errors for each item, it is possible to identify potential sector-specific deficits in particular skills. As noted by Raven, Court, and Raven (1990), incorrect responses are not produced randomly and can be used to infer the processes and strategies employed in solving the task, allowing differences between groups to emerge and helping to understand their underlying causes. Based on item analysis, four types of errors have been identified (Raven, 1938, 1980): incomplete solutions, arbitrary lines of reasoning, over-determined choices, and repetition.

2.5. Digital adaptations and alternative tools for fluid intelligence

Even in the domain of fluid intelligence (FI), significant progress has been made in the development of computerized tests to overcome the limitations of traditional paper-and-pencil assessments. Fluid intelligence, as extensively illustrated in this dissertation, represents the capacity to reason logically, process novel information, learn, solve problems in unfamiliar situations, and perceive relationships among elements. Deficits in executive functions (EF) and FI are commonly observed in neurodevelopmental disorders as well as in conditions resulting from brain injury, psychiatric

disorders, and age-related cognitive decline. In these clinical populations, responding to traditional assessments and maintaining the required attention load can be particularly challenging. Consequently, there is a critical need for tools that provide personalized, adaptive, and timely assessments to support early and individualized rehabilitative interventions.

Early computerized adaptations of FI tests involved projecting the same stimuli used in traditional Raven Matrices test and asking participants to select the correct response by pressing a button (Calvert & Waterfall, 1982). Later studies administered the Standard Progressive Matrices (SPM; Raven, 2009) and the Colored Progressive Matrices (CPM; Raven, 1982) via computer to school-aged children, high school students, psychiatric patients, university students, and other populations. However, these early digital versions differed only in format and did not reduce test completion time. Styles and Andrich (1993) developed a single test using items from the SPM and Advanced Progressive Matrices (APM) for children aged 10 to 16, employing a simple adaptive algorithm: if the participant answered an item incorrectly, the next item was presented in order of difficulty, and the test ended after five consecutive errors. More recently, the Hansen Research Services Matrix Adaptive Test (HRSMAT; Hansen, 2019) has provided an adaptive assessment of nonverbal intelligence, primarily targeting individuals with autism spectrum disorder but also applicable to typically developing populations. The HRSMAT algorithm selects items based on the participant's age and ability, limiting administration to a maximum of 15 minutes.Beyond adaptations of Raven's matrices, other computerized and web-based tools have emerged to assess fluid intelligence in neuropsychological contexts. The NeuroCognitive Performance Test (NCPT) offers a web-based platform that evaluates multiple cognitive domains, including fluid reasoning, memory, and attention, and has been used in large-scale research and clinical studies (Morrison et al., 2015). The Cogstate Brief Battery (CBB) provides computer-administered cognitive assessments, including nonverbal reasoning tasks, and has demonstrated reliability and validity in

both clinical and healthy populations (Maruff et al., 2009). Similarly, BrainCheck offers remote cognitive testing, facilitating assessment in populations with limited access to traditional evaluation settings (Ye et al., 2022). These tools exemplify the ongoing shift toward adaptive, efficient, and ecologically valid measures of fluid intelligence, particularly suited for populations with attentional or cognitive weaknesses.

3. The assessment of executive functions and fluid intelligence in Specific Learning Disorders (SLD)

3.1. Overview of Specific Learning Disorders

Specific Learning Disorders (SLD) comprise a constellation of clinical conditions—in particular dyslexia, dysorthography, dysgraphia, and dyscalculia—which often tend to co-occur, although they may also appear in isolation, likely due to shared genetic bases and partially overlapping neurofunctional anomalies involving reading, writing, and arithmetic abilities (American Psychiatric Association (APA), 2013). SLD are disorders restricted to specific cognitive domains that do not affect general cognitive functioning; nevertheless, their consequences can be pervasive, impacting multiple areas of cognition as well as personal and social adaptation (CC-ISS, 2022). In the International Classification of Diseases, tenth edition (ICD-10; 1992), SLDs are listed as Specific learning disorders(F81), under the categories of neurodevelopmental disorders, referring to conditions in which typical acquisition of learning abilities is disrupted from the early stages of development. Based on the functional deficit, the following clinical conditions are commonly distinguished:

- Specific reading disorder (Dyslexia): persistent deficit in decoding and reading skills (ICD-10, F81.0).
- Specific spelling disorder (Dysorthography): persistent deficit in phonographic coding and orthographic skills (ICD-10, F81.1).
- Specific arithmetic disorder (Dyscalculia): persistent difficulty in acquiring numerical understanding and calculation skills (ICD-10, F81.2).
- Mixed disorders of scholastic skills: combination of Dyslexia, Dysorthography, and/or Dyscalculia (ICD-10, F81.3).

- Disorder of written expression (Dysgraphia): persistent deficit in graphomotor skills (ICD-10, F.81.8).

Their expression is highly heterogeneous and may involve various components of the cognitive-linguistic system—such as attention, executive functions, memory, and lexical access—and they may also co-occur with other neurodevelopmental disorders underpinned by these functions, including attention-deficit/hyperactivity disorder (ADHD), developmental language disorder (DLD), or developmental coordination disorder (DCD) (Cornoldi & Tressoldi, 2014;CC-ISS, 2022). The frequent comorbidity with other neurodevelopmental conditions has, in recent years, prompted a reconsideration of the defining feature of "specificity" in these disorders and of the "IQ-discrepancy" criterion, an issue highlighted in the Consensus Conference of the Italian National Institute of Health (CC-ISS, 2010; 2022). Consequently, the existence of single "core deficits" underlying each disorder has been questioned, and more recent theoretical models converge in describing their multifactorial and multidimensional nature (Pennington, 2006). This partial reconceptualization helps refine the understanding of SLD without undermining its significant clinical impact. The symptomatology of SLD can present differently across individuals and may vary over time as a function of numerous biological and environmental factors, including schooling, rehabilitative or support interventions, and family-related stimulation, all of which contribute to shaping their expression in diverse ways (Snowling & Hulme, 2012). Although there is broad consensus on their neurobiological origin, no reliable biological markers for their identification are currently available, and diagnosis continues to rely primarily on behavioral observation and psychometric assessment of reading, writing, and arithmetic abilities (APA, 2013; World Health Organization (WHO), 1992). Although SLDs are generally agreed to have a neurobiological origin, no reliable biological markers currently exist for diagnosis. Identification continues to rely primarily on behavioral observation and standardized assessment of reading, writing, and arithmetic skills (CC-ISS, 2022).

According to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) diagnostic criteria for SLDs are as follows:

A. Persistent difficulties in learning and using academic skills, evident for at least six months despite targeted interventions, including at least one of the following:

- Inaccurate or slow and effortful word reading.
- Poor comprehension of the meaning of reading material.
- Difficulties with spelling.
- Impairments in written expression.
- Difficulties with number sense, numerical operations, or calculation.
- Problems with mathematical reasoning.

B. Academic skills are substantially below expectations for chronological age, interfering significantly with academic, occupational, or daily activities, as confirmed by standardized tests and comprehensive clinical evaluation. For individuals aged 17 or older, documented history of disabling learning difficulties may substitute for standardized assessment.

C. Learning difficulties begin during school years but may not manifest fully until academic demands exceed the individual's limited capacities.

D. Learning difficulties are not better explained by intellectual disability, sensory deficits, other mental or neurological disorders, psychosocial adversity, lack of instruction in the school language, or inadequate education.

Clinical practice recommends that SLD evaluation includes thorough collection of developmental history (including prenatal, perinatal, and early milestones), family history of learning disorders, academic history across subjects, previous neuropsychological assessments, access to specialized services, interventions, and perceived functional difficulties in daily life (CC-ISS, 2022).

This approach raises conceptual challenges, since the definition of a critical threshold for deficient performance (e.g., –2 standard deviations or the 5th percentile) is ultimately a conventional choice, albeit widely applied in clinical and healthcare practice (Fluss et al., 2009). In essence, it reflects the need to discretize a continuous variable in order to group individuals into dichotomous diagnostic categories (pathological vs. non-pathological). Despite significant progress in neuroimaging and genetic research, a sufficiently clear and coherent understanding of the pathogenic mechanisms underlying these disorders is still lacking, particularly regarding the interplay between genetic bases, epigenetic factors, neuroanatomical anomalies, and dysfunctions in cognitive processes that produce observable deficits in reading, writing, and arithmetic (Gialluisi et al., 2021). Furthermore, beyond the recognized neurobiological underpinnings, it must be acknowledged that learning to read, write, and calculate are culturally mediated activities, inevitably shaped by environmental determinants. This complex interplay of biological and environmental factors makes the expression of SLD even more variable across cultures and orthographic systems, demanding greater effort from clinicians to interpret how these elements interact in shaping the overall clinical presentation and to design effective intervention plans (Ziegler & Goswami, 2005). The clinical expression of SLD is also strongly influenced by the timing of assessment. Learning disorders should not be conceptualized as static or fixed conditions, but rather as developmental difficulties that evolve over time through the interaction of cognitive and linguistic processes with the practice of instrumental learning. A large body of literature emphasizes the importance of studying developmental trajectories in SLD, instead of relying exclusively on cross-sectional analyses (Lyytinen et al., 2015). From a diagnostic perspective as well, it is crucial to situate SLD within the broader developmental history of the child, rather than limiting the evaluation to a purely taxonomic framework. The prevalence of SLD in the population is deeply affected by the phenotypic and pathogenic complexity described above and, as with all neurodevelopmental disorders, shows wide variability depending on definitional criteria, the

age at assessment, and the orthographic characteristics of the language. DSM-5 reports prevalence rates ranging between 5% and 15% (APA, 2013). In Italy, no national database on neuropsychiatric disorders exists; the only official source is school certifications issued under Law 170, which in the 2018/19 academic year identified SLD in 4.9% of the school population, with marked variability across school levels (3.1% in primary school) and geographic regions (Ministero dell'Istruzione, 2020). Following the recommendations of the previous Consensus Conference on SLD, a national epidemiological study was conducted between 2008 and 2013, involving 9,964 children aged 8–10 years equally distributed across the country's regions. The prevalence of specific reading disorders in this cohort was found to be 3.5% (95% CI: 3.2–3.9%), yet only 1.3% of the sample had previously received an official SLD diagnosis (Tretti et al., 2014). It is important to highlight that SLD often exerts persistent and disabling effects on the lives of affected individuals, leading, for example, to premature school dropout and occupational difficulties. Furthermore, when not promptly identified, they may trigger a cascade of secondary difficulties in the psychopathological domain—ranging from low self-esteem and learned helplessness to more severe forms of anxiety and depression—as well as behavioral issues, particularly when SLD co-occurs with ADHD (Cornoldi & Tressoldi, 2014). These effects can substantially compromise personal functioning and social adaptation, especially when the disorder is present in its more severe forms.

3.2. Planning abilities in Specific Learning Disorders (SLD)

Although there is considerable heterogeneity in the cognitive profiles of children with Specific Learning Disorders (SLD), and not all of them show executive function (EF) difficulties, including planning, a substantial body of research highlights vulnerabilities in this domain within clinical populations. The relationship between EFs and poor academic achievement is well documented (Mulder & Cragg, 2014), and recent studies have emphasized their predictive role not only in school performance (Cortés Pascual et al., 2019) but also in emotional and motivational variables (Quílez-

Robres et al., 2021). Within this framework, SLD represents a particularly relevant condition in which to investigate EF, as these disorders are strongly intertwined with academic achievement. Children with SLD often display deficits in central executive functioning (Landerl et al., 2004; Pickering, 2006), especially in working memory (WM) (Mammarella et al., 2013, 2018; Moll et al., 2016; Peng & Fuchs, 2016). Verbal and visuospatial WM have been linked to the early acquisition of reading and math skills (Passolunghi et al., 2008; Peng et al., 2018b), with verbal WM becoming more implicated in reading performance and comprehension in later stages (Peng et al., 2018a), and visuospatial WM being associated with more complex mathematical achievement (Giofrè et al., 2014; Caviola et al., 2020). Evidence regarding inhibition deficits in children with reading and math disabilities remains mixed, likely due to differences in paradigms used (De Weerdt et al., 2013), whereas meta-analyses suggest that shifting is substantially associated with both reading and math performance (Yeniad et al., 2013). Within the EF framework, planning abilities deserve particular attention. Planning, defined as the capacity to set goals, select strategies, and monitor one's performance (Lezak et al., 2004), plays a critical role in educational contexts. Several studies have reported planning impairments in children with SLD, which manifest in disorganized approaches to problem-solving, inefficient allocation of cognitive resources, and greater difficulties in handling multi-step academic tasks (Helland & Asbjørnsen, 2000; Meltzer, 2007; Varvara et al., 2014). These weaknesses are evident in reading comprehension, written composition, and mathematical problem-solving, where effective planning is required to integrate information, select strategies, and sustain goal-directed behavior (Montague, 2008). Research employing the Tower of London (TOL) task has provided further evidence of such impairments. Children with dyslexia and dyscalculia have been found to perform worse than typically developing peers in terms of the number of attempts required to reach the correct solution (Condor, Anderson & Saling, 1995) and the time needed to complete the task (Reiter et al., 2005). Similar findings confirm planning difficulties in children with dyscalculia (Mirmehdi et al.,

2009; McLean & Hitch, 1999; Khosrorad et al., 2015). However, contradictory evidence also exists: several studies found no significant differences between children with dyslexia and typically developing peers in overall planning scores, planning time, or execution time (Lima et al., 2011; Marzocchi et al., 2008; Brosnan et al., 2002). Likewise, Moura, Simoes, and Pereira (2014) reported no significant differences in planning between children with mathematical learning disorders and controls. These discrepancies may be explained either by methodological issues, such as variations in test administration and scoring, or by the heterogeneity of SLD manifestations. The role of planning becomes even more evident when considering comorbidity. Studies have highlighted an additive effect of cognitive deficits when multiple disorders co-occur (Seidman et al., 1997, 2001; Willcutt et al., 2013; Horowitz-Kraus, 2015). For instance, children with both SLD and ADHD display more pronounced EF impairments, including planning and organizational difficulties, than children with ADHD alone (Seidman et al., 2001; Mattison & Mayes, 2012). This suggests that planning deficits may serve as a transdiagnostic marker of vulnerability, contributing to the additive effect observed in comorbid profiles and supporting the view that EF dimensions exist on a continuum across neurodevelopmental disorders. Overall, these findings suggest that planning, as part of the broader EF construct, is a key factor in understanding the variability of academic outcomes in children with SLD. Its assessment is therefore crucial not only for diagnostic purposes but also for designing tailored interventions that address learning weaknesses while enhancing metacognitive awareness and self-regulated learning. In this perspective, the implementation of effective intervention strategies may help reduce learning disparities between typically developing children and those with SLD, with a positive impact on their academic performance (Diamond & Lee, 2011; Akyurek & Bumin, 2019).

3.3. Fluid intelligence in Specific Learning Disorders

As extensively discussed in this dissertation, intelligence represents one of the fundamental constructs in psychology and is frequently assessed in both research and clinical practice (Gottfredson, 1997a).

I). In Europe, among the most commonly used tools to evaluate intellectual abilities are the WISC, the WAIS, and the Raven Progressive Matrices, ranking first, second, and fourth in frequency of use, respectively (Evers et al., 2012). This widespread application highlights the importance of intelligence, which has been shown to predict significant outcomes in educational, occupational, and everyday life contexts (e.g., Deary et al., 2007; Gottfredson, 1997b; Schmidt & Hunter, 2004).

Recent research has increasingly focused on fluid intelligence (Gf), working memory, and processing speed due to their central role in learning (Swanson, 1994). According to Cattell's investment theory, fluid intelligence serves as a foundational precursor for the development of crystallized intelligence and academic skills (Cattell, 1971, 1987), providing the cognitive resources necessary for problem-solving and learning (Dehn, 2017). Deficits in fluid intelligence, especially when associated with neurodevelopmental disorders, may contribute to learning difficulties (Mano et al., 2018).

Children with Specific Learning Disorders (SLD), despite having normal overall IQ, often exhibit heterogeneous cognitive profiles, particularly showing weaknesses in working memory and processing speed. The WISC-IV (Wechsler, 2003), the most widely used instrument for assessing cognitive functioning in developmental age, measures four indices: Verbal Comprehension (VCI), Perceptual Reasoning (PRI), Working Memory (WMI), and Processing Speed (PSI). In children with SLD, scores are typically higher on VCI and PRI and lower on WMI and PSI (Giofrè & Cornoldi, 2015), suggesting that their intelligence may be "organized" differently compared to typically developing peers, especially with regard to working memory and processing speed. Tasks assessing processing speed (e.g., Coding, Symbol Search) and working memory (e.g., Digit Span, Letter-Number Sequencing) require attentional control, which is frequently compromised in SLD (Giofrè & Cornoldi, 2015).

The WISC-IV indices can be further grouped into the General Ability Index (GAI; VCI + PRI) and the Cognitive Proficiency Index (CPI; WMI + PSI) (Prifitera et al., 2008). In children with SLD, GAI

scores are typically higher than CPI scores, reflecting that the abilities composing the GAI are more strongly associated with fluid intelligence, while CPI components are less related to the general factor g. This uneven preservation of cognitive domains explains why children with dyslexia or dyscalculia struggle with tasks involving letters or numbers, whereas tasks involving abstract symbols may be less challenging (Giofrè et al., 2019). Such patterns suggest that the general factor g may be relatively weak in SLD due to differential domain preservation.

Fluid intelligence plays a critical role in academic skills, influencing them directly or indirectly through mediating factors such as phonological awareness and rapid automatized naming (RAN) (Mano et al., 2018). Reading, for example, can be conceptualized as a complex problem-solving task gradually mastered by children. Fluid intelligence contributes indirectly via language functions (vocabulary, word knowledge) and directly via phonological processing. Weaknesses in fluid intelligence may therefore partially explain difficulties in acquiring literacy skills, consistent with Cattell's investment theory (Mano et al., 2018). Strong fluid intelligence may enable children to compensate for cognitive weaknesses, developing alternative strategies for reading, writing, and arithmetic, whereas limited fluid intelligence can increase the risk of academic failure (Giofrè & Cornoldi, 2015; Mano et al., 2018).

Non-verbal assessments provide additional insight into fluid intelligence in children with SLD. Studies using Raven Progressive Matrices (RCPM) have shown that children with moderate SLD often perform comparably or even better than typically developing peers, despite more frequent inadequate identification errors (IC-type) and fewer visuo-perceptual errors (WP-type) (Belelli et al., 2010). These findings highlight that non-verbal measures may offer a clearer evaluation of cognitive potential in populations with verbal vulnerabilities. Patterns of errors evolve with age: for instance, in older children (10–12 years), some errors (e.g., R-type errors) become less frequent in SLD while remaining high in other clinical groups, emphasizing the differential cognitive profiles of these

children.

Overall, these findings underscore the importance of assessing fluid intelligence comprehensively in children with SLD. Fluid intelligence is not only an indicator of general cognitive potential but also a predictor of children's ability to cope with learning difficulties and benefit from targeted educational interventions.

Chapter 2: Empirical studies

Study 1

PsycAssist: A Web-Based Artificial Intelligence System Designed for Adaptive Neuropsychological Assessment and Training

de Chiusole, D., Spinoso, M., Anselmi, P., Bacherini, A., Balboni, G., Mazzoni, N., Brancaccio, A., Epifania, O. M., Orsoni, M., Giovagnoli, S., Garofalo, S., Benassi, M., Robusto, E., Stefanutti, L., & Pierluigi, I. (2024). PsycAssist: A Web-Based Artificial Intelligence System Designed for Adaptive Neuropsychological Assessment and Training. *Brain Sciences*, *14*(2), 122. https://doi.org/10.3390/brainsci14020122

Abstract

Assessing executive functions in individuals with disorders or clinical conditions can be challenging, as they may lack the abilities needed for conventional test formats. The use of more personalized test versions, such as adaptive assessments, might be helpful in evaluating individuals with specific needs. This paper introduces PsycAssist, a web-based artificial intelligence system designed for neuropsychological adaptive assessment and training. PsycAssist is a highly flexible and scalable system based on procedural knowledge space theory and may be used potentially with many types of tests. We present the architecture and adaptive assessment engine of PsycAssist and the two currently available tests: AdapToL, an adaptive version of the Tower of London-like test to assess planning skills, and MatriKS, a Raven-like test to evaluate fluid intelligence. Finally, we describe the results of an investigation of the usability of AdapToL and MatriKS: the evaluators perceived these tools as appropriate and well-suited for their intended purposes, and the test-takers perceived the assessment as a positive experience. To sum up, PsycAssist represents an innovative and promising tool to tailor evaluation and training to the specific characteristics of the individual, useful for clinical practice.

Keywords: adaptive assessment; PsycAssist; planning skills; fluid intelligence; knowledge space theory

1. Introduction

The prototype of a web-based artificial intelligence (AI) system designed for neuropsychological adaptive assessment and training, named PsycAssist (psychological assistant), is presented. The system was designed to overcome some limitations that traditional paper- and-pencil tests and some existing computerized tests have. In fact, the system introduces several noteworthy features compared to existing ones: (a) it presents potential applicability in clinical settings, offering both quantitative and qualitative automatic feedback; (b) it utilizes adaptive assessment algorithms that mitigate the learning effect, address monotony for skilled examinees, and minimize frustration for examinees within clinical populations; (c) it incorporates rigorous and innovative methodological approaches; and (d) its scalability allows for potential enrichment by incorporating additional tests, enhancing its versatility and the scope of the assessments. This is in line with the attention that neuropsychology researchers have recently given to the need for integration between modern psychometric theories and technological advances in neuropsychological assessment (see, e.g., Germine et al., 2019; Marcopulos & Lojek, 2019).

PsycAssist was developed and applied within an Italian project funded by the Ministry of Research and University. This project aimed at the development of computerized tools for the adaptive and personalized assessment of executive functions (EFs) and fluid intelligence (FI).

"Executive functions" is an umbrella term that identifies multiple interrelated neurocognitive processes mainly controlled by the prefrontal cortex, such as cognitive flexibility (or set shifting), working memory, planning, problem-solving, inhibition, and interference control (Diamond, 2013). EFs are regarded as high-level mental abilities that regulate lower-level processes (Craig et al., 2016; Cristofori et al., 2019), enabling individuals to plan and organize thoughts in a goal-directed way, to understand complex or abstract concepts, to evaluate and make decisions, and to repress inappropriate behavior (Craig et al., 2016; Cristofori et al., 2019; Jurado & Rosselli, 2007). For these reasons, EFs

contribute to an individual's adaptation to the environment, everyday living, and academic and occupational successes (Jurado & Rosselli, 2007). FI represents the ability to think logically, process new information, learn, and solve problems in novel situations (Cattell, 1963, 1987; Nisbett et al., 2012), other than perceiving relations among elements (Jensen, 1974). Deficiencies in EFs and FI occur frequently in neurodevelopmental disorders (e.g., attention-deficit/hyperactivity disorder, autism, learning disabilities), after brain damage, in psychiatric disorders, and in relation to aging (e.g., (Craig et al., 2016; Carlin et al., 2000; Xiang et al., 2021). For these populations, the accurate assessment of FI and EFs is extremely relevant since it provides valuable information to plan personalized interventions. However, assessment among these populations might be difficult due to the combination of the length and the need for extended administration time of traditional tests, along with the low attention capabilities of people with the aforementioned conditions. This might lead to assessments that are poorly reliable, marginally accurate, and inefficient. Therefore, the availability of tools that allow for personalized and adaptive assessments would be crucial in order to plan early and tailor made rehabilitation interventions.

Several tests and questionnaires already exist for the assessment of both EFs and FI. Some examples of the assessment of EFs are the Tower of London test (Shallice, 1962), attentive matrices (Spinnler, 1987), and the Wisconsin card sorting test (Grant, 1948). FI assessment examples include Raven's progressive matrices (Raven 1938; 1962). For some of them, a computerized version exists that allows obtaining immediate feedback on the results. Still, it does not overcome an important inefficiency issue, which is that individuals should answer all the test items regardless of their characteristics. These tests have followed a rigorous validation process with statistical techniques belonging to the classical test theory (see, e.g., Novick, 1965) or item response theory (Mohai et al., 2022). In both cases, limited efforts were made to create adaptive versions of the tests. Falmagne & Doignon (2011) introduced an adaptive battery assessing several cognitive functions, but it did not include a measure

for planning ability. In contrast, Mohai and colleagues (2022) developed the Tower of London Adaptive Test (ToLA) to assess the abilities of individuals with neurodevelopmental disorders efficiently and accurately. However, this version has not undergone clinical studies despite its design. For a complete and in-depth description of these studies, see Section 2.1.

An alternative and more recent methodological approach for building assessment tools is knowledge space theory (KST, see, e.g., Falmagne & Doignon, 2011). It is a mathematical theory for the efficient assessment of individual knowledge aimed at personalized learning. One of the most important fields of application of KST is the adaptive assessment of knowledge. Adaptive assessment tests allow the evaluation process to be personalized on the basis of the responses an examinee gave to previous questions. The advantages that come from adaptive assessments are many and concern both the examinees and the examiner. Indeed, the number of questions asked decreases, as well as the effort required from each examinee. The effect is that the "quality" of each answer increases, allowing the examiner to collect more reliable data, and, thus, leading to more accurate feedback. Moreover, most EFs and FI standard tests have a so-called "learning effect" problem (Culbertson, 1998). This describes a situation in which an individual learns strategies for solving test items during the filling of the test. The consequence is that the assessment is contaminated by the learning skills of the individual other than the variable that the particular test measures.

This unwanted effect can be reduced a lot with adaptive administration of the test since a minimum number of questions is administered to individuals. Furthermore, a KST-based assessment generates an output that encompasses both quantitative and qualitative insights. Quantitative feedback facilitates a comparison between the examinee's performance and their reference population. Meanwhile, qualitative feedback is valuable for pinpointing the strengths and weaknesses of the examinee, aiding in the identification of areas that warrant further exploration. This dual nature of output enhances the depth and comprehensiveness of the assessment process.

Almost all the applications of KST were carried out in the area of knowledge assessment. Nevertheless, recent extensions of the theory have shown that it can be successfully applied in fields like psychological assessment (see, e.g., (Spoto et al., 2010) and the assessment of human problem-solving and planning (Stefanutti, 2019). This last extension is called procedural KST (PKST; Stefanutti, 2019) and provides deterministic and probabilistic models for partially ordering individuals based on their performances in problem-solving tasks. The deterministic models account for both the accuracy of the responses and the sequence of actions made by the problem solver. Probabilistic models consist of latent Markov models that are used as the basis of adaptive assessment algorithms.

PsycAssist is an innovative web-based system featuring a highly flexible and scalable adaptive assessment engine based on KST and PKST algorithms. Its adaptability enables the efficient handling of increasing amounts of data, users, and transactions without compromising performance. The system's scalability allows it to dynamically adapt and expand in response to growing demands and workloads. Section 3 explains how PsycAssist serves as a comprehensive solution for a diverse range of psychological and neuropsychological assessments, hosting and managing various test types to meet a wide spectrum of assessment needs.

The paper is organized as follows. The state of the art of a specific EF (i.e., planning) and of FI assessment is briefly described in Section 2.1, and those of knowledge space theory and PKST are provided in Section 2.2. Section 3 introduces the architecture of PsycAssist, with particular attention placed on its adaptive assessment algorithms. The two web apps available in the system, named "Adap-Tol" and "MatriKS", are described in Section 4. The former is an adaptive version of the Tower of London-like test for the assessment of planning skills. The latter is an adaptive version of a Raven-like test for the assessment of FI. In Section 5, results on the usability of AdapToL and MatriKS are discussed, from the point of view of both the respondent and the evaluator. Section 6

concludes the argumentation.

2. Background

2.1 State of the Art on the Assessment of Planning Skills and Fluid Intelligence

EFs are high-level mental abilities that regulate lower-level processes (Craig et al., 2016; Cristofori et al., 2019). They allow individuals to plan and organize thoughts in a goal-directed manner, to understand complex or abstract concepts, to evaluate and make decisions, and to suppress inappropriate behaviors (Craig et al., 2016; Cristofori et al., 2019; Jurado & Rosselli, 2007). For these reasons, EFs contribute to individual adaptations to the environment, daily life, and academic and professional successes (Jurado & Rosselli, 2007).

Among EFs, planning plays a central role. It is regarded as the ability to identify and organize the sequence of steps required to achieve a goal, typically without a predetermined path (Injoque-Ricle et al., 2014; Lezak, 2004). Traditional planning tasks, such as tower tests, like the Tower of London (ToL; Shallice, 1982) and Tower of Hanoi (Simon, 1975), are commonly used in neuropsychological assessments. These tasks involve moving objects from one position to another with specific rules and constraints (Willcutt, 2005). Therefore, they require individuals to visualize the necessary course of action before manipulating the materials (Riccio et al., 2004). Successful completion of these tasks involves the ability to consider the overall situation, define sub-goals, and generate a sequence of moves to achieve them.

Tower tasks have been widely used in research on the planning skills of individuals with typical development (Injoque-Ricle et al., 2014;Phillip et al., 2021) or clinical conditions, such as neurodevelopmental disorders, focal brain lesions, frontal lobe dementia, Parkinson's and Huntington's diseases, and psychiatric disorders (Craig et al., 2016; Carlin et al., 2000; Xiang et al., 2021; Owen et al., 1990; Watkins et al., 2000). Among tower tests, ToL is likely the most used with individuals of different chronological ages and conditions.

Despite the popularity of the ToL task in cognitive and clinical studies of EFs and some attempts at standardization (see, e.g., (Culbertson et al., 1998; Anderson et al., 1996; Krikorian et al., 1994; Schnirman et al., 1998) the literature shows a notable lack of uniformity in the procedure across studies (38). Several variants of the ToL have been introduced and used to examine planning abilities in different populations (Fancello et al., 2006; Unterrainer et al., 2005). Although some of them are close to the original version, others have implemented major changes concerning, for example, the administration procedure, outcome measures, the number of items, and the material itself, so that the comparison of results across studies is challenging (Anderson & Anderson, 1996; Krikorian et al., 1994; Berg & Bird, 2002; McKinlay & McLellan, 2011).

A growing interest has been placed on the computerized versions of these tests. One widely used computerized version of the ToL is the Stockings of Cambridge (SOC) task, a test of the Cambridge neuropsychological test automated battery (CANTAB; Robbins et al., 1994). Despite the advantages of computerized ToL tasks (e.g., ease of presentation, ease of use, and accurate data acquisition), they involve critical differences in the task itself, both physically and conceptually. Moreover, computerized versions proposed so far do not solve the "inefficiency issue" that can significantly affect specific clinical populations. This issue concerns the fact that individuals are asked to answer all the test questions. Conversely, adaptive assessment allows for the personalization of administration by using stimuli that dynamically adapt to an individual's appropriate challenge level. They offer the potential to shorten test times, improve test accuracy, and reduce the effect of irrelevant variance caused by the frustration of having to answer items that are too difficult (Mills & Stocking, 1996).

Only a few attempts have been made to develop adaptive tools for the assessment of EFs, particularly planning tasks. For example, in a recent study, a novel adaptive battery for the assessment of EFs, composed of eight tasks, was tested with individuals in middle childhood to measure working

memory, context monitoring, and interference resolution, but not planning ability (Younger et al., 2023). On the contrary, Mohai and colleagues (2022) developed the Tower of London Adaptive Test (ToLA) to efficiently and accurately assess planning in individuals with neurodevelopmental disorders. However, the stimuli presented are apparently different compared to Shallice's one. To develop the test, the authors first created a precisely calibrated item bank using item response theory and then set a suitable algorithm capable of estimating the participant's skill level during the test. Unfortunately, the version was only designed but was not subjected to clinical studies.

Concerning FI, its assessment generally occurs through reasoning ability tests, which are considered more culture-fair and less affected by differences in learning experiences, test familiarity, or sociocultural status (Balboni et al., 2010; Happé et al., 2013). One of the most well-known tests of this type is the Raven's Matrices test (Raven, 1938). This test requires the individual to identify the piece that best completes a visual-spatial matrix from a series of given options. Currently, the Raven's Matrices test is available in three forms, each composed of stimuli that differ in number and type, and presented in a booklet format. Colored progressive matrices (CPM; Raven, 1962) consist of 36 colored items and are intended for kindergarten to middle school-aged children and the elderly; standard progressive matrices (SPM; Raven, 1938) are composed of 60 black and white items, designed for individuals from 6 years old and above. Advanced progressive matrices (APM; Raven, 1962) are composed of 48 more complex items, allowing for the differentiation of individuals with high cognitive abilities, and are intended for adolescents and adults. Within each form, stimuli are placed in order of increasing difficulty and are organized into different series, assessing specific competencies. For instance, CPM are organized into three series, which evaluate the ability to identify similarities (based on shape, dimension, direction, quantity, orientation, figure/background, and density criteria), the ability to detect symmetry, and conceptual thinking skills (i.e., detection of abstract relations according to 'operant-deductive' logic and their retention in working memory),

respectively (Villardita, 1985). Efforts to limit administration times have led to the testing of reduced versions with adults (Arthur & Day, 1994; Bilker et al., 2012; Bors & Stokes, 1998; Caffarra et al., 2003; Van der Elst et al., 2013; Wytek et al., 1984), high school and university students (Chiesti et al., 2012), and individuals of developmental age (Kramer et al., 2023; Langener et al., 2022), as well as versions with a fixed administration time (Hamel et al., 2006).

Other attempts have been made to automate the administration of Raven's Matrices, mainly to standardize administration and scoring procedures. The initial efforts involved automating the SPM using slides and a projector, where the exact same stimuli of the original SPM were used. The selected response to each item was indicated by pressing a button. The time and accuracy of each response were registered (Calvert & Waterfall, 1982; Gilberstadt et al., 1976; Watts et al., 1982). Later on, SPM (Arce-Ferrer & Guzman, 2009; Kubinger et al., 1991; Williams & McCord, 2006) and CPM (Rock et al., 1982) were administered using computers to school-aged children, high school students, psychiatric patients, federal employees, and undergraduate students. These automated versions differed only in their presentation format from the original ones, hence providing no advantages regarding the test length and the duration of attention focus required to the test-takers.

Attempts to create computerized adaptive versions of Raven's Matrices date back to the 1980s and 1990s (Watts et al., 1982; Styles et al., 1993), and continue today, since computerized adaptive testing provides more accurate FI estimations using a reduced number of administered items (see, e.g., Odeh & Obaidat, 2013). For instance (Styles & Andrich, 1993), using Rasch models, created a unique test with SPM and APM items for individuals aged 10 to 16. This version consisted of five practice items from SPM followed by the first test item that was relatively easy for that person's chronological age. If the individual responded incorrectly to this first item, then the second one—in order of difficulty—was presented. Conversely, the algorithm skipped to the third item, and so on, ending the test when the person answered five items incorrectly in a row.

Using item response theory, the Hansen Research Services Matrix Adaptive Test (HRS- MAT) was developed specifically for individuals with autism spectrum disorder but is also usable with the general population (Hansen et al., 2019). It measures nonverbal intelligence using tasks similar to Raven's Matrices. Similar to the tool developed by Styles & Andrich (1993), the instrument's adaptive algorithm selects items to administer that are appropriate for the participant's age (children or adults) and ability level, taking no more than 15 min to complete. However, this HRS-MAT is only available for research purposes and it is not used in clinical settings. To conclude, given the importance of planning skills and FI and the evaluation of their manifestations in both clinical and non-clinical populations, having precise and efficient tools for their assessment is highly relevant. Furthermore, a helpful resource would be a tool with clear standard administration and scoring procedures that is applicable to different

types of populations and capable of adapting to the characteristics of the individuals.

2.2 Knowledge Space Theory, Procedural Knowledge Space Theory, and Adaptive Assessment Algorithms

Knowledge space theory (KST, see, e.g., (Falmagne & Doignon, 2011; Doignon & Falmagne, 1985; 1999) is a mathematical theory developed for the efficient assessment of knowledge and personalized learning. One of the most prominent features of KST is that no attempt is made to compute linearly ordered numerical scores for sorting individuals. Rather, the goal of the assessment is to describe "what an individual masters", which is their knowledge state , and "what they are ready to learn", which is named the outer fringe of their knowledge state. Formally, a knowledge state is the set $K \subseteq Q$ of items that an individual masters in a particular knowledge domain Q, and the outer fringe of a knowledge state K is the set of all problems in $Q \setminus K$, such that $K \cup \{q\}$ is a knowledge state. From a pedagogical point of view, the outer fringe is the set of all problems that the individual can learn individually, allowing for the increase of their own knowledge state.

The collection of all knowledge states existing in a population of individuals is the so called knowledge structure (Q, K). The knowledge structure reflects the precedence relations (e.g., prerequisites) existing among the problems in Q. Thus, K is a subset of the power set 2Q (i.e., the collection of all the subsets of Q) of problems. Let $\prec$ be a precedence relation defined on Q. A subset of Q is a knowledge state whenever, for every pair q, r $\in$ Q of problems, if r $\prec$ q and q $\in$ K then r $\in$ K. From a practical point of view, this means that r is a prerequisite of q. As a consequence, the state containing r but not q does not exist.

To provide an example, consider the knowledge domain Q = {a, b, c, d, e} composed of five problems, and assume a precedence relation $\prec$ for which a $\prec$ b $\prec$ d and a $\prec$ c $\prec$ e. The following knowledge structure contains all those subsets of Q that are consistent with the precedence relation $\prec$:

K = {∅, {a}, {a, b}, {a, c}, {a, b, c}, {a, b, d}, {a, c, e}, {a, b, c, d}, {a, b, c, e}, Q}.

Figure 1 shows the Hasse diagram of K. In the diagram, each node represents a knowledge state, and an arrow from a left node to a right node indicates that the state on the left is a subset of the state to the right.

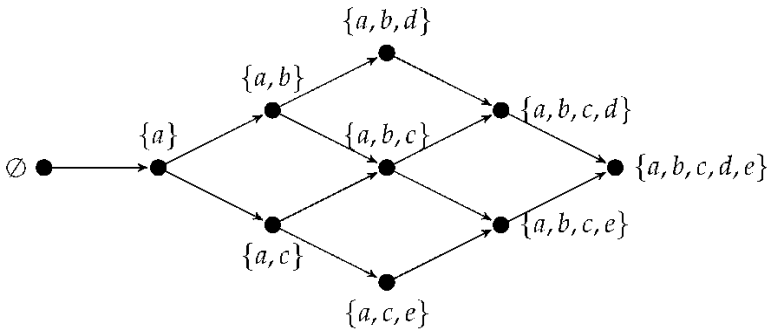


Figure 1. The knowledge structure corresponding to the partial order a $\prec$ b $\prec$ d, a $\prec$ c $\prec$ e.

It can be noted that the empty set (i.e., the situation in which an individual cannot solve any problem) and the full set Q (i.e., the situation in which an individual knows everything) are knowledge states. Moreover, among the 25 = 32 different subsets that can be formed with the 5 problems in Q, only 10 belong to K. Given that a $\prec$ b $\prec$ d, all states containing b also contain a, and all states containing d

also contain both a, b. A similar reasoning is applied to the other precedence relations.

Assuming that the knowledge state of an individual is K = {a, b}, its outer fringe K = {c, d} gives information about what the individual is ready to learn (i.e., problems c and d). Thus, the outer fringe is useful for personalizing learning and reaching the educational needs of each individual.

Knowledge structures can be empirically validated via probabilistic models. Several probabilistic models are available in the literature (Anselmi et al., 2012, 2016, 2017; de Chiusole et al., 2013, 2015; de Chiusole & Stefanutti, 2013; Heller et al., 2015, 2016; Robusto et al., 2010; Stefanutti et al., 2011) that can be used, depending on the particular application context. Almost all of them are generalizations of the so called basic local independence model (BLIM; (Falmagne & Doignon, 1988). BLIM is a latent class model, where the latent classes are knowledge states. What is observed is a response pattern (i.e., the subset R of items correctly solved) for each individual. The prediction of the knowledge state behind a response pattern is provided by estimating two parameters for each item $q \in Q$, representing the careless error (i.e., $q \in K$, $q \notin R$) and the lucky guess (i.e., $q \notin K$, $q \in R$) probabilities. Throughout the years, BLIM was studied in depth from both theoretical (Heller, 2017; Spoto et al., 2012, 2013; Stefanutti et al., 2018) and practical/application (de Chiusole et al., 2013; Heller & Wickelmaier, 2013; Stefanutti & Robusto, 2009) perspectives.

It is worth noting that the original development of KST is useful when dichotomous (correct/wrong) responses are available, and it is specifically meant for the knowledge assessment field of study. In recent years, some extensions of KST were proposed in order to generalize the theory to different fields of applications and different response type formats. For example, the polytomous KST (Heller, 2021; Stefanutti, Anselmi, et al., 2020; Stefanutti, de Chiusole, Anselmi, et al., 2020; Stefanutti, de Chiusole, Gondan, et al., 2020; Stefanutti et al., 2023) allows for nominal, ordered, and partially ordered polytomous response scales.

An extension of KST that is relevant to the project presented in this paper is termed the procedural

knowledge space theory (PKST; (Stefanutti, 2019). PKST is useful for the assessment of human problem-solving skills. To this aim, the novelty of the approach consists of modeling the whole solution process made by a problem solver, rather than considering the problem's accuracy only. PKST is based on both problem space theory (Newell & Simon, 1972) and KST. In fact, it provides a formal representation of the notion of problem space and makes an algorithm for deriving a knowledge structure starting from a problem space available. The main concepts at the basis of the theory are briefly introduced below.

In PKST, all the problems in Q are pairs (s, g), and the objective is to transform an initial configuration s into a target configuration g. In a problem space, a solution path for a problem (s, g) is a pair $s\pi$ where π is the sequence of observable operations required to transform s into g. It is worth mentioning that a problem (s, g) may have multiple alternative solution paths. Moreover, the solution paths that solve the problems in Q are partially ordered. In particular, a solution path $s\pi$ is a subpath of another solution path $t\sigma$, denoted as $s\pi \sqsubseteq t\sigma$ if there are two sequences of operations α, and β, such that σ is the sequence $\alpha\pi\beta$, and the sequence α transforms t into s. The psychological interpretation of the relationship between a solution path and its subpaths is that "if an individual knows a solution path, then they also know all of its subpaths". This interpretation is named the problem–subproblem assumption. The straightforward implication is that an individual who knows how to solve problem (s, t) through the solution path $s\pi$ knows how to solve all the problems that are solved by any subpaths of $s\pi$. A whole description of the deterministic foundation of PKST can be found in (Stefanutti, 2019) and in (Stefanutti et al., 2021).

Just like KST, the deterministic models of PKST need to be empirically validated. The latent class Markov solution process model (MSPM; Stefanutti et al., 2019) can be used for this purpose. The MSPM is useful for making predictions on both the observable solution process and the unobservable knowledge state (the latent class) on which the solution process is built. Under the MSPM, the

probability of moving from one configuration of the problem to another one might take one of two alternative forms. One of them is conditional to the belonging of a problem to the knowledge state (i.e., the problem solver is able to plan at least one of the problem's solution paths) and the other one is conditional to the 'non-belonging' of a problem to the knowledge state (i.e., the problem solver is not able to plan any of the problem's solution paths).

One of the main fields of application of both KST and PKST is the adaptive assessment of knowledge. In this respect, precedence relations in KST and the problem–sub-problem assumptions in PKST play a central role. In brief, an adaptive assessment consists of administering items selected on previously observed responses. This has the effect of personalizing the assessment and minimizing the number of questions. In the KST framework, an example is as follows: If $a \prec b$ and an individual provides a correct answer for problem b, then it is plausible to assume (in a situation without noise) that she is able to solve item a.

In the PKST framework, an example is as follows: if problem (s, g) is solved by the path $s\pi$, and $t\sigma \sqsubseteq s\pi$ (i.e., $t\sigma$ is a subpath of $s\pi$), then an individual who is able to solve (s, g) is able to solve all those problems that are solved by $t\sigma$. Precedent relations, on the one hand, and problem–subproblem assumptions, on the other hand, allow for the prediction of the answer.

In the KST framework, several types of adaptive algorithms have been proposed (see, e.g., Falmagne & Doignon, 1988;de Chiusole et al., 2020; Degreef et al., 1986; Dowling & Hockemeyer, 2001; Falmagne & Doignon, 1988; Hockemeyer, 2002). The most used one is the continuous Markov procedure proposed by (Falmagne & Doignon, 1988). The basic idea is that a likelihood function over the knowledge structure expresses the plausibility of the states. At each step of the assessment process, the likelihood function is updated according to the observed correct or wrong answer via a Bayesian rule that takes into account the careless error and lucky guess parameters of the items. The assessment terminates when the mass of the likelihood is concentrated on a single state, which is

regarded as the uncovered state of the individual.

A similar algorithm was proposed by (Brancaccio, 2022) in the PKST framework. The main difference between the KST-based and the PKST-based algorithms is that the latter updates the likelihood on the basis of the whole observed solution process instead of using only the correct/wrong answer. In (Brancaccio et al., 2022), it was shown that the PKST-based algorithm outperforms the KST-based algorithm in terms of efficiency and accuracy since the former one employs more information from the observed data than the latter one.

3. The Architecture of PsycAssist

In this section, a description of the information technology architecture of the system is given. PsycAssist is a highly flexible and scalable web-based system whose adaptive assessment engine is based on KST and PKST algorithms. The flexibility and scalability of the system allow it to handle a growing amount of data, users, or transactions without compromising performance. Mostly, the scalability allows the system to adapt and expand as demand and workload increase. Indeed, as explained in the following section, PsycAssist can host and manage many different types of tests for psychological and neuropsychological assessments, making it a comprehensive solution for a wide range of assessment needs. Moreover, the web-based design of PsycAssist facilitates global accessibility for evaluators, allowing them to utilize the system from any location worldwide and on any device, simply connecting to the web page at https://psycassist.fisppa.unipd.it/research_project/). The only requirement is the availability of an internet connection.

The system is designed to cater to a diverse range of professionals who are licensed or authorized to administer psychological tests, including psychologists and neuropsychologists. Users have their own credentials to access the system. Once logged, a user can add a list of test-takers who can be assessed with all the tests available on the system. The system collects and organizes the results of any assessment made by any test-takers. In fact, at the end of any assessment, an automatically generated

report describes in detail the results of the assessment by using several types of information, both quantitative and qualitative, which might be of support for clinicians.

It is worth noting that the system is inherently an explainable AI system in two different aspects. In the former, it is a rule-based system in the user interaction since the rules are conditional statements that establish how the system should behave or make decisions in various situations, depending on the answers provided by the user. Section 3.1 describes the rule-based algorithm at the core of the system. The other aspect that makes PsycAssist an AI-based system concerns the so-called knowledge structures. Each test implemented in the system is referred to a specific knowledge structure that represents the organization and the connections of a particular knowledge (or cognitive) domain. These knowledge structures are sometimes constructed by using machine learning methods, such as advanced clustering algorithms (see, e.g., de Chiusole, 2017; 2020).

The subsequent sections offer detailed descriptions of both the assessment engine and the assessment module. Additionally, comprehensive information regarding the compatible types of devices for system usage and the security measures implemented within the system are provided.

3.1. The System Engine: A PKST-Based Adaptive Algorithm

The assessment engine of PsycAssist is a generalized version of the PKST-based adaptive algorithm proposed by (Brancaccio et al., 2022). Figure 2 shows the flowchart of the algorithm.

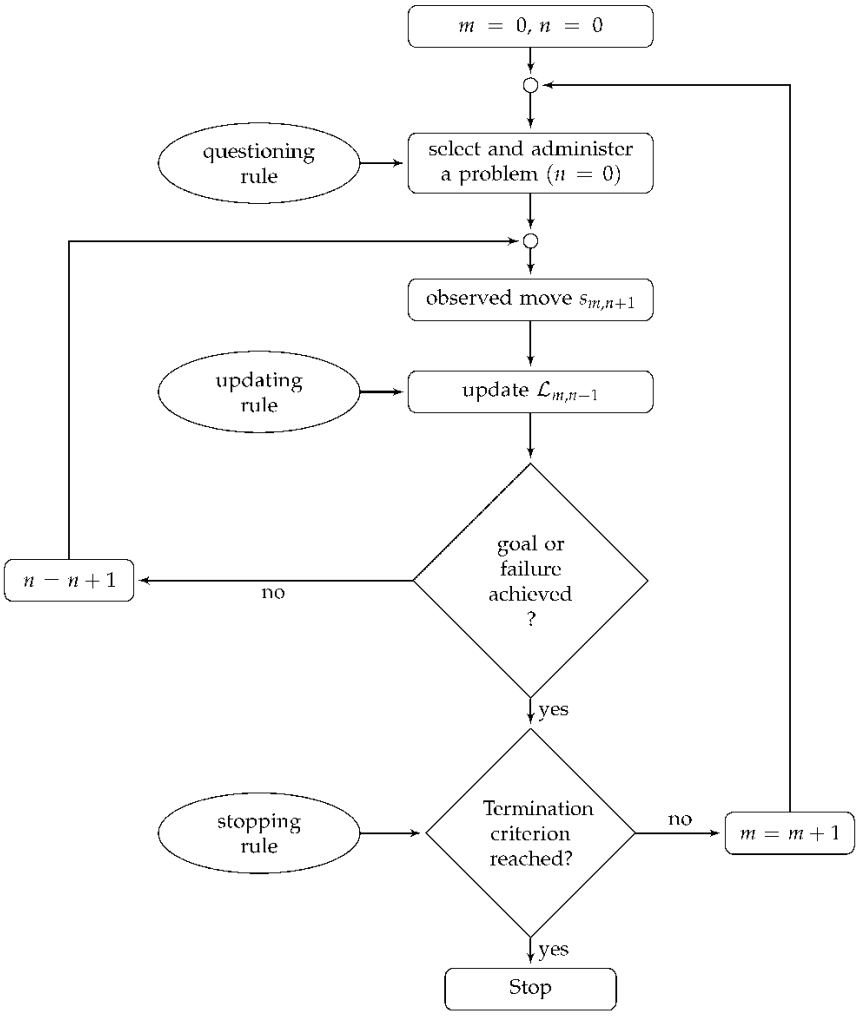


Figure 2. Algorithm of the adaptive probabilistic procedure implemented in PsycAssist. See text for more details.

The algorithm consists of two nested loops. The outer loop handles the administration of a new problem, whereas the inner one handles and records the steps of the solution for the problem. It is worth noting that, in situations in which a single-step answer is required for the problem, the algorithm does not enter the inner loop. In this last case, the PKST-based algorithm is equal to the KST-based one.

The assessment begins with the initial values m=0 and n=0 representing the iterations throughout the outer and inner loops, respectively. At the same starting step, the likelihood $\mathscr{L}$0,0 is set to be a uniform distribution among the knowledge states. Starting from $\mathscr{L}$0,0, the assessment is carried out iteratively.

At each iteration of the external loop, a questioning rule, an updating rule, and a stopping rule are applied.

At each step, m, the most informative problem, $q \in Q$, is selected according to a questioning rule named half-split. Let $\mathcal{K}q$ be the collection of all states containing problem q; the problem selected by the half-split is the one that minimizes the difference.

$$|\mathcal{L}_{m,n}(\mathcal{K}_q) - \frac{1}{2}|.$$

If there are several problems that minimize that quantity, then one of the problems is selected at random. In other words, the half-split rule selects the problem for which the two probabilities of the individual responding correctly or incorrectly are as similar as possible. From a psychological point of view, this amounts to selecting a problem that is neither too difficult for the individual (and, thus, demotivating) nor too easy (and, thus, boring).

The participant is presented with the initial configuration of the problem sm,n, selected by the questioning rule, and the likelihood $\mathscr{L}m,n+1$ is updated on the basis of the participant's move from the configuration sm,n of the problem to the configuration $sm,n+1$. The updating rule computes $\mathscr{L}m,n+1$ from the $\mathscr{L}m,n$ by

where $P(sm,n+1|sm,n,q,K)$ is the conditional probability of moving from sm,n to $sm,n+1$, given the knowledge state $K \in \mathcal{K}$ and problem $q \in Q$. Practically, the Bayesian updating rule updates the likelihood of all knowledge states in the knowledge structure as follows.

If the individual provides a correct answer to the problem, the likelihood of all states containing it is increased, and the likelihood of all states not containing it is decreased. If the individual fails the problem, the likelihood of all states containing it is decreased and the likelihood of all states not containing it is increased.

The procedure continues by selecting and administering problems and by updating the likelihood of

$$\mathcal{L}_{m,n+1}(K) = \frac{P(s_{m,n+1}|s_{m,n},q,K)\,\mathcal{L}_{m,n}(K)}{\sum_{K' \in \mathcal{K}} P(s_{m,n+1}|s_{m,n}q,K)\,\mathcal{L}_{m,n}(K')},$$

the knowledge states until the likelihood of one of them reaches a pre-specified termination criterion. In particular, as soon as the likelihood $\mathscr{L}_{m,n+1}(K)$ of any knowledge state $K \in \mathcal{K}$ is greater than a termination criterion $p \in (0.5,1)$, the assessment terminates. Otherwise, an additional problem is administered. Upon termination, the knowledge state whose likelihood is maximum provides an estimation of the knowledge state of the individual.

It is worth noting that (Stefanutti et al., 2021; Brancaccio et al., 2022) presented three possible explicit forms for the conditional probability in Equation (1), which are based on three different assumptions, describing different solution behaviors.

What the three assumptions have in common is that the conditional probability $P(sm_{,n+1}|sm_{,n},q,K)$ has two interpretations, depending on the belonging (or not) of problem q to the knowledge state K. Suppose that $sm_{,n+1}$ is a problem solution, q. The probability of making a move from $sm_{,n}$ to $sm_{,n+1}$, given that the individual masters the item (i.e., $q \in K$) is interpreted as a non-careless error. Conversely, the probability of making a move from $sm_{,n}$ to $sm_{,n+1}$, given that the individual does not master the item (i.e., $q \notin K$), is interpreted as a lucky guess. In this respect, these two interpretations are the same as the careless error and lucky guess parameters of the BLIM.

Instead, the three assumptions differ in the type of planning that is considered. The first assumption, named the pre-planning assumption, describes a situation in which a participant meticulously plans out the entire solution process for the problem and applies it at every step. The second assumption, named the interim-planning assumption, describes a situation in which several planning instances may take place during the solution process. Finally, the third assumption, named the mixed-planning assumption, describes the situation in which both pre-planning and interim-planning might occur. The formal characterization of such assumptions can be found in (Brancaccio et al., 2022).

The algorithm at the core of PsycAssist is versatile, allowing for personalizations based on the specific

aim and task of the test. The two tests currently available on the system were implemented by applying the pre-planning assumption.

3.2. The Assessment Module

Regardless of the particular test to be administered, the administration procedure consists of the following four phases: (1) instruction; (2) practice; (3) testing; and (4) conclusion.

The instruction phase consists of a short video tutorial to be watched by the test-taker. The tutorial includes all the information useful to complete the task characterizing a test. Specifically, a human-like avatar reads the text on the screen, describing the task, the type of responses the test-taker must provide, and the rules to be respected for accomplishing the task. Concurrently, the video demonstrates the correct execution of the task while the instructions are narrated. To ensure universal accessibility, the video tutorial was crafted by incorporating insights from the literature, drawing from various sources, such as (a) previously developed computerized tests (Salaffi et al., 2013; Zelazo, 2014), (b) self-reported measures tailored to individuals with intellectual and developmental disabilities (Lancioni et al., 2007; Walton et al., 2022), and (c) established guidelines in the field (Harniss et al., 2007), including the widely recognized Universal Design for Learning (see, e.g., https://udlguidelines.cast.org/). Key features adopted for enhanced accessibility include the use of a sans-serif font, appropriate font sizes, high-colored contrast between letters and background, spacious layouts without narrow spacing, the absence of hyphenations, the incorporation of visual markers to emphasize essential words (e.g., bold, different colors) or images (e.g., circles, arrows), and the integration of animations without flashes, sparkles, or flickers. Furthermore, the video allows for personalized volume adjustments to cater to individual preferences.

After instructions, a practice phase follows. Regardless of the specific test, the trial block comprises a limited number (three or four) of basic test items. The responses to these items do not contribute to the final individual's assessment. However, they play a crucial role in assessing whether the test-taker

masters the essential skills to complete the test. Specifically, if more than 75% of the observed responses are incorrect during the practice phase, the evaluator may opt to forego administering the test.

In the assessment phase, individual test items are presented sequentially, one at a time. The modality of responding to these items is contingent on the specific nature of the test. For instance, responses may involve selecting an option or engaging with objects by moving them from one position to another. If a test-taker is unable to provide an answer for a particular item, the evaluator can decide to skip it by pressing the designated button. Throughout the assessment phase, several variables are recorded in a database. The specific variables to be recorded are contingent on the characteristics of the particular test. However, common variables include the order of items, the responses of the test-takers, and the time spent on each item. In terms of sociodemographic information, the system can capture gender, age, nationality, and years of schooling for each test-taker. Notably, the system does not allow the registration of other personal data. A concluding message appears on the screen once the test is completed.

The conclusion phase pertains to the evaluator and consists of consulting a report that is automatically generated by the system. In fact, at the end of each administration, the system automatically generates a comprehensive report, offering valuable insights into the performance of the test-taker. The content of the report is contingent on the particulars of the administered test. Regardless of the test, the report furnishes tables and figures illustrating the frequency and percentage of correct, incorrect, and unanswered responses. Furthermore, the report includes more specific descriptions of the results, organized into distinct sections. These sections provide a nuanced understanding of the test-taker's performance, enhancing the clinicians' insights into various facets of the assessment.

3.3. Device Typologies

As highlighted in the literature, several advantages are linked to computerized tests. These advantages

concern (a) the ease of administration (less training, reduced time, accuracy in recording responses) and scoring (no effects of the evaluator's bias, automated calculation, and presentation of the findings); (b) the reduction of long-term costs; (c) the possibility of instant feedback on individual performance, which is especially useful in a clinical setting; (d) the ease of importing data into statistical software, which is useful in a research setting; (e) instantaneous recording of the performance, which can also be used to tailor the task difficulty to the individual performance level (Mead & Drasgow, 1993;Witt et al., 2013). Furthermore, due to the time-efficient method of administration, computerized tests appear ideal if an individual needs to be tested multiple times, for example, to monitor changes and effects of interventions in clinical populations (Witt et al., 2013).

Concerning the device to be used for administering a computerized test, tablets are particularly useful. Indeed, they are readily available, relatively cheap, and portable, and, compared to smartphone technology, have an adequate screen size for administering cognitive tools (McHenry et al., 2023). Touchscreen use requires less eye-motor coordination than indirect input devices (mouse/keyboard), and is, thus, particularly useful for patients with mild motor dysfunctions (Deguchi et al., 2023). Moreover, the touchscreen provides a more engaging experience and seems more intuitive even for those unfamiliar with digital devices, as well as for patients with mild cognitive impairment (Deguchi et al., 2023; Robinson et al., 2016).

For all these reasons, the usage of a tablet is suggested to administer tests via PsycAssist. However, a computer can be used as well.

3.4. Security of the Platform

The platform is equipped with the most recent "safety parameters" according to the European legislation on privacy (see, e.g., the General Data Protection Regulation, GDPR, at https://gdpr-info.eu/, and the Artificial Intelligence Act at https://artificialintelligenceact.eu/)). The precautions and strategies adopted to protect the personal data being processed include techniques such as disk

encryption, encryption of individual files, encryption of web communications via the HTTPS protocol, database, and backup encryption, and data pseudonymization to prevent the association between personal data and sensitive data; moreover, access control is enforced through the use of JSON Web Tokens for authentication and authorization operations.

4. Web Apps for the Neuropsychological Assessment

Two web apps are currently available on PsycAssist: Adap-ToL, a test for the assessment of planning skills, and MatriKS, a test for the assessment of FI. The following sections describe these two tests.

4.1. Adap-ToL: Assessment of Planning Skills

Adap-ToL is a test designed to assess the planning skills of individuals who are four years old and above (children, adolescents, and adults), with or without clinical conditions. Similar to the Tower of London test (Shallice et al., 1982) and other classic tower tests (Shallice et al., 1982; Simon et al., 1975; Delis et al., 2001; Korkman, 1998), the task is to move beads from an initial configuration to a goal configuration, using a given maximum number of moves and respecting specific rules.

In developing Adap-ToL, some precautions were adopted to maximize comparability with the traditional ToL test (Berg & Bird, 2002). First, the computerized testing material was created to be as faithful as possible to the physical one, with the same number of pegs and beads to be manipulated and, consequently, the one-to-one ratio of beads to empty spaces. Furthermore, the required task and recorded responses (correctness, planning time, execution time, total time, violations, number of moves) are the same. Regarding the rules, those that can be replicated via a digital device were maintained, e.g., (a) the highest peg can hold three beads; the central peg can hold a maximum of two beads, and the shortest peg can hold only one bead; (b) a bead cannot be moved if it has another bead on its top. At the same time, the tool shows some differences from the traditional ToL, which are described below.

Unlike the traditional ToL, in which the initial state is the same for all the items, but the goal state

changes, in Adap-ToL, the goal state is fixed for all the problems, while the initial state changes. The reason why this modification was introduced is strictly connected with the probabilistic model MSPM on which the PKST-based assessment procedure is based. In fact, the conditional probabilities used in the adaptive assessment procedure (see Section 3.1) require that the two problems share the same goal.

Moreover, in the original ToL test (Shallice, 1982) an indirect measure of the difficulty of a problem is obtained as the minimum number of moves necessary to solve it. However, recent studies (Fancello et al., 2006; McKinlay et al., 2011; Kaller et al., 2004) found that other factors affect the difficulty of a problem. Some of them are the number of alternative solutions for the problem, the initial configuration of the beads on the pegs (named "start hierarchy"), and the final configuration (named "goal hierarchy"). PKST goes beyond the notion of the minimum number of moves, considering partial orders of the item's difficulty instead of a total one.

The partial order of item difficulty built for Adap-ToL takes into account several features of the items. Table 1 lists the 35 items included in Adap-ToL, specifying, for each of them (Columns 1 and 5), the start hierarchy (Columns 2 and 6), the number of moves required for reaching the goal state with the minimum number of moves (Columns 3 and 7), and the number of alternative solutions (Columns 4 and 8). Concerning the start hierarchy, a "1" specifies the so-called ambiguous configuration, a "2" specifies the so-called partial ambiguous configuration, and a "3" specifies the so-called unambiguous configuration. The reader can refer to (114) for a comprehensive description of these configurations.

Table 1. The 35 problems of Adap-ToL with their characteristics. Columns 1 and 5 display the number of the problem, Columns 2 and 6 display the hierarchy of the initial state (see text for more details), Columns 3 and 7 display the minimum number of moves for solving the problem. Finally, Columns 4 and 8 display the number of alternative solution paths.

Problem	Hierarchy	# Moves	# Paths	Problem	Hierarchy	# Moves	# Paths
1	2	1	1	19	2	5	1
2	2	1	1	20	1	5	2
3	2	2	1	21	3	6	1
4	1	2	1	22	2	6	1
5	1	2	1	23	3	6	2
6	2	3	1	24	2	6	2
7	2	3	1	25	3	6	1
8	2	3	1	26	2	6	1
9	2	3	1	27	2	6	2
10	2	3	1	28	2	7	1
11	2	3	1	29	1	7	3
12	1	4	1	30	1	7	1
13	2	4	1	31	3	7	2
14	3	4	2	32	2	7	2
15	2	4	1	33	2	8	2
16	2	4	1	34	2	8	3
17	2	5	1	35	2	8	3
18	2	5	2				

Adap-ToL can be administered in three different versions. Adap-ToL 4–8 is used with children between the ages of 4 and 8 years old and includes Items 1 to 20 of Table 1. Adap-ToL 9–13 is used with children of age between 9 and 13 years old and includes Items 1 to 27. Finally, Adap-ToL 14+ is used with individuals who are more than 14 years old and includes all 35 items listed in Table 1.

In Adap-ToL, the task is to move from an initial state to a goal state using a given maximum number of moves and respecting specific rules. When a tablet is used, orienting the screen vertically is recommended. Two images are presented to the test-taker, one in the upper part (i.e., the goal state) and one in the lower part (i.e., the initial configuration) of the screen. Figure 3 shows the screenshot of one of the Adap-ToL problems.

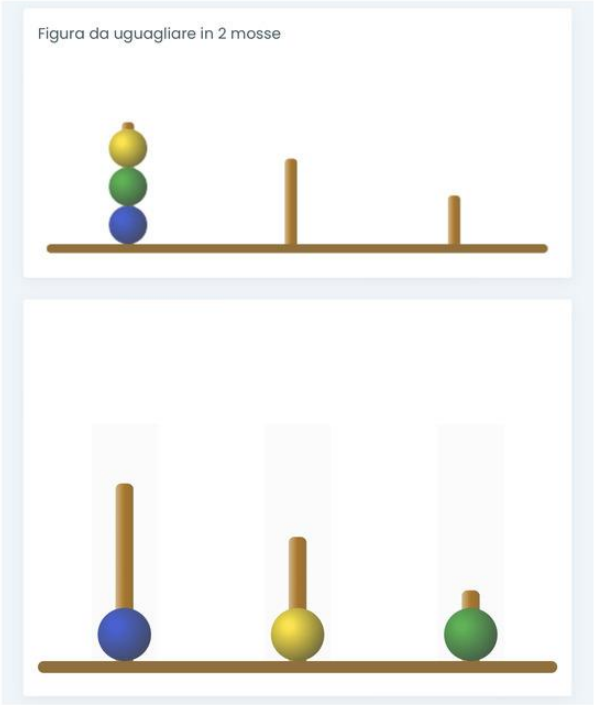


Figure 3. One of the items included in Adap-ToL. The stimulus is composed of two images. The one in the upper part is the goal state and the one in the lower part is the initial configuration. The initial configuration is used by the test-taker for making the moves. The text in Italian “Figura da uguagliare in due mosse” on the left top corner of the figure means “Figure to be matched in two moves”.

The initial configuration is used by the test-taker for making the moves. Each of the two images represents a base from which three pegs of different heights rise, into which three colored beads are inserted. Differently from the original ToL, the beads are blue, green, and yellow, with yellow used instead of red to make the instrument accessible to individuals with redgreen dyschromatopsia. The number of moves needed to solve the problem successfully appears at the upper left of the screen.

The required movement in Adap-ToL is a double tap on the monitor. With a first touch, the individual selects the bead to move; with a second touch, the individual selects the peg where they want to insert the bead. The problem is correctly solved when the goal state is reproduced using no more than the indicated number of moves.

The system incorporates three types of feedback pop-ups that may appear for the test-taker. Feedback

is provided in cases where a rule is violated, when an excessive number of moves are made to complete the task, or when the task is correctly executed. Each feedback is presented within a colored box (red, orange, or green, respectively) to enhance the intuitiveness of the feedback, especially for children who cannot read yet or who may face challenges in reading. In such instances, the evaluator reads the message aloud. Unlike the traditional ToL, in which three attempts are required for each item, in Adap-ToL, only one attempt is allowed. Throughout the execution of Adap-ToL, the system records various parameters, including (a) the accuracy of the item, which reflects whether the task is completed correctly or incorrectly; (b) the pre-planning time, which is the duration from the presentation of the stimulus to the initial touch of a bead, indicating the time taken for pre-planning before problem-solving begins; (c) the execution time, which is the duration from the first touch of the bead to the completion of the problem, measuring the time dedicated to executing the task; (d) the total time, which is the cumulative time encompassing both pre-planning and execution phases; (e) the number and type of violations that capture instances where test-takers deviate from established rules, with information specifying the type and frequency of violations; (f) the number and sequence of moves, which detail the quantity and order of movements made by the test-takers during the task. These recorded parameters offer a comprehensive assessment of the test-taker's performance during the execution of Adap-ToL.

4.2. MatriKS: The Assessment of Fluid Intelligence

MatriKS is a test designed to assess FI, specifically targeting reasoning abilities. It was developed considering the principles of Raven's Matrices. Like Adap-ToL, the idea behind the creation of MatriKS was to have a unique tool usable with different populations: children, adolescents, and adults, with or without clinical conditions.

As in traditional tests that use Raven-like matrices as stimuli, the MatriKS task aims to identify the piece that best completes a visuospatial matrix among the proposed options. Response options in

MatriKS can be five or eight, depending on the matrix type. The incorrect options, referred to as "distractors", were intentionally crafted to replicate the same types of errors observed in traditional Raven's Matrices (Kunda et al., 2016). These errors are classified into the following four categories:

1) a difference error occurs when the selected distractor exhibits a qualitative difference from the other choices, making it visually distinctive and likely to "pop" among the response options. Examples include a completely blank option or an option featuring extraneous shapes not present elsewhere in the item;
2) a repetition error arises when the chosen distractor replicates a matrix tile adjacent to the blank;
3) a wrong principle error occurs when the distractor selected is a copy or composition of elements occurring in various tiles of the matrix, combined according to an incorrect rule;
4) an incomplete correlate error occurs when the chosen distractor is almost correct but deviates by only one rule.

Categorizing erroneous responses using these classifications facilitates a comparison with the examination of specific error patterns observed in traditional Raven's Matrices, spanning diverse age groups and clinical conditions (Babcock, 2002; Gunn & Jarrold, 2004; Van Herwegen et al., 2011). This approach allows for a meaningful analysis of error trends and patterns, contributing to a broader understanding of cognitive performance across different contexts and populations.

The matrices included in MatriKS, along with their corresponding response lists, were generated using a novel R package designed for the automated creation of Raven-like matrices (MatriKS; Brancaccio et al., 2023). In brief, MatriKS generates stimuli based on the specified parameters, including (a) the dimension of the matrix (e.g., 2×2 or 3×3); (b) the objects to be used (e.g., square, circle, etc.); (c) the rule that guides the manipulation (e.g., change in shape, orientation, size, etc.); (d) the direction of the manipulation (e.g., vertical, horizontal, or diagonal). For a comprehensive

understanding of the package and its functionalities, readers can refer to the documentation accompanying the MatriKS package.

Table 2 lists the features of the stimuli included in MatriKS 4–11. Concerning the stimuli included in MatriKS 12+, all of them are in black and white, only three have dimension 2×2 with five response options, whereas all of the others have dimension 3×3 with eight response options. The two versions share 11 stimuli. Stimuli of MatriKS 4–11 require visual–perceptual, elaboration of the general configuration, reasoning, and elementary inference skills. In addition to these skills, stimuli of MatriKS 12+ require advanced skills of reasoning and logic inference.

# Matrix	Color	Dim.	# Options	# Matrix	Color	Dim	# Options
1	yes	Mono	5	21	yes	2 × 2	5
2	yes	Mono	5	22	yes	2 × 2	5
3	yes	Mono	5	23	no	2 × 2	5
4	yes	Mono	5	24	no	2 × 2	5
5	yes	Mono	5	25	no	2 × 2	5
6	yes	2 × 2	5	26	yes	3 × 3	8
7	yes	2 × 2	5	27	yes	3 × 3	8
8	yes	2 × 2	5	28	yes	3 × 3	8
9	yes	2 × 2	5	29	yes	3 × 3	8
10	yes	2 × 2	5	30	yes	3 × 3	8
11	yes	2 × 2	5	31	yes	3 × 3	8
12	yes	2 × 2	5	32	yes	3 × 3	8
13	no	2 × 2	5	33	no	3 × 3	8
14	no	2 × 2	5	34	no	3 × 3	8
15	no	2 × 2	5	35	no	3 × 3	8
16	no	2 × 2	5	36	no	3 × 3	8
17	no	2 × 2	5	37	no	3 × 3	8
18	no	2 × 2	5	38	no	3 × 3	8
19	no	2 × 2	5	39	no	3 × 3	8
20	yes	2 × 2	5	40	no	3 × 3	8

Table 2. The 40 stimuli included in MatriKS 4–11. For each stimulus, information concerning color (columns 2 and 6), dimension (columns 3 and 7), and the number of response options (columns 4 to

8) are provided.

The stimuli of MatriKS are composed of two parts: the matrix and the response options. The matrix is presented on the left of the screen and can be of three types: monothematic matrices (named "Mono" in the table), representing a single figure from which a piece is missing; 2 × 2 matrices, series of four cells, one of which is missing; and 3 × 3 matrices, series of nine cells, one of which is missing. The missing cell is at the low-right corner of the matrix. Response options are presented on the right of the screen. To provide the answer, the test-taker selects one of the options. Figure 4 shows one of the stimuli of MatriKS.

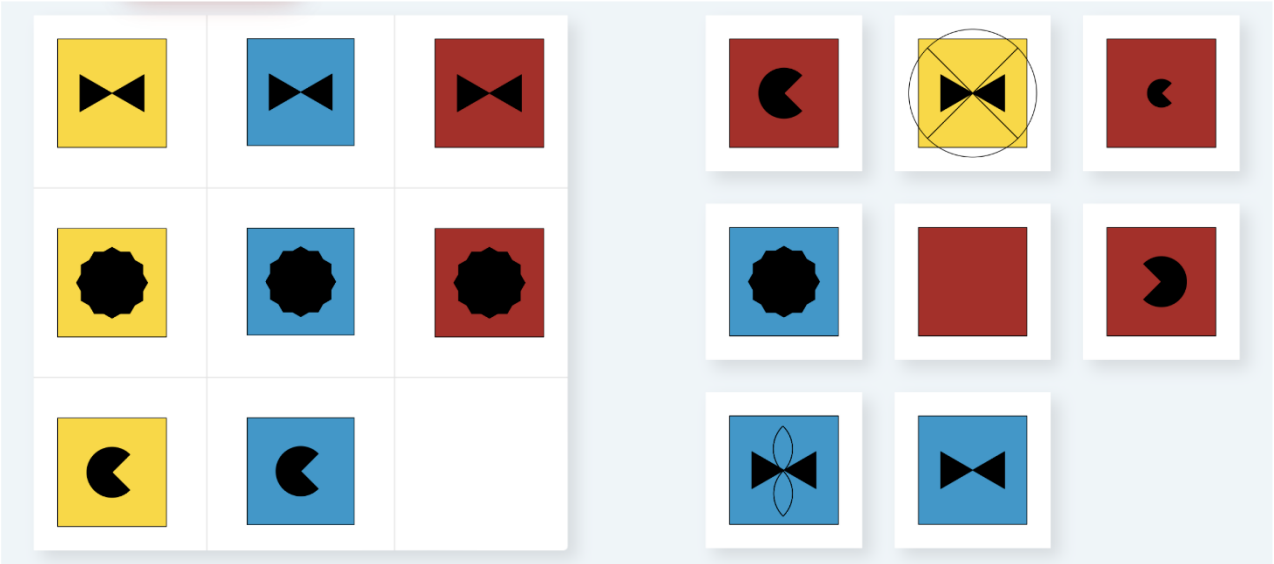


Figure 4. One of the stimuli of MatriKS 4–11, which is a colored 3 × 3 matrix with eight response options. The matrix is presented on the left and the response options are presented on the right. The missing cell is at the low-right corner of the matrix.

When a tablet is used, MatriKS should be administered by orienting the screen horizontally for a better view of the stimuli. The MatriKS trial block consists of two monothematic matrices and two 2×2 matrices. In case of an error in at least one of the two monothematic matrices, the evaluator shows the video tutorial again and presents the same trial items one more time before moving on to the test items.

During the task, the system automatically records correct, incorrect, or non-provided answers. In case of an incorrect answer, the type of response given (i.e., the distractor) is also recorded. Moreover, the time taken to solve each problem, the total time to complete the test, and the average time per item are recorded.

5. Usability Study

The perceived "ease of use" and attitude toward new technologies are particularly relevant for the effectiveness and implementation of new digital neuropsychological tests. Understanding and evaluating these aspects is crucial to ensure that technology not only improves assessments but is also perceived as facilitating and effectively used by clinicians and patients. Specifically, acceptability refers to the degree to which a user finds a computer system suitable, agreeable, and satisfactory. It is strictly connected to the concept of usability, which is the extent to which a product can be used by specified users to achieve specific goals with effectiveness, efficiency, and satisfaction in a given context of use (ISO, 1998).

To investigate the usability and acceptability of the new digital assessment web apps Adap-ToL and MatriKS, we developed the Usability and Attitude Scale (UAS), which is a revised version of the system usability scale (SUS; Brooke, 1996) and the technology acceptance model (TAM; Venkatesh, & Bala, 2008) questionnaires.

The system usability scale (SUS) is a widely used questionnaire for measuring usability (Brooke, 1996), operatively defined as the subjective perception of interaction with a system. The SUS is specifically designed to generate a comprehensive single measure of perceived usability. It is known as a "quick" survey scale that allows practitioners to easily and efficiently access the usability of novel technologies. The original SUS includes a mix of positive and negative items and it was structured to control for acquiescence bias and to identify respondents who did not pay attention to the statements. However, several studies highlighted the fact that the inclusion of both positive and

negative items can be problematic (Barnette, 2000; Stewart & Frye, 2004). For instance, it can lower internal reliability (Stewart & Frye, 2004), distort the factor structure (Pilotte & Gable, 1990; Schmitt & Stuits, 1985; Schriesheim & Hill, 1981), and increase interpretation problems with cross-cultural use (Wong et al., 2003). Moreover, mixed items may lead to difficulty in switching users' response behaviors and it can increase the cognitive load (Kortum et al., 2021). This could be particularly true for the clinical and younger population, whose EFs may be impaired or still in development.

Notably, some authors suggested that questionnaires for usability assessment should avoid the inclusion of a mixture of positive and negative items and that researchers who do not specifically need to use the standard SUS should consider using the positive version to reduce the likelihood of response or scoring errors (e.g., (Lewis, 2018). A version of the SUS that includes only positively worded items has been created (Sauro & Lewis, 2011). Evidence about the positive version of the SUS reliability, validity, and sensitivity has been provided (Lewis et al., 2015). Moreover, more recently, the equivalence in terms of confidence to measure the subjective usability of two versions of the SUS (mixed and only positive) has been shown, together with the fact that the scores generated from the two SUS versions are comparable (e.g., (Kortum et al., 2021). The positive SUS appears, hence, as a valuable alternative to the standard SUS and offers the advantages of reduced cognitive load and shifting ability.

6. Materials and Methods

6.1 The Usability and Attitude Scale (UAS)

UAS is a new questionnaire based on SUS (Brooke, 1996; Sauro & Lewis, 2011) and TAM (Venkatesh & Bala, 2008; Davis, 1989) that has been structured to achieve brevity and ease of comprehension. It is an 8-item questionnaire, adapted in the Italian language, currently undergoing validation. To date, there are no instruments specifically targeted for developmental age that encompass both the features of usability and acceptability. Given the breadth of our sample (ranging

from 4 to 70 years of age), it was deemed beneficial to develop a scale suitable even for younger participants (e.g., brief, employing easy language, incorporating emoticons). Therefore, we decided to design the UAS to be administered to all test-takers, regardless of their chronological age. On the other hand, the original versions of SUS and TAM were administered to the evaluators to assess their perceived usability and acceptability of Adap-ToL and MatriKS, given that they were adults capable of adequately responding to the original versions of the questionnaires.

Concerning the UAS, four items assess the usability based on SUS items and four items assess the acceptability based on TAM items. As mentioned above, UAS was designed primarily for developmental age, especially for children ranging from 4 to 13 years of age. Nevertheless, for the study's purposes, it was administered to all the study participants (i.e., ranging from 4 to 70 years of age).

The SUS authors advocated for a flexible application of the SUS items according to the context of interest. Thus, only 4 out of 10 SUS items were considered. Specifically, referring to the positive SUS version of (Sauro & Lewis, 2018), only positive assertions were included, as they are more comprehensible for the targeted audience of children (i.e., Items 1, 2, 3, and 7). Moreover, for maximizing the comprehension of the response valence, both for younger and older participants, in the proposed UAS questionnaire a 5-point Likert scale was used as the response scale, where "1" corresponded to "strongly disagree" and "5" corresponded to "strongly agree". The verbal response options were paired with colored emoticon images. This visual aid was designed to enhance comprehension and engagement in the rating process.

A similar approach was followed in selecting items related to acceptability. The original TAM version consists of 12 items, with 6 evaluating perceived usefulness (PU) and 6 assessing perceived ease of use (PEU). PU refers to the degree to which a person believes that technology will enhance job performance, while PEU is defined as the extent to which a person believes that using technology

will be effortless (Lewis et al., 2015). However, MatriKS and Adap-ToL are digital assessment tools, designed to enhance the job performance of clinicians and experimenters but they do not aim to improve that of the tested participants. For this reason, the UAS questionnaire intended for the test-takers exclusively included items related to PEU, while the usability and acceptability questionnaire filled out by the experimenters included both PU and PEU items. Within the PEU items, the UAS questionnaire selectively incorporated just four items. Specifically, Items 5 and 6 pertained to the perceived ease of use for both oneself and others; Item 7 focused on the perceived pleasantness/fun, while Item 8 addressed intentionality.

Concerning the questionnaire administered to the evaluators, it was the original version of the SUS and TAM questionnaires composed of 22 items overall. The responses to the SUS items were provided by using a 5-point Likert scale, while responses to TAM items were provided by using a 7-point Likert scale.

6.2 Participants and Procedure

The UAS intended for the test-takers, was administered to a sample of 1239 participants aged 4 to 70 (47% males and 53% females, age mean =19.43, age standard deviation SD=16.86) of the general population. The descriptive statistics of the test-takers sample are reported in Table A1 of Appendix A (See de Chiusole et al., 2024).

The study was approved by the ethical committee for the psychological research of the University of Padua. Before participating in the study, participants, or their parents if minors, received detailed information about the study purposes and procedure, providing a signed informed consent, in accordance with the Declaration of Helsinki recommendations.

Data were collected in different regions of the North, Centre, and South of Italy (i.e., Abruzzo, Basilicata, Calabria, Emilia Romagna, Liguria, Lombardia, Marche, Puglia, Toscana, Umbria, Veneto). Participants aged 4 to 18 were recruited in schools (i.e., from kindergarten to high school),

while adults were recruited using a snowball sampling procedure. The school principals were contacted beforehand to inform them about the aim of the project and to ask about their willingness to be involved in the project. The informed agreements were given to the parents of the children in each of the interested classrooms. Only the children whose parents signed the informed consent were included in the study. The exclusion criteria were the presence of motor, visual, and auditory impairments that prevented the test-takers from completing the task.

Participants completed different versions of MatriKS and Adap-ToL according to their chronological age, that is 4–11 or 12+ years old for MatriKS, and 4–8, 9–13, or 14+ years old for Adap-ToL. Two UAS were filled out by all participants, one after the completion of MatriKS and another one after the completion of Adap-ToL.

To assess the usability and acceptability of Adap-ToL and MatriKS as perceived by the evaluators, the original version of the SUS and TAM questionnaires were filled out by 12 evaluators overall, aged 23 to 37 (100% females, age mean =26.69, age SD=3.92). Evaluators were asked to complete the SUS and TAM twice, namely the first day and the last day of the data collection (pre-test and post-test hereafter, respectively). Evaluators were interns, doctoral students, and research fellows in the field of psychology. All study participants and evaluators completed the questionnaires using a tablet.

6.3 Data Analysis

$$x_{ij} = \frac{\mu_{ij} - \alpha_i}{\beta_i},$$

To determine the UAS score, we adopted two different procedures for usability and acceptability. In usability, we adopted the method described in (Lewis, 2018) and the curved grading scale of (Sauro & Lewis, 2012, 2016), both reported below. Concerning acceptability, we applied the procedure outlined by the original authors of TAM (Lewis, 2019) to provide acceptability scores consistent with usability metrics. The Sauro–Lewis curved grading scale (CGS), as proposed by (Sauro & Lewis,

2012, 2016), offers a useful framework for interpreting these overall scores. CGS utilizes an 11-grade scale, ranging from F (an SUS total score of less than 51.6) to A+ (a SUS total score of less than 84.1). The average value on this scale is 68, corresponding to a grade of C. Table A2 in the Appendix A of the published paper shows the numerical score range corresponding to each grade. However, this method does not allow for the establishment of targets for other and more specific experience attributes. Lewis (2018) proposed item-level benchmarks estimated through regression analysis to consider specific items from the SUS without adversely affecting the overall score. The authors estimated regression coefficients for each item in the original SUS version to establish relationships between the items and their corresponding SUS total scores. Thus, in accordance with (Lewis, 2018), we employed item-level benchmarks to assess the different weight of usability items within the UAS for both Adap-ToL and MatriKS and for each different version of the tests, by using the following formula:

where μij is the mean of item i for the test j, and αi and βi are, respectively, the intercept and the regression coefficient estimated for item i, according to (Lewis, 2018).

The application of this procedure allows for determining the estimated value for each SUS item integrated into the usability and attitude scale (UAS), by utilizing the item-level regression coefficients proposed by (Lewis, 2018). Then, we compared the obtained score with the regression coefficients provided by the authors, in order to identify the most relevant items for evaluating the usability of our novel assessment tool. To provide acceptability scores consistent with usability metrics, first, the initial 1 to 5 responses were converted into scores ranging from 0 to 4. Then, the average obtained on the four items was multiplied by 100/4, where 4 is the maximum value of the scale 0 to 4. The outcome of this calculation yields an acceptability average score that spans from 0 (suggesting very poor perceived usability) to 100 (indicating excellent perceived usability). Moreover, the item-level acceptability score was computed. To the best of our knowledge, there are

no studies that estimated the item-level benchmark for the TAM, as (Lewis, 2018) did for the SUS. However, to be consistent with the usability results provided here, and for descriptive purposes, we calculated the single-item mean values of the acceptability as perceived by the participants of the sample. The scores were averaged separately for Adap-ToL and MatriKS, in every version of the two tests, both for the users and experimenters. Then, to obtain 0–100 scores that were comparable with the usability ones, first, the initial 1 to 5 responses were converted into scores ranging from 0 to 4, then these values were multiplied by 100/4 for the users who completed the UAS and by 100/6 for the examiners who completed the original TAM (4 and 6 were, respectively, the maximum value of the response scale). Finally, we assigned the respective A–F grades (Sauro & Lewis, 2012, 2016).

7. Results

7.1 Perspective of the Test-Takers

Table 3 displays the results concerning the system's perceived usability of both Adap-ToL and MatriKS, separately for each version of the instruments. To facilitate comprehension, the results are described referring to the CGS solution proposed by (Sauro & Lewis, 2012, 2016).

Version	Item 1	Item 2	Item 3	Item 4	U-Score
Adap-ToL 4–8	97.50	80.75	72.46	78.24	82.24
Adap-ToL 9–13	94.09	72.02	75.71	82.78	81.15
Adap-ToL 14 +	92.27	66.00	71.79	81.13	77.80
MatriKS 4–11	93.41	76.11	72.73	79.69	80.43
MatriKS 12 +	75.76	51.19	58.04	65.56	62.64

Table 3. Results concerning the users' usability of Adap-ToL (Rows 2 to 4) and MatriKS (Rows 5 and 6). Column 1 lists the version of the test. Columns 2 to 5 list the item-level usability scores. Columns 6 lists the average usability (U-score) scores.

The results suggest that both Adap-ToL and MatriKS are well-received by users, regardless of the test's version. In particular, the perceived usability of Adap-ToL ranged from "more than excellent"

(Grade A: Adap-ToL 4–8 and 9–13 versions) to “excellent” (Grade B+: Adap-ToL 14+ version). Concerning MatriKS, the perceived usability resulted in “excellent” for the version tailored to the younger population (Grade A−: MatriKS 4–11), while the usability score of the other version resulted between “sufficient” and “good” (MatriKS12+: Grade between C− and D).

The results obtained with the item-level benchmark method confirmed different weights of usability items within the UAS for both Adap-ToL and MatriKS. Specifically, in the item scores of all three versions of Adap-ToL, Item 1 (“I would gladly take the test again”, translated to Italian as: “Rifarei volentieri la prova”) obtained an A+ grade score, demonstrating consistent excellent values. In contrast, Item 2 (“The test was very easy”, in Italian: “La prova è stata molto facile”) obtained varying scores, ranging from A− in the 4–8 version to C+ in the 9–13 version, and C in 14+ version. Item 3 (“I think that anyone could take this test”, in Italian: “Penso che tutti possano fare questa prova”) obtained a C+ both in the 4–8 and in 14+ versions and a B grade in the 9–13 version. Finally, Item 4 (“It was easy to understand what to do during this test”, in Italian “E’ stato facile capire cosa fare durante la prova”) received a B+ grade in the 4–8 version and A grade in both 9–13 and 14+ versions. In the MatriKS item scores, Item 1 achieved an outstanding A+ grade in the 4–11 version and a commendable B grade in the 12+ version. Item 2 received a B grade for the 4–11 version but performed poorly, earning an F, in the 12+ version. Item 3 garnered a B− grade for the 4–11 version and a D grade for the 12+ version. Lastly, Item 4 obtained an A− grade for version 4–11, while registering a score between C− and D for version 12+.

Overall, the item score results for Adap-ToL and MatriKS indicate satisfactory usability of the instruments, especially for the population with lower school grades, as evidenced by the greatest UAS usability scores for the Adap-ToL and MatriKS versions designed for this demographic.

Concerns arise regarding the usability scores for higher school grades, as particularly highlighted by Item 2, which showed low scores in Adap-ToL 14+ and MatriKS 12+. Additionally, a notably poor

score is observed in Item 3 of MatriKS 12+. A potential explanation for the low score in Item 2 might be attributed to a misinterpretation of the sentence, which could have directed respondents' attention more toward the task itself rather than the tools. A similar consideration applies to Item 3, where the complexity of the task might have influenced respondents and guided their responses, diverting their attention away from the instruments themselves.

Generally, the lower scores observed in the higher school-grade versions of both tests may be attributed to the extended duration and increased difficulty of their items when compared to those featured in the versions designed for younger grades.

Table 4 displays the results concerning the perceived system acceptability of both Adap-ToL and MatriKS, separately for each version of the instruments.

Version	Item 5	Item 6	Item 7	Item 8	A-Score
Adap-ToL 4–8	83.70	70.26	88.93	81.92	81.20
Adap-ToL 9–13	85.85	71.81	87.98	77.68	80.83
Adap-ToL 14+	86.86	74.34	86.82	81.87	82.47
MatriKS 4–11	84.65	70.75	86.5	81.05	80.74
MatriKS 12+	83.14	71.62	84.65	79.22	79.66

Table 4. Results concerning the users' acceptability of Adap-ToL (Rows 2 to 4) and MatriKS (Rows 5 and 6). Column 1 lists the version of the test. Columns 2 to 5 list the acceptability item scores. Column 6 lists the average acceptability (A-score) scores.

In both Adap-ToL and MatriKS, the results obtained with the CGS method confirm positive ratings similar to those found for usability. Moreover, the results obtained with the item-level benchmark method show that in all three versions of Adap-ToL, Item 5 "I think that taking the test on the tablet was easy" (in Italian: "Penso che fare la prova sul tablet sia stato semplice") consistently demonstrated excellent values, from A to A+ grade for all three versions. In contrast, Item 6 "I think that it would be easy for anyone to take the test on the tablet" (in Italian: "Penso che per tutti sia facile

fare la prova con il tablet") exhibited scores in versions 4–8 corresponding to C grades and version 14+ received a B grades. Item 7 "I felt good taking the test on the tablet" (in Italian: "Sono stato bene a fare la prova sul tablet") attained an A+ grade for all the tests. Finally, Item 8 "I would like to always use the tablet for this test" (in Italian "Vorrei usare sempre il tablet per fare questa prova") received an A grade for versions 4–8, a B+ grade for versions 9–13, and A grade for 14+ versions.

In the item scores of both versions of MatriKS, Item 5 achieved an outstanding A+ grade in the 4–11 versions and an A grade in the 12+ versions. Item 6 received a C grade for the 4–11 versions and a C+ grade, in the 12+ versions. Item 7 showed an A+ grade for both the 4–11 and 12+ versions. Lastly, Item 8 obtained an A grade for the 4–11 versions, while registering a score of A− for the 12+ versions.

7.2. Perspective of the Evaluator

Table 5 shows the results of the perceived usability scores provided by the evaluators for Adap-ToL and MatriKS.

Original Item	Adap-ToL		MatriKS	
	Pre-Test	Post-Test	Pre-Test	Post-Test
1. I think that I would like to use the system frequently	100.00	100.00	98.25	100.00
2*. I found the system unnecessarily complex	92.06	87.95	80.35	83.69
3. I thought the system was easy to use	92.38	93.97	83.61	90.63
4*. I think that I would need the support of a technical person to be able to use the system	92.28	75.61	89.33	86.37
5. I found the various functions in this system were well integrated	97.88	99.41	88.28	100.00
6*. I thought there was too much inconsistency in this system	85.43	83.92	90.16	83.06
7. I would imagine that most people would learn to use this system very quickly	91.76	91.01	89.70	83.49
8*. I found the system very awkward to use	88.33	87.87	88.31	84.93
9. I felt very confident using the system	80.05	94.91	87.56	95.08
10. I needed to learn a lot of things before I could get going with this system	81.80	82.83	79.55	90.84
U-score	90.20	89.75	87.51	90.08

Table 5. Results concerning the evaluators' usability of Adap-ToL (Columns 2 and 3) and MatriKS (Columns 4 and 5). Columns 2 and 4 refer to the first day of the administration (pre-test), whereas

Columns 3 and 5 refer to the last day of the administration (post-test). Rows 2 to 11 display the usability item scores. Row 12 displays the average usability scores (U-score). The scores of the items marked with a star are inverted.

The results suggest that the evaluators perceived excellent usability of both Adap-ToL and MatriKS. Specifically, upon examining potential differences in U-scores between the pre-test and post-test experiences of the evaluators, a decrease of −0.1% was observed for Adap-ToL and an increase of 2.85% for MatriKS. The slight decrease for Adap-ToL in the post-test was associated with Item 4, which states, “I think I would need the support of someone who already knows how to use it” (translated from Italian: “Penso che avrei bisogno del supporto di una persona che sia già in grado di utilizzarlo”). This item exhibited a reduction of 16.04% in the post-test compared to the pre-test, while item 9: “I felt very confident with the test during use” (translated from Italian: “Ho avuto molta confidenza con il test durante l’uso”), showed the principal improvement, demonstrating an increase of 18.77%. The increase in the post-test scores for MatriKS was primarily influenced by the improvements in Items 3 (8.40 %), 5 (11.72 %), 9 (7.91%), and 10 (12.43), whereas minor reductions were noted in Items 6 (−8.55 %), and 7(−7.44 %).

Table 6 shows the results of the acceptability perceived by the evaluators for both Adap-ToL and MatriKS.

Original Item	Adap-ToL		MatriKS	
	Pre-Test	Post-Test	Pre-Test	Post-Test
1. My interaction with the system is clear and understandable	96.43	97.62	92.86	96.43
2. Interacting with the system does not require a lot of my mental effort	67.86	80.95	83.33	84.52
3. I find the system to be easy to use	94.05	97.62	90.48	94.05
4. I find it easy to get the system to do what I want it to do	79.76	91.67	82.14	86.90
5. Computers do not scare me at all	94.05	95.24	90.48	90.48
6*. Working with a computer makes me nervous	90.48	98.81	91.67	97.62
7*. Computers make me feel uncomfortable	95.24	95.24	90.48	96.43
8. I find using the system to be enjoyable	84.52	84.52	84.52	78.57
9*. The actual process of using the system is pleasant	96.43	91.67	90.48	83.33
10. I have fun using the system	95.24	86.90	89.29	80.95
11. Assuming I had access to the system, I intend to use it	97.62	92.86	90.48	89.29
12. I plan to use the system in the next four months	92.86	57.14	94.05	53.57
A-score	90.38	89.19	89.19	86.01

Table 6. Results concerning the evaluators' acceptability of Adap-ToL (Columns 2 and 3) and MatriKS (Columns 4 and 5). Columns 2 and 4 refer to the first day of the administration (pre-test), whereas Columns 3 and 5 refer to the last day of the administration (post-test). Rows 2 to 13 display the acceptability item scores. Row 14 displays the average acceptability scores (A-score). The scores of the items marked with a star are inverted.

The acceptability perceived by the evaluators was excellent and obtained an A+ grade for both Adap-ToL and MatriKS. A decrease between the pre-test and post-test acceptability scores was observed for both Adap-ToL (−1.32%) and MatriKS (−1.32%). Nevertheless, for both Adap-ToL and MatriKS, a within-subjects ANOVA did not reveal statistically significant differences between the pre-test scores and the post-test scores.

8. Discussion and Conclusion

PsycAssist is an innovative and flexible web-based AI system designed for neuropsychological adaptive assessment and training. The adaptive assessment engine of PsycAssist draws on PKST (Stefanutti, 2019), a formal approach that furnishes deterministic and probabilistic models for partially ordering individuals based on their performances in problem-solving tasks. Moreover, the web-based design of PsycAssist allows users (e.g., psychologists and neuropsychologists) to utilize

the system from any location worldwide and on any device.

The system's adaptivity and serviceability are helpful, especially for the psychological assessment of specific clinical populations. Indeed, evaluating individuals with specific disorders or conditions can be challenging, often due to their limitations in answering conventional test formats. Adaptive assessments allow for the personalization of the administration, shortening test times, improving test accuracy, and reducing the effect of frustration (Mills & Stocking, 1996).

When introducing new digital neuropsychological assessment tools, it is crucial not only to conduct psychometric validation of the assessment tests but also to assess the perceived "ease of use" and "attitudes" toward the software by both the test-takers and the evaluators. These aspects are crucial to ensure that technology not only enhances assessments but is also perceived as a facilitative and effective tool by both evaluators and test-takers.

The paper presented here describes the general functioning of the system and the initial results regarding the usability and acceptability of the two web apps available on the system. The results confirmed the positive reception of these two tests, indicating that evaluators perceived these tools as appropriate and well-suited for their intended purposes, and that test-takers viewed the assessment as a positive experience. Despite the overall positive satisfaction, some results provided valuable insights for enhancing and optimizing the usability and acceptability of Adap-ToL and MatriKS in future applications. Furthermore, with regard to the sample of test-takers, it is important to acknowledge a limitation in the usability study. Most of the participants were recruited from schools (N=902), and only a fourth of the participants were adults (N=337). To enhance the comprehensiveness of our findings and gather more insights into perceived acceptability and usability among adults, future research should aim to include a broader representation of older participants.

Moreover, because the examiners were not healthcare providers, conducting a further study with practicing clinicians would be appropriate to verify their willingness to use these instruments in their

daily practice. Furthermore, it must be noted that the versions of MatriKS and Adap-ToL used in this usability study were preliminary versions whose validations are in progress. Despite these limitations, feedback provided by both test-takers and evaluators suggests that the new digital neuropsychological assessment tools are promising and deserve further research.

The use of the system might offer various clinical advantages that warrant attention. First, its automatic report generation and scoring alleviate the evaluator's burden, significantly reducing the likelihood of errors. This not only enhances the efficiency of the assessment process but also ensures more reliable outcomes. Secondly, the system serves as a versatile tool that complements traditional tests. It offers a detailed exploration of the areas under investigation, providing a comprehensive understanding of the individual's cognitive profile. The integration of this system with traditional assessments enhances the richness of the evaluative process. A noteworthy aspect is the motivational impact on examinees. The digital format, as opposed to traditional paper–pencil tests, tends to engage examinees more effectively. This is particularly beneficial for skilled examinees who might find traditional methods monotonous. Additionally, it minimizes frustration among examinees within clinical populations, contributing to a more positive testing experience. Lastly, the personalized and qualitative nature of the feedback generated by the system holds significant potential. This tailored feedback not only aids in understanding an individual's strengths and weaknesses but also provides valuable insights for training and treatment strategies. The ability to monitor the effectiveness of interventions and make adjustments based on personalized feedback adds a dynamic dimension to the potential clinical applications of the system.

Study 2

Psychometric Properties of the Usability and Attitude Scale (UAS) for Evaluating Digital Tools in Children and Adolescent Users

Spinoso, M., Benassi, M., Mazzoni, N., Orsoni, M., Stefanutti, L., Anselmi, P., de Chiusole, D., Bacherini, A., Pierluigi, I., Garofalo, S., Balboni, G., & Giovagnoli, S. (2025). Psychometric properties of the Usability and Attitude Scale (UAS) for evaluating digital tools in children and adolescent users [*Manuscript under review*]

Abstract

The rapid integration of digital tools in educational and clinical settings highlights the need for assessing their usability and acceptability, particularly among populations of developmental age. This study aims to validate the Usability and Attitude Scale (UAS), a new scale tailored for children aged 4 to 18, derived from integration and an adaptation of the Technology Acceptance Model and System Usability Scale. The UAS was administered to a sample of 908 participants. Results of Exploratory and Confirmatory Factor Analyses consistently supported a bifactorial model encompassing the two dimensions of usability and acceptability. Reliability was also assessed using Cronbach's α and McDonald's ω, with overall results indicating acceptable internal consistency. These findings suggested that the UAS is a valid and reliable questionnaire for evaluating digital tools among younger users, offering valuable insights for developers and educators aiming to create child-friendly technologies.

Keywords: usability, acceptability, psychometric properties, digital tools, attitude, children and adolescents, usability questionnaire, users scale.

1. Introduction

Nowadays, technological advancements are progressing at an unprecedented rate compared to previous times. Technology plays an increasing role in children's lives, shaping their learning, entertainment, and social interactions. In this context, the perceived ease of use and attitude toward new technologies are particularly important for the effectiveness and implementation of new digital tools, such as computerized neuropsychological tests (Davis, 1989; Venkatesh & Davis, 2000). Concerning the latter, it is essential to ensure that technology not only enhances assessment quality, but it is also perceived as accessible and easily usable by children (Druin, 2002). Acceptability refers to the extent to which a user finds a computer system suitable, agreeable, and satisfactory. It is closely linked to the concept of usability, which measures the extent to which specified users can use a product to achieve specific goals with effectiveness, efficiency, and satisfaction in a defined context of use (ISO, 1998). Assessing the acceptability and usability of digital technologies in children is crucial for several reasons. Children have distinct cognitive and motor skills compared to adults, which significantly influence their interaction with technology (Piaget, 1970; Diamond, 2000). Therefore, tools designed for children must be tailored to their specific needs and capabilities (Markopoulos et al., 2008; Hourcade, 2007). Additionally, children's attitudes toward technology can affect their willingness to engage with digital tools, impacting the effectiveness of educational and therapeutic interventions (Druin, 2002). Moreover, understanding the usability and acceptability of these technologies can help developers create more engaging and effective products, ultimately enhancing the overall user experience for children.

To effectively assess usability and acceptability, various methods are employed, including usability testing (e.g., the System Usability Scale, SUS, and task completion rates; Brooke, 1996; Rubin & Chisnell, 2008), direct observation (e.g., think-aloud protocols; Nielsen, 1994; van den Haak, de Jong,

& Schellens, 2003), heuristic evaluation (e.g., Nielsen's heuristics; Nielsen & Molich, 1990; Jeffries & Desurvire, 1992), cognitive walkthrough (e.g., Polson et al., 1992; Wharton et al., 1994), focus groups (e.g., contextual interviews and group discussions; Kuniavsky, 2003). Among these methods, questionnaires are one of the most widely used tools in this field (Brooke, 1996; Bangor, Kortum, & Miller, 2008).

However, a significant gap remains, as no specific questionnaires have been developed for the developmental age population. To address this need, the present study aims to create and validate a new questionnaire, the Usability and Attitude Scale (UAS), specifically tailored for children and adolescents aged 4 to 18. By focusing on this age group, the UAS seeks to enhance our understanding of how children engage with technology, ultimately contributing to the development of more effective and engaging digital tools. The UAS is an adapted version of the Technology Acceptance Model (TAM; Venkatesh & Bala, 2008) and System Usability Scale (SUS; Brooke, 1996) questionnaires. In this study, we administered the UAS to assess the usability and acceptability of new digital tools for neuropsychological evaluation, such as MatriKS, a computerized test measuring fluid intelligence (de Chiusole et al., 2024).

Introduced by Davis (1989), TAM is one of the most influential models for understanding user acceptance of technology. The model considers perceived ease of use and perceived usefulness as key determinants of technology adoption. TAM has been extensively applied in various contexts, including educational technology (Dizon, 2016; Lee et al., 2009; Ratna & Mehra, 2015; Teo et al.,2008), healthcare (Nikolaus et al., 2014), and consumer software (Lilley et al., 2004), due to its easiness and predictive power (Davis et al., 1989; Lee et al., 2003). However, as technology evolves and new factors that influence user acceptance emerge, the need for more comprehensive models becomes evident. TAM2 (Venkatesh & Davis, 2000) extended the original model by including additional variables such as social influence and cognitive instrumental processes, further refining the

understanding of technology adoption. Building upon this, TAM3 (Venkatesh & Bala, 2008) introduced an even more nuanced framework by integrating the determinants of perceived ease of use from the original TAM with the determinants of perceived usefulness from TAM2. TAM3 considers a wider range of factors, including computer self-efficacy, perceptions of external control, and perceived enjoyment, making it particularly suitable for evaluating complex and interactive technologies. This model provides a robust framework for assessing the likelihood of technology adoption by addressing both user perceptions and contextual factors that may influence technology acceptance (Davis et al., 1989).

The SUS (Brooke, 1996) is a widely used standardized questionnaire for measuring usability, specifically designed to generate a comprehensive single measure of perceived usability. It is known as a "quick" survey scale that allows practitioners to assess the usability of novel technologies easily and efficiently. Since its introduction, SUS has been incorporated into a variety of commercial usability toolkits and technology-based systems, including consumer software, websites, applications, and hardware products (Bangor, Kortum, & Miller, 2008; Blažica & Lewis, 2015; Brooke, 2013; Lewis, 2018; Marzuki, Azwany, & Yaacob, 2018).

The SUS items have been developed according to the three usability criteria defined by ISO 9241-11: (1) the ability of users to complete tasks using the system and the quality of the output of those tasks (i.e., effectiveness); (2) the level of resource consumed in performing tasks (i.e., efficiency); and (3) the users' subjective reactions using the system (i.e., satisfaction).

The original SUS (Brooke, 1996) includes a mix of positive and negative items, and it was structured to control for acquiescence bias and to identify respondents who did not pay attention to the statements. However, several studies have highlighted the potential issues associated with the inclusion of both positive and negative items (Barnette, 2000; Stewart & Frye, 2004). For example, it can result in a reduction in internal reliability (Stewart & Frye, 2004), a distortion of the factor

structure (Pilotte & Gable, 1990; Schriesheim & Hill, 1941), and an increase in the number of interpretation problems encountered when using the scale in cross-cultural contexts (Wong et al., 2003).

Moreover, mixed items may lead to difficulty in switching users' response behaviors and it can increase the cognitive load (Kortum et al., 2021). This could be particularly true for clinical and younger populations, whose executive functions may be impaired or still in development.

Notably, some authors suggested that questionnaires for usability assessment should avoid the inclusion of a mixture of positive and negative items and that researchers who do not specifically need to use the standard SUS should consider using the positive version to reduce the likelihood of response or scoring errors (e.g., Lewis, 2018). A version of the SUS that includes only positively worded items has been created (Sauro & Lewis, 2011) and it has been found to have satisfactory reliability, validity, and sensitivity (Lewis; 2018). Thus, the positive SUS emerges as a valuable alternative to the standard SUS, offering the advantages of reduced cognitive load and shifting ability.

Since 2014, several translations and psychometric evaluations of SUS have been published, including Arabic (AlGhannam et al., 2018), Danish (Hvidt et al., 2020), Chinese (Wang, 2020), French (Gronier & Baudet, 2021), Italian (Borsci et al., 2009), Malay (Marzuki et al., 2018), Persian (Dianat et al., 2014), Polish (Borkowska & Jach, 2017), Portuguese (Martins et al., 2015), Slovene (Blažica & Lewis, 2015), and other languages. In addition, translations of SUS into German (Perrig, 2024), and Finnish (Raita & Oulasvirta, 2011) have been conducted, although there is no available data regarding their psychometric properties.

Despite the wide adoption of SUS, there remains a significant need for validations in different populations (Lewis, 2018). To the best of our knowledge, there are currently no instruments specifically designed for the developmental age population that encompass both the dimensions of usability and acceptability.

Based on these premises, we developed and validated the UAS, a new and comprehensive scale designed to assess both the usability and acceptability of a digital tool in the developmental population. Starting from assessing face validity and content validity of the scale, the present investigation aimed to validate the UAS in a group of children aged 4-18. In particular, we investigated the construct validity of the UAS examining the factorial structure of the scale using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). Then we investigated the reliability of UAS measurement checking the internal consistency Cronbach's α, McDonald's ω, and corrected item-total correlation. Moreover, we verified if the factorial structure was invariant for sex. This study would provide evidence about the reliability and validity of this new tool for assessing the usability and acceptability of digital technologies in children and adolescents, thereby enhancing the development and implementation of child-friendly digital assessments.

2. Materials and Methods

The Usability and Attitude Scale (UAS)

The UAS is an eight-item questionnaire, based on the Italian adaptations of the SUS (Brooke, 1996) and TAM (Davis, 1989; Venkatesh & Bala, 2008). It has been specifically designed to assess usability and acceptability in children and adolescents. It has been structured to achieve brevity and ease of comprehension, simplifying the language, and using emoticons to make it easier for the younger users to understand and answer the questions. In the UAS, participants are asked to rate their agreement with eight statements using a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree." Each statement is presented in a child-friendly format using simple language accompanied by icons. Specifically, the verbal response options were paired with coloured emoticon images, this visual aid was designed to increase comprehension and engagement in the rating process (see Figure 1).

Figure 1. Example of verbal response option and visual response option of UAS. The verbal responses are presented in Italian. Below is the corresponding English translation: Decisamente no! = Definitely not!; No = Not; Non so. = I don't know.; Sì= Agree; Decisamente sì! = Definitely yes!.

In the present study, we asked 10 experts to evaluate the content validity of the questionnaire. Experts were chosen based on their expertise in the field of usability and acceptability and their experience with children in neuropsychological assessment. A background questionnaire, the UAS scale, an introductory letter, and the content validation form were created and sent via email. On completion, all items were returned electronically and the comments from the expert panel were integrated into the questionnaire. Following the analysis of expert feedback, the questionnaire was revised to be evaluated by subsequent participants, incorporating changes such as the addition, removal, or modification of certain items. The original questionnaire consisted of $N = 22$ items, and after the adaptation process, the updated version was finalized. The final version of the UAS consists of eight items: four items assessing the usability dimension, based on SUS items, and four items evaluating the acceptability dimension, derived from TAM3 items. The SUS authors advocated for a flexible application of the SUS items according to the context of interest. Thus, only four out of 10 SUS items were considered. Specifically, referring to the positive SUS version (Sauro & Lewis, 2011), only positive assertions were included (i.e., Items 1, 2, 3, and 7), as they are more comprehensible for the targeted audience of children.

A similar approach was followed in selecting items related to acceptability. The original TAM version consists of 12 items, with six evaluating perceived usefulness (PU) and six assessing perceived ease of use (PEU). PU refers to the degree to which a person believes that technology will enhance job

performance, while PEU is defined as the extent to which a person believes that using technology will be effortless (Davis, 1989). However, we aimed to use the UAS for assessing the usability and acceptability of MatriKS, a new digital assessment tool for assessing fluid intelligence designed to enhance the job performance of clinicians and experimenters, but not to improve that of the tested participants. For this reason, the UAS questionnaire was intended for the test-takers only and included items related to PEU. Within the PEU items, the UAS questionnaire selectively incorporated just four items. Specifically, Items 5 and 6 pertained to the perceived ease of use for both oneself and others, Item 7 focused on the perceived pleasantness/fun, while Item 8 addressed intentionality.

Additionally, the study considered face validity, defined as the extent to which the scale's purpose is apparent to the target users and appears to be an appropriate measure for its intended purpose (Salkind, 2010; Mosier, 1947). Face validity encompasses aspects such as feasibility, readability, consistency in style and formatting, and the clarity of the language used (DeVon et al., 2007). Moreover, $N = 10$ participants were asked to select the statement they thought best reflected the construct in question and to change the wording, delete statements, or add new aspects they thought relevant.

Procedure

In this study, we administered the UAS to assess the usability and acceptability of MatriKS, a new digital tool for the assessment of fluid intelligence based on the theory of knowledge structures (Doignon & Falmagne, 1985; Falmagne & Doignon, 2010; Heller & Stefanutti, 2024). MatriKS is available in two versions, differentiated according to participants' age, namely between 4 and 11 years old and 12 or more years old. Participants completed a different version of MatriKS according to their chronological age. The test is part of PsycAssist (https://psycassist.fisppa.unipd.it/en/), a platform for the assessment of neuropsychological functioning (de Chiusole et al., 2024).

Participants completed first MatriKS and then the UAS using a tablet (IoS operating system with a 10.9-inch screen). All participants completed the questionnaire independently, except for the youngest

children, aged 4 to 6. For these children, examiners assisted by reading the items aloud, as the children's reading skills were not yet fully developed. Children then responded autonomously to the items, with visual support provided by emoticons.

Participants

The UAS was administered to a total sample of $N = 908$ participants of the general population aged 4 to 18 (Female = 53%, Male = 47%, Mean age: 10.53 ± 3.50). The descriptive statistics of the general sample are reported in Table 1. Participants were randomly divided into two groups to conduct the EFA (EFA group, $n = 359$) and CFA (CFA group, $n = 549$). The two groups did not differ in terms of age ($t_{(906)} = -0.29$; $p = 0.79$), sex ($\chi^2_{(1)} = 1.30$; $p = 0.25$), or educational level ($\chi^2_{(3)} = 0.63$; $p = 0.89$). The sample size of the two groups was established a priori according to the minimum criteria to have a subject to an item ratio of 10:1 in the EFA (Nunnally, 1978) and at least 10 observations for each freely estimated model parameter in the CFA (Kline, 1998; 2023). The data were collected in different regions of the North, Centre, and South of Italy. Schools of all levels were involved, from kindergarten to high school. The school principals were contacted in advance to inform them about the aim of the project and to inquire about their willingness to participate in the project. Subsequently, the parents or legal guardians of the children enrolled in the participating classrooms were provided with informed consent in accordance with the Declaration of Helsinki recommendations (Williams, 2008). Only children whose parents provided informed consent were included in the study, and subsequently, the children themselves decided whether or not to participate. The exclusion criteria were the presence of motor, visual, and auditory impairments that prevented the participants from completing the task and the questionnaire. An expert team of psychologists handled the assessments and managed communications with both teachers and parents.

The study was approved by the ethical committee for the psychological research of the University of Padua (protocol code E047A9B0520E9732A8365DC72335EE90/ date of approval: 8 July 20222).

Table 1. Descriptive statistics of the total sample of participants (N = 908) and Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) groups.

Variables		Total sample (n=908)	EFA group (n=359)	CFA group (n=549)
Sex (n)	M-F	47-53	49-51	46-54
Age (Years)	Range	4-18	4-18	4-18
	M (SD)	10.52 (3.50)	10.01 (3.45)	10.91 (3.48)
Educational level (n)	Kindergarten	107	42	65
	Primary school	346	142	204
	Middle School	251	98	153
	High School	204	77	127

Note: M-F: male-female; *M(SD)*: Mean (Standard Deviation).

Data Analysis

All statistical analyses were performed on the variables related to the final version of the UAS (see Table 2 for details on items) using IBM SPSS 25 (SPSS Inc., Chicago, IL, United States) and JASP version 0.18.3.0 (JASP Team, 2024).

As suggested by Tabachnick and Fidell (2013), for the EFA and CFA groups we checked for the presence of univariate outliers (scores more than 3.29 standard deviations above or below the corresponding group mean were excluded), normalized the UAS total score distribution, and checked for the presence of multivariate outliers. Each time we excluded participants, we normalized the UAS score distribution.

Table 2. Items of the final version of Usability and Acceptability Scale (UAS).

UAS Scale	Italian version	English version
Item 1	Rifarei volentieri la prova	I would gladly take the test again
Item 2	La prova è stata molto facile	The test was very easy
Item 3	Penso che tutti possano fare questa prova	I think that anyone could take this test
Item 4	È stato facile capire cosa fare durante la prova	It was easy to understand what to do during this test

Item 5	Penso che fare la prova sul tablet sia stato semplice	I think that taking the test on the tablet was easy
Item 6	Credo che per tutti sia facile fare la prova con il tablet	I think that it would be easy for anyone to take the test on the tablet
Item 7	Sono stato bene a fare le prove sul tablet	I felt good taking the test on the tablet
Item 8	Vorrei usare sempre il tablet per fare questa prova	I would like to always use the tablet for this test

Exploratory Factor Analysis

The Exploratory Factor Analysis (EFA) was conducted using the Weighted Least Squares (WLS) technique, which aims to capture most of the variance across a set of variables with a reduced number of factors (Fabrigar, 2012). An oblimin rotation was applied, suitable when factors are expected to be correlated (Fabrigar et al., 1999; Costello & Osborne, 2005).

Parallel analysis, scree test, incremental variance, and interpretability of the pattern of factor loadings were employed to choose the number of factors to retain (Fabrigar et al., 1999). Items with loadings greater than 0.40 and cross-loadings less than 0.10 were considered for inclusion in a factor.

Confirmatory Factor Analysis

The goodness of fit of the factorial structure of the scale identified in the EFA was tested. Maximum likelihood estimation with a mean-adjusted Chi-square test (MLM estimator), which is robust to non-normal score distributions, was used. The metric of the latent variables was set by fixing the factor loading of the first item to one, for each factor. Overall model fit was determined by using the Satorra-Bentler scaled Chi-square statistic (S-Bχ^2), robust comparative fit index (rCFI), robust root mean square error of approximation (rRMSEA) with associated 95% confidence intervals (CIs), and standardized root mean square residual (SRMR) (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Values close to .95 for rCFI, smaller than .05 for rRMSEA, and smaller than .08 for SRMR suggest a reasonable fit (Byrne, 2011).

The factorial structure found in EFA was compared with alternative nested models, which were theoretically plausible. To this aim, ΔS-Bχ^2 and ΔrCFI were used as fit indices. To indicate that the null hypothesis of equivalence should be rejected (i.e., the EFA factorial structure model had a better fit than the alternative model), a significant ΔS-Bχ^2 and a value of ΔrCFI (which is less affected by sample size) higher than .01 are required (Cheung & Rensvold, 2002). Finally, Akaike's information criterion (AIC) was calculated, with a lower value indicating a better fit of the model to the data (Schermelleh-Engel et al., 2003; Sterba & Pek, 2012).

Measurement invariance

To verify the measurement invariance of the UAS, we tested the model across the sex of participants (female vs. male), performing three levels of invariance: configural, metric, and scalar (i.e., same item intercepts) (Vandenberg & Lance,2000). Measurement invariance was assessed specifically within the CFA subsample.

For the assessment of configural invariance, no constraints were imposed, enabling a test of whether the pattern of fixed and freely estimated parameters remained consistent across sub-groups. Regarding the metric invariance (weak invariance), the factor loadings of each item on the corresponding factor (i.e., the scale unit and the metric of each item) were constrained to be invariant across subgroups. Moreover, we performed the scalar invariance (strong invariance) where the factor loadings plus the intercepts of each item on the corresponding factor (i.e., the origin of the scale of each item) were constrained to be invariant (Byrne, 2008; Vandenberg & Lance, 2000).

Internal Consistency

As a measure of internal consistency Cronbach's α (cutoff $\geq$ 0.70; Nunnally, 1978), McDonald's ω (cutoff $\geq$ 0.70; McDonald, 1999), and corrected item-total correlations (cutoff $\geq$ 0.30; Streiner and Norman, 2008) were computed for each factor and for the total scale (Dunn, Baguley, & Brunsden, 2014; Soĉan, 2000). These measures were assessed within the total sample.

3. Results

Exploratory Factor Analysis (EFA)

The EFA was conducted on the first subsample (n = 359). Using parallel analysis, scree test, incremental variance, and interpretability of item factor loadings, all methods consistently extracted two interrelated factors, which were then rotated using oblimin rotation. The total explained variance was 42%. All items met the inclusion criteria for factor retention, with loadings greater than 0.40 and cross-loadings less than 0.10. Table 3 presents the item loadings for the two-factor solution (Model 1). The first factor consists of four items related to usability (items 1, 2, 3, and 4), while the second factor includes four items representing acceptability (items 5, 6, 7, and 8). Skewness and kurtosis computed on the two-factor scores indicated approximately normal univariate distributions, as all values were lower than |2| (Tabachnick and Fidell, 2013); skewness ranged from -1.347 to 0.149 (SE = 0.18), and kurtosis ranged from -0.853 to 3.153 (SE = 0.357).

Table 3. Exploratory Factor Analysis (EFA) results.

EFA factor loadings (n=359)			
Item content	Mean ± *SD*	F1	F2
1. I would gladly take the test again	3.95 ± 0.88	**0.694**	0.010
2. The test was very easy	3.59 ± 0.98	**0.771**	-0.031
3. I think that anyone could take this test	3.71± 0.97	**0.456**	0.002
4. It was easy to understand what to do during this test	3.94 ± 0.96	**0.689**	0.055
5. I think that taking the test on the tablet was easy	4.34 ± 0.67	0.100	**0.584**
6. I think that it would be easy for anyone to take the test on the tablet	3.81 ± 0.98	-0.047	**0.548**
7. I felt good taking the test on the tablet	4.46 ± 0.61	0.082	**0.641**
8. I would like to always use the tablet for this test	4.15 ± 0.87	-0.091	**0.661**
Eigenvalues after Oblimin Rotation		1.8 (23%)	1.1(19%)

Confirmatory Factor Analysis (CFA)

The two-factor model selected in the EFA was tested on the second subsample (n = 549) using CFA. Results indicated an acceptable fit to the data, with all indices close to the expected value: S-Bχ^2 = 89,76, p < .05; rCFI = .91; rRMSEA = .082, 95% CI [.066, .100]; SRMR = .044. Aiming to evaluate possible alternative models explaining specificity or overlapping between the investigated domains, the two-factor model obtained by EFA (Model 1, where the items of the questionnaire cluster around domains of usability and acceptability) was compared to a one-factor model solution (Model 2, representing unique general factor in which all items of the UAS were loaded on a single dimension). The CFA findings strongly support the validity of the UAS's two-factor structure showing that Model 1 was the most representative model of the questionnaire structure with the best fit and better results than the alternative nested model: ΔS-Bχ^2 (Δdf) range = 253.718(20) - 89.76 (19); ΔrCFI range = .72–.91. Concerning the other parsimony fit indices, it is possible to note as the two-factor model (Model 1) showed the lowest values for AIC (Model 1= 10424.668; Model 2=10586.617) confirming that Model 1 turned out to be the model that best fits the data and best explains the dimensionality of the analyzed data. All factor loadings were statistically significant. Each item loaded highly (>0.70) and significantly (p < 0.001) on its designated factor, with factor loadings from 0.71 to 0.97 (see Figure 2).

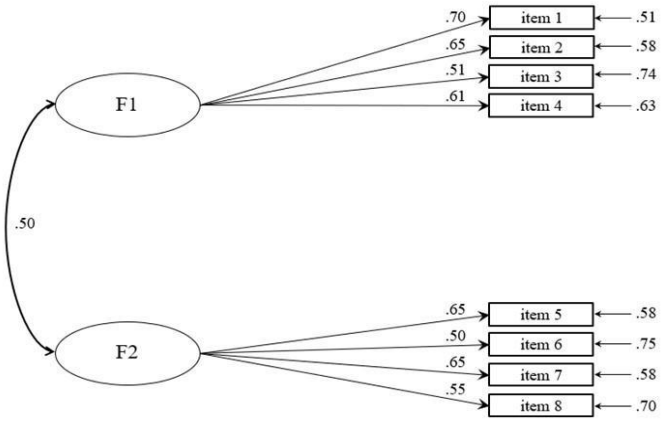


Figure 2: Measurement model with standardized parameters (n = 549)

Measurement Invariance

The two-factor model exhibited full configural, metric, and scalar invariance (Table 4 presents the results of the sex invariance test, and fit indices for the model across males and females). All factor loadings were statistically significant and comparable between groups ($p < .001$). The models assessing configural, metric, and scalar invariance demonstrated good fit indices that did not vary significantly from one model to the others. The UAS showed configural, metric, and scalar invariance across sexes (see Table 3), indicating that latent structure, factor loadings, and intercepts are equivalent for both groups.

Table 4. Results of the measurement invariance test across groups (sex: male vs female)

Model	GFI	NFI	PNFI	SRMR	RMSEA
Model (a) – Configural: Factor structure constrained to be equal	0.998	0.881	0.598	0.050	0.082
Model (b) – Metric: Factor loadings constrained to be equal	0.998	0876	0.735	0.052	0.072
Model (c) – Scalar: Item intercepts constrained to be equal	0.997	0.860	0.844	0.050	0.069

Note: GFI: Goodness of Fit Index; NFI: Normed Fit Index; PNFI: Parsimony-Adjusted Normed Fit Index; SRMR: Standardized root means square residual; RMSEA: Root Mean Square Error of Approximation.

Internal consistency

The internal consistency of the UAS was assessed using both Cronbach's α and McDonald's ω, yielding promising results. The overall reliability was found to be acceptable, with Cronbach's α at .74 and McDonald's ω at 0.77, suggesting that the scale is internally consistent. When examining the individual factors, Factor 1 (F1) showed solid reliability with McDonald's ω at 0.71 and Cronbach's α at 0.70. However, Factor 2 (F2) had slightly lower internal consistency, with McDonald's ω at 0.65 and Cronbach's α at 0.66. While these values indicate somewhat lower reliability, they still fall within acceptable ranges, suggesting that both factors can be considered reliable, though further investigation might be warranted for Factor 2.

4. Discussion and conclusion

This study aimed to validate and analyze the psychometric properties of the Usability and Attitude Scale (UAS), a questionnaire designed to assess the usability and acceptability of digital technologies among children and adolescents.

Our findings from EFA and CFA suggest that the UAS is a reliable and valid tool for evaluating usability and acceptability among children aged 4 to 18 years. The EFA revealed a clear factor structure for the UAS, indicating distinct dimensions for usability and acceptability. The results demonstrated that the items loaded appropriately onto the hypothesized factors, with eigenvalues supporting a two-factor solution. This is consistent with the theoretical foundations of the Technology Acceptance Model (TAM) and System Usability Scale (SUS), which posit that perceived ease of use and attitude towards technology are critical determinants of user acceptance. The CFA confirmed the two-factor model structure identified in the EFA, with fit indices indicating a good model fit. These findings reinforce the scale's structural validity and suggest that the UAS is a reliable instrument for measuring the constructs it was designed to assess. The internal consistency, as indicated by Cronbach's α, was satisfactory for both factors, providing further evidence of the reliability of the scale. Our study aligns with previous research highlighting the importance of usability and acceptability in the adoption of digital technologies by children. The TAM framework has been extensively validated in adult populations, but its application to children is relatively novel. The successful adaptation and validation of the UAS demonstrate that these theoretical constructs are applicable and relevant to younger users as well.

The development and validation of the UAS for children represents a significant advancement in the assessment of digital technologies tailored for younger users. The UAS could be a valuable tool for developers, educators, and researchers engaged in the implementation of digital technologies for children. It provides a reliable measure of usability and acceptability, thus enabling the identification

of strengths and improvement areas in digital products with the goal of better meeting the needs of young users. This can enhance the effectiveness of educational and therapeutic interventions delivered through digital platforms. For developers, the UAS provides insights that guide the design of user-friendly and engaging products. For educators and researchers, it assesses the impact of digital technologies on learning outcomes and engagement, promoting the evidence-based integration of technology into educational settings.

Although the UAS demonstrates strong psychometric properties, we acknowledge some study limitations. The data collection, though in diverse regions, was limited to Italy, and cultural factors may influence the generalizability of the findings. Therefore, future research should explore the applicability of the UAS in different cultural contexts to ensure its broader validity. Additionally, while the UAS captures key aspects of usability and acceptability, it may benefit from further refinement and expansion to include additional dimensions such as long-term engagement and user satisfaction. Moreover, longitudinal studies could provide important insights into how these perceptions evolve over time and with prolonged use of digital technologies. Concerning future directions, it would be beneficial to verify the convergent evidence correlating this instrument with SUS and TAM or other relevant questionnaires tailored for developmental age.

To conclude, our results support the two-factor structure of the UAS, aligning with the theoretical constructs of usability and acceptability. The EFA results support a two-factor structure for the UAS, capturing distinct but related constructs of usability and acceptability. The identified factors are both meaningful and interpretable, reflecting the theoretical underpinnings of the scale. The CFA results provide robust evidence supporting the validity and reliability of the two-factor structure of the UAS, making it a valuable instrument for evaluating digital tools designed for children. The good internal consistency and significant factor loadings indicate that the UAS could be a reliable tool for assessing children’s usability and attitudes toward digital tools like MatriKS.

The development and validation of the UAS represent a significant contribution to the field of child-centred digital technology assessment. By providing a reliable and valid measure of usability and acceptability, the UAS helps ensure that digital technologies designed for children are both effective and engaging. This, in turn, can enhance the educational and developmental outcomes associated with the use of these technologies, ultimately contributing to better learning experiences and improved quality of life for young users.

Study 3

Multi-method validation of the new computerized test of fluid intelligence MatriKS

de Chiusole, D., Epifania, O. M., Anselmi, P., Brancaccio, A., Mazzoni, N., Spinoso, M., Orsoni, M., Giovagnoli, S., Pierluigi, I., Bacherini, A., Benassi, M., Balboni, G., Stefanutti, L. (2025, August 22; under review). Multi-method validation of the new computerized test of fluid intelligence MatriKS. https://doi.org/10.31219/osf.io/akd26_v1

Abstract

This paper introduces MatriKS, a new computerized tool for the assessment of fluid intelligence. Based on knowledge structure theory (KST), a mathematical framework initially designed for efficient assessment and personalized learning, MatriKS is the first large scale application of KST to fluid intelligence assessment. The validation results for MatriKS, suitable for Italian individuals aged 4 to 11 (N = 568) are presented. A multi-method approach incorporating classical test theory (CTT), the Rasch model, and KST was employed. CTT confirmed the reliability and validity of the test, the Rasch model ensured precise measurement across different ability levels, and KST provided nuanced insights into fine-grained individual cognitive profiles and developmental trajectories. The study concludes by exploring the methodological and practical benefits of using KST in assessment test construction and multi-method analysis.

Keywords: MatriKS, multi-method validation, fluid intelligence, Raven's matrices, knowledge space theory, test development

1. Introduction

It is not new that the advancement of science is strictly dependent on newer, more sophisticated techniques to measure and analyze the variables under investigation. This is particularly true for psychology, a century-and-a-half-year-old discipline that is inherently complex, with multiple components, most of which are latent. As in other empirical sciences, multi-method measurement allows researchers to explore different dimensions and perspectives of a scientific phenomenon. This approach can provide a more comprehensive understanding of psychological constructs and offers stronger evidence for psychological theories compared to single-method studies. Moreover, researchers can cross-verify their findings by comparing results across different methods, reducing the likelihood of errors or biases associated with a single approach. This enhances the credibility and reliability of the conclusions drawn from a research and contributes to the validity and robustness of a research study (see, e.g., Eid & Diener, 2006).

In psychology, at least three different measurement approaches are available nowadays: classical test theory (CTT; see, e.g., Lord & Novick, 1968; Novick, 1966), modern test theory (MTT; see, e.g., Andrich, 2011), and knowledge structure theory (KST; see, e.g., Doignon & Falmagne, 1999; Falmagne, Albert, Doble, Eppstein, & Hu, 2013a). According to Andrich (2011), MTT encompasses both item response theory with continuous latent variables (Embretson & Reise, 2013) and Rasch measurement (Rasch, 1960). Among the three approaches, KST is the newest and less known.

In the field of knowledge assessment and learning, KST marks a significant departure from traditional psychometric theories. Indeed, KST is not aimed at quantifying knowledge or learning in a conventional numerical sense. Rather, it focuses on finding the specific items (or skills) a student masters within a particular field of knowledge, at a given moment. Thus, the feedback of a KST-based assessment is nonnumerical, assuming the form of a set.

Contrarily to KST, CTT and MTT have a long tradition in the field of test and questionnaire validation, and are commonly used for this purpose. While CTT focuses on classical measures such as test scores, MTT offers a more nuanced analysis by considering item-level properties and latent traits. It is not uncommon to see these two methods used together for comprehensive test validation (see, e.g., Bacherini, Anselmi, Havercamp, & Balboni, 2024; Bechger, Maris, Verstralen, & Béguin, 2003; Eluwa, Eluwa, & Abang, 2011).

Instead, until a few years ago, KST has been applied almost exclusively to the educational context (see, e.g., Falmagne, Albert, Doble, Eppstein, & Hu, 2013b). In the last decade, some extensions of KST have been proposed that moved it from its native landscape to the wider field of psychological and neuropsychological assessment. In particular, Bottesi, Spoto, Freeston, Sanavio, and Vidotto (2015) and Donadello et al. (2017) applied KST to build and validate a questionnaire for the psychological assessment of depression. Moreover, Stefanutti (2019) proposed the so-called procedural knowledge space theory and showed that it can be successfully applied in the field of the assessment of human problem-solving and planning (Stefanutti, de Chiusole, & Brancaccio, 2021; Brancaccio, de Chiusole, & Stefanutti, 2023). Nevertheless, the application of KST in the field of cognitive psychology is still underexplored.

Along this line, this paper presents MatriKS, a new computerized tool to measure fluid intelligence developed using KST and validated with a multi-method approach, including CTT, the Rasch model and KST.

Thus, the novelty of this research is double-faced. First, it applies KST to fluid intelligence assessment, marking a novel use of the theory in this context. Second, it uses CTT, the Rasch model, and KST together, with the purpose of demonstrating that a multi-method approach can enhance the robustness, reliability and comprehensiveness of test validation. By combining these diverse

methodologies, researchers can benefit from the distinct advantages that each method offers, providing a more holistic and effective tool for the assessment of fluid intelligence.

The paper is organized into the following sections. Section 2 summarizes the key KST-based methods and procedures used in constructing and validating a cognitive test. Section 3 provides a detailed description of the principles underlying the creation of the new Raven-like matrices integrated into MatriKS. The multi-method validation analysis is illustrated in Section 4. This section outlines the administration procedure and participants, followed by a comprehensive presentation of data analyses and results from the perspectives of CTT, the Rasch model, and KST. A brief comparison of the results obtained by the three approaches with some suggestions on how to integrate them is given in Section 4.5. The paper concludes with some final remarks.

2. Knowledge structure theory: construction and validation of cognitive tests

The theory of knowledge structure (Doignon & Falmagne, 1985, 1999; Falmagne & Doignon, 2011) is a mathematical approach conceived for the non-numerical assessment of knowledge. Defining a knowledge domain Q as the set of all problems useful for assessing knowledge in a given field (e.g., mathematics, chemistry, statistics, etc.), the individual's knowledge is represented by the subset of problems in Q they are able to solve. This subset is named knowledge state and it is denoted by K. When a population of individuals is considered, it is possible to define the collection K of all knowledge states existing in that population, named knowledge structure.

It is worth noticing that a knowledge structure reflects the precedent relations (e.g., prerequisites) existing among the problems in Q. Thus, K is a subset of the power set 2^Q (i.e., the collection of all the subsets that can be obtained from Q). In particular, a subset of Q is a knowledge state whether, for every pair q, r ∈ Q of problems, if r is a prerequisite of q (r ≺ q) and q belongs to the knowledge

state K ($q \in K$), then also r belongs to the knowledge state K ($r \in K$). As a consequence, the knowledge state containing r without q does not exist.

KST was initially developed as a behavioral theory that do not provide any assumptions or descriptions of cognitive processes or skills behind the solution of a problem. Later, the theory was extended to a "competence" level for the assessment of cognitive skills (Doignon, 1994; Düntsch & Gediga, 1995; Falmagne, Koppen, Villano, Doignon, & Johanessen, 1990; Gediga & Düntsch, 2002; Stefanutti & de Chiusole, 2017; Ünlü et al., 2013; Heller, Stefanutti, Anselmi, & Robusto, 2015; Korossy, 1997, 1999). Such extension is known as competence-based knowledge structure theory (CbKST; Heller, Ünlü, & Albert, 2013a; Heller, Augustin, Hockemeyer, Stefanutti, & Albert, 2013; Stefanutti & Albert, 2003). Given a set S of skills, the competence state is the set $C \subseteq S$ of skills mastered by a individual and the collection C of all the competence states is the competence structure. The behavioral and competence levels can be connected by two functions: the skill map and the problem function (Doignon, 1994; Düntsch & Gediga, 1995; Heller, Ünlü, & Albert, 2013b). The skill map is a triple (Q, S, μ) where $\mu : Q \rightarrow 2^S$ is a function assigning a nonempty subset of skills in S to each item in Q. In this way, it connects the performance level to the competence level. Notably, skill maps have two alternative interpretations called conjunctive and disjunctive (Doignon, 1994). In the former, all the skills in μ(q) are necessary for solving q. In the latter, any skill in μ(q) is sufficient for solving q.

Under the conjunctive model (i.e., the approach used in the following), the problem function $p : C \rightarrow 2^Q$ is defined by

$$p(C) = \{q \in Q : \mu(q) \subseteq C\},$$

for each state $C \in C$. The collection of all the p(C) induced by the states $C \in C$ is, indeed, a knowledge structure consisting of knowledge states that are delineated by μ.

The knowledge and the competence structures are deterministic models that need to be empirically validated. This phase requires testing the fit of a probabilistic model to some empirical data. The model mostly used for this purpose is the basic local independence model (BLIM; Doignon & Falmagne, 1999; Falmagne & Doignon, 1988). The BLIM is a probabilistic model that makes a distinction between the latent knowledge state K of an individual and their observable response pattern R (i.e., the subset of the item correctly solved). This distinction between unobservable knowledge states and observable response patterns is posited following the assumption that observed data are noisy and that they might mask the true knowledge states.

Under the BLIM, the relation between K and R is probabilistic and it is given by a model in which the marginal probability P(R) of the response patterns can be computed by

$$P(R) = \sum_{K \in \mathcal{K}} P(R|K)\pi_K,$$

where π_K is the probability of K in the population. Assuming that the responses provided by a student to the items are locally independent, given the true knowledge state of that student, the conditional probability P(R|K) can be computed by

$$P(R|K) = \left(\prod_{q \in K \setminus R} \beta_q\right)\left(\prod_{q \in K \cap R} (1-\beta_q)\right)\left(\prod_{q \in R \setminus K} \eta_q\right)\left(\prod_{q \in Q \setminus (K \cup R)} (1-\eta_q)\right)$$

,where β_q ∈ [0, 1) is a careless error probability and η_q ∈ [0, 1) is a lucky guess probability. The careless error parameter reflects the conditional probability that an individual does not solve item q given that q belongs to their knowledge state K. On the contrary, the lucky guess parameter reflects the conditional probability that an individual solves q given that q does not belong to their K. The careless error and lucky guess parameters of an item provide a measure of random error in the two directions of, respectively, a false negative and a false positive. Therefore, they also provide an

indirect measure of reliability of each item. In this respect, it is typical to check if condition $\beta_q + \eta_q < 1$ holds. If it does not hold for some items, they should be removed from the data analysis.

It is worth mentioning that two distinct procedures exist for deriving parameter estimates for the BLIM: by maximum-likelihood (ML) via the expectation-maximization (EM) algorithm adapted for KST (Stefanutti & Robusto, 2009; Anselmi, Robusto, Stefanutti, & de Chiusole, 2016; de Chiusole, Stefanutti, Anselmi, & Robusto, 2015) and by minimum discrepancy (Heller & Wickelmaier, 2013). Moreover, the "pks" R package and a MATLAB toolbox provide the functions for applying these two estimation procedures (Brancaccio, de Chiusole, & Wickelmaier, 2024).

Concerning the goodness-of-fit of the model to the data, it can be performed by using the Chi-square statistic and by applying a likelihood ratio test between the estimated and the saturated models. If the model fits the data there is empirical evidence that the items, the skills, and the structure built for collecting the data are plausible in the real word.

Over time, thorough investigations have delved into this model, establishing it as a robust framework for validating knowledge structures. A (non-exhaustive) list of studies on this model is Stefanutti, Heller, Anselmi, and Robusto (2012); de Chiusole, Stefanutti, Anselmi, and Robusto (2013); Stefanutti and Robusto (2009); Stefanutti, Spoto, and Vidotto (2018); Anselmi, Stefanutti, Chiusole, and Robusto (2017); de Chiusole, Anselmi, Stefanutti, and Robusto (2013); de Chiusole and Stefanutti (2013); Heller et al. (2015); Heller, Stefanutti, Anselmi, and Robusto (2016).

One of the most important results obtained about the BLIM concerns its local identifiability. If the domain Q has a moderate size (e.g., up to about 20 items) a general test of the local identifiability of the BLIM might be performed by using the MATLAB function blimit (Stefanutti et al., 2012). This function tests the local identifiability of the BLIM for arbitrary knowledge structures. In case the model is not locally identifiable for a given knowledge structure, the blimit function provides precise diagnostics about the local identifiability of every single parameter in the model. When the number

of items in Q is greater than 20, no analytic procedures can be applied, and the only possibility to test the local identifiability of the model is via an empirical procedure. This last consists of estimating the BLIM with the same data set, each time starting from a different point in the parameter space. If the model is identifiable, the estimated parameter values do not depend on their initial values, and they remain the same (up to round-off error) all the time. If the model is unidentifiable, then the estimated values depend on the initial values and the obtained estimated values will change each time. Thus, the standard deviation of the parameter estimates and the difference between the maximum and the minimum values of the estimates can be analyzed. If parameters are identifiable, their estimates have zero variance (up to round-off error), otherwise they have a nonzero variance. For a similar approach, see, e.g., Stefanutti, de Chiusole, Anselmi, and Spoto (2020) and Stefanutti et al. (2021).

Having available a model that fits the data and whose item parameter estimates are reliable and identifiable, the assessment of the underlying skills can be performed by assigning to each participant their knowledge and competence states. For each response pattern R and each knowledge state $K \in \mathcal{K}$, the conditional probability P(K|R) is computed by an application of the Bayes theorem, with the following formula:

$$P(K|R) = \frac{P(R|K)\pi_{\mathcal{K}}(K)}{P(R)},$$

where P(R) and P(R|K) are computed, respectively, by Equations (2) and (3), and π_K are the state probability estimates. For each individual, the posterior probability distribution among the states is obtained by computing P(R|K) for all $K \in \mathcal{K}$. Then, modal state $\hat{K}$ in the posterior probability distribution is taken to be the knowledge state of the individual.

It is worth noticing that the same knowledge state could be mapped into different competence states. This happens whenever a 1-to-1 correspondence between knowledge states in K and competence states in C does not hold. In these situations, the "minimum competence state" can be computed for

each knowledge state (Stefanutti & de Chiusole, 2017; de Chiusole, Stefanutti, Anselmi, & Robusto, 2020; Heller, Anselmi, Stefanutti, & Robusto, 2017) which includes only the skills that the individual certainly masters. The minimum competence state $\hat{C}$ that corresponds to a particular knowledge state $\hat{K}$ is built by taking the union of the subsets of rules assigned to each item $q \in \hat{K}$ via the skill map μ, that is:

$$\hat{C} = \bigcup_{q \in \hat{K}} \mu(q).$$

In this way, for each individual of the sample, the modal knowledge state $\hat{K}$ and the corresponding modal competence state $\hat{C}$ are obtained.

3. Automatic rule-based generation of the stimuli of MatriKS

In this section, the principles that guided the generation of MatriKS stimuli are given and justified in light of the most relevant literature that studied the structure of Raven's stimuli and the properties that contribute to their difficulty. For an overview, see, e.g., Carpenter, Just, and Shell (1990), DeShon, Chan, and Weissbein (1995), Matzen et al. (2010), and Primi (2001).

A review of the studies cited above allowed for the development of a new and simplified taxonomy that individualizes the so-called "transformation rules" mostly used for creating new Raven-like stimuli. These rules are listed in Table 1 and refer to the following five macro-categories: elaboration of the general configuration, visuospatial transformation, pre-inference reasoning, inference reasoning, and directional logic.

Table 1: Taxonomy of the "transformation rules" most commonly used in the literature.

Macro-category	Rule	Definition
Configuration elab.	Completion	Identification of the missing portion of an object
Visuospatial	Orientation	Manipulation of the spatial orientation
	Shape	Manipulation of the shapes
	Filling	Manipulation of the shading
	Size	Manipulation of the size
Pre-inference	Object addition	Visual combination of two objects into a whole
	Object subtraction	Visual deletion of the objects across cells
Inference	Conjunction (AND)	The object in the third cell is obtained via the intersection between the objects in the first two cells
	Disjunction (OR)	The object in the third cell is obtained via the union of the elements in the first two cells
	Exclusive disj. (XOR)	The object in the third cell is obtained via the symmetrical difference between the elements in the first two cells
Directional logic	Horizontal	Rules are applied horizontally (across columns)
	Vertical	Rules are applied vertically (across rows)
	Diagonal	Rule are concurrently applied vertically and horizontally

Note: The third cell can be the third cell of either a row or a column. The diagonal directional rule can follow the main diagonal of the matrix from the top-left corner to the low-right corner (TL-LR directional logic) or the secondary diagonal of the matrix from the low-left corner to the top-right corner (LL-TR directional logic).

Figure 1 shows one Raven-like matrix as an example of the way the transformation rules apply in a matrix.

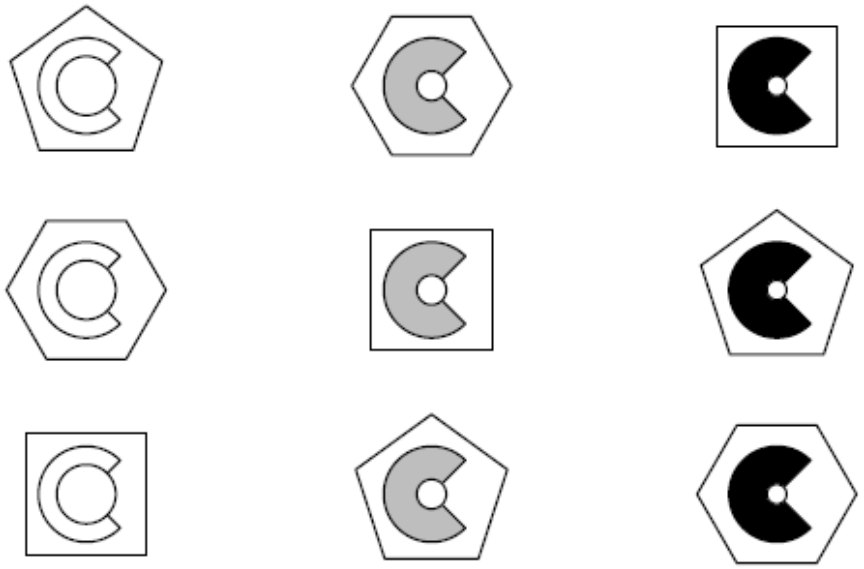


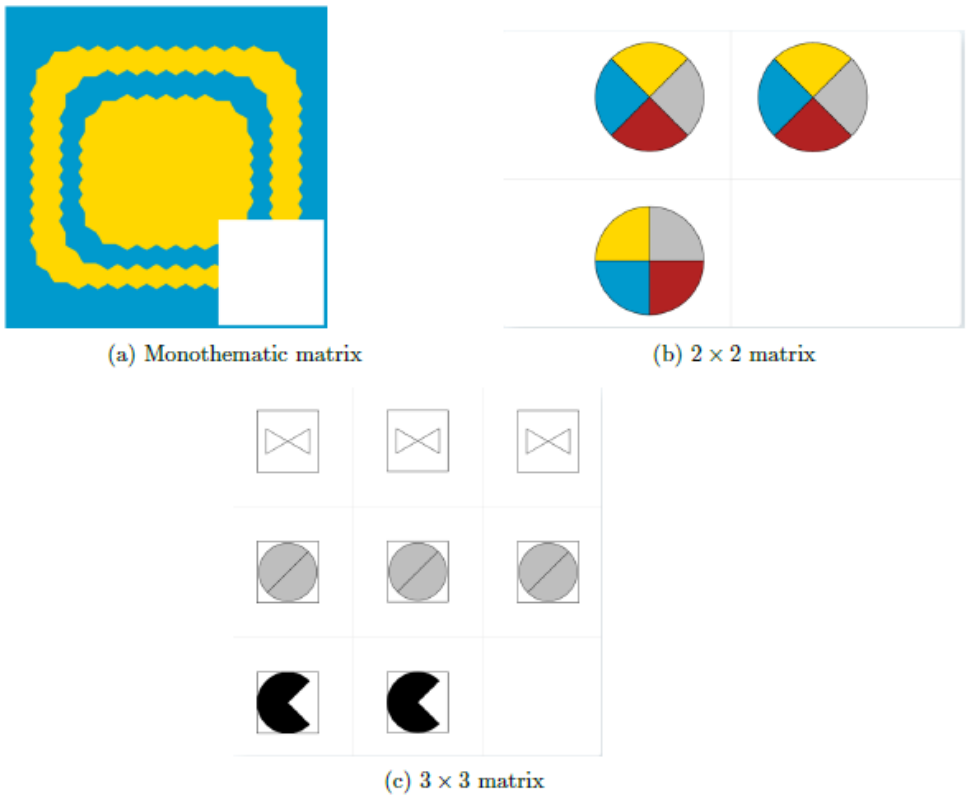
Figure 1: Example of a Raven-like matrix automatically generated manipulating the external shape (diagonal logic), filling (horizontal logic) and size (horizontal logic) transformation rules.

The manipulated transformation rules are: (i) the external shape, that varies with the so-called top-left to low-right diagonal logic; (ii) the filling of the "pacman", that varies with a horizontal logic; and (iii) the size of the circle, that varies with a horizontal logic, too.

Individuals aged 4 to 11 were chosen as the target group for the test administration. Thus, all the generative rules belonging to the category inference reasoning were excluded. This is in line with the stages of cognitive development proposed by Piaget (Piaget, 1978) establishing that logical thinking (i.e., formal operational stage) is expected to come after 12 years old. All the other rules were taken into account in building the stimuli of MatriKS.

Other than the transformation rules, Raven-like matrices can be created by using different number of cells and different types and numbers of response options. Dimensions used in current Raven's tests are: (a) the so-called "Monothematic" or "jigsaw puzzles" matrices, representing a single figure from which a piece is missing; (b) 2 × 2 matrices, representing series of four cells, one of which is missing;

(c) 3 × 3 matrices, representing series of nine cells, one of which is missing. The missing cell is typically at the low-right corner of the matrix. Figure 2 shows an example for each of the three types of matrices. MatriKS contains all three types of matrices.



(a) Monothematic matrix

(b) 2 × 2 matrix

(c) 3 × 3 matrix

Figure 2. Three examples of Raven-like matrices having different dimensions. Namely a monothematic matrix (Panel 2a), a 2 × 2 matrix (Panel 2b), and a 3 × 3 matrix (Panel 2c) are depicted. The answer to each Raven-like matrix has to be chosen from a list of response options. Only one option is correct and all the others are wrong. Wrong options are named distractors. Studies on common response errors in the Raven's tests (see, e.g., Kunda, Soulieres, Rozga, & Goel, 2016) led to group distractors into the following four categories: (i) repetition refers to an error response corresponding to a cell adjacent to the missing one; (ii) wrong principle refers to an error response that applies an incorrect combination of rules, (iii) difference refers to an error response characterized by a pop-out effect since it is perceptively different from all the other responses; and (iv) incomplete

correlate refers to an error response with a missing or modified transformation rule. In MatriKS, the response options were designed to represent each of the four categories of distractors. Figure 3 shows one distractor for each of the four types for the matrix depicted in Figure 1.

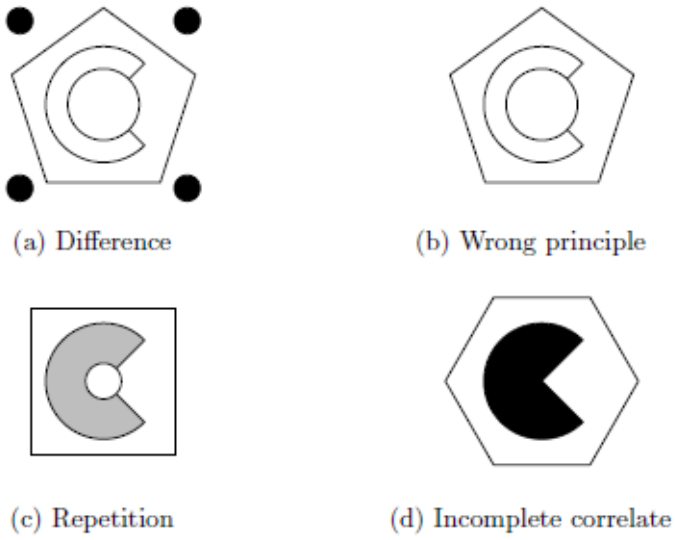


Figure 3. Example of the plausible distractors for the matrix depicted in Figure 1.

In MatriKS, response options of the 2 × 2 and monothematic matrices include one distractor for each macro-category along with the correct response. Thus, an overall of five response options was considered for these matrices. In the 3 × 3 matrices, two distractors for each category were included in the response options, except for the difference distractor, for which only one distractor was included. This was done intentionally to emphasize the pop-out effect of this distractor. Thus, an overall of eight response options was considered for 3 × 3 matrices.

In choosing the characteristics of the stimuli for MatriKS, the aim was to create an item pool that is: (1) sufficiently big to leave the possibility to exclude some items (if needed); (2) representative of several combinations of the transformation rules; and (3) composed by items with gradually increasing complexity.

Table 2 displays the features of each of the 40 items of MatriKS. Columns 1 to 4 indicate, respectively, the number of the item, the type of matrix among Monothematic (Mono in the table), 2 × 2, or 3 × 3,

if the matrix is colored or not, and the number of response options. The last column displays the subset of transformation rules manipulated in each matrix.

Table 2: The 40 stimuli included in MatriKS. See text for more details.

#	Color	Dim.	# Options	$\tau(q)$
1	yes	Mono	5	$\{C\}$
2	yes	Mono	5	$\{C\}$
3	yes	Mono	5	$\{C\}$
4	yes	Mono	5	$\{C, V\}$
5	yes	Mono	5	$\{C, V, H, VH\}$
6	yes	2×2	5	$\{C, O, V\}$
7	yes	2×2	5	$\{C, O, V, H, VH, D\}$
8	yes	2×2	5	$\{C, O, F, V, H, VH\}$
9	yes	2×2	5	$\{C, F, Sh, V\}$
10	yes	2×2	5	$\{C, F, Sh, V, H, VH\}$
11	yes	2×2	5	$\{C, O, Sh, V, H, VH, D\}$
12	yes	2×2	5	$\{C, O, Sh, V, H, VH\}$
13	no	2×2	5	$\{C, O, V\}$
14	no	2×2	5	$\{C, O, V, H, VH, D\}$
15	no	2×2	5	$\{C, O, Sh, V, H, VH\}$
16	no	2×2	5	$\{C, F, Sh, V\}$
17	no	2×2	5	$\{C, F, Sh, V, H, VH\}$
18	no	2×2	5	$\{C, O, Sh, V, H, VH\}$
19	no	2×2	5	$\{C, O, Sh, V, H, VH\}$
20	yes	2×2	5	$\{C, Sh, OA, V\}$
21	yes	2×2	5	$\{C, Sh, V, H, VH, OS\}$
22	yes	2×2	5	$\{C, Sh, V, H, VH, OS\}$
23	no	2×2	5	$\{C, Sh, OS, V\}$
24	no	2×2	5	$\{C, Sh, V, H, VH, OA\}$
25	no	2×2	5	$\{C, Sh, V, H, VH, OA\}$
26	yes	3×3	8	$\{C, F, V\}$
27	yes	3×3	8	$\{C, S, H\}$
28	yes	3×3	8	$\{C, F, H\}$
29	yes	3×3	8	$\{C, Sh, V\}$
30	yes	3×3	8	$\{C, Sh, S, V, H, VH\}$
31	yes	3×3	8	$\{C, Sh, V\}$
32	yes	3×3	8	$\{C, Sh, V\}$
33	no	3×3	8	$\{C, Sh, V\}$
34	no	3×3	8	$\{C, Sh, S, V, H, VH\}$
35	no	3×3	8	$\{C, Sh, V\}$
36	no	3×3	8	$\{C, F, Sh, V, H, VH\}$
37	no	3×3	8	$\{C, O, F, H\}$
38	no	3×3	8	$\{C, O, F, V, H, VH\}$
39	no	3×3	8	$\{C, O, Sh, S, V, H, VH\}$
40	no	3×3	8	$\{C, O, F, Sh, V, H, VH\}$

Note: C = Completion; O = Orientation; F = Filling; Sh = Shape; S = Size; V = Vertical Logic; H = Horizontal Logic; VH = Vertical & Horizontal Logic; D = Diagonal Logic; OA = Object Addition; OS = Object Subtraction.

It is worth noticing that, assuming that an individual is able to solve the matrix if they recognize each of the rules manipulated in it (later on named "rule-to-skill correspondence assumption"), the last

column of Table 2 represents what in KST is named skill map (Q, S, μ), where Q is the set of the 40 matrices, S is the set of the 11 transformation rules, and μ is the function assigning to each matrix q ∈ Q the subset of rules/skills required for solving it.

Table 2 guided the creation of the stimuli. More in detailed, the stimuli were created using an ad-hoc developed R package designed for the automatic rule-based generation of Raven-like matrices (i.e., the matRiks package, Brancaccio, Epifania, & de Chiusole, 2023). It generates stimuli based on specified parameters, including: (a) the type of matrix (e.g., 2 × 2 or 3 × 3); (b) the objects to be used (e.g., square, circle, etc.); (c) the rules that guide the manipulation (e.g., change in shape, orientation, size, etc.); (d) the direction of the manipulation (e.g., vertical, horizontal, or diagonal). For a comprehensive understanding of the package and its functionalities, readers can refer to the documentation accompanying the matRiks package.

4. Multi-method validation analysis

4.1 Participants

The sample was composed of n = 609 children (47% female). Data from 40 children with developmental disorders (e.g., ADHD, dyslexia), as well as data from one child with a missing response due to the skipping of an item were excluded from the analyses. The final sample was composed of n = 568 children (47.89% female) aged 4 to 11 (mean = 8.33, sd= 2.19). The lower and upper bounds of the age interval are included. This means that the test can be administered from the first day of the 4th year of age to the last day of the 11th year of age.

Data were collected among several Italian primary and middle schools. The informed agreements were given to the parents of the children. Only the children for whom the informed agreement signed by their parents could be retrieved were included in the study. To avoid any form of discrimination, no exclusion criteria were applied at this level, such that every child could participate given the signed agreement of their parents.

The descriptive statistics of the sample are reported in Table 3, along with the total proportion of correct responses for female and male children according to the different schooling years. The schooling years refer to the number of school years that the children have successfully completed. For instance, 0 schooling years means that children have not yet successfully completed an entire school year (e.g., they are enrolled either in kindergarten or in the first year of elementary school). Three schooling years identify children that had completed three entire school years and are currently enrolled in the fourth class of elementary school.

Table 3: Descriptive statistics of the sample considering schooling years (Column 1), gender (Column 2), along with the proportion of correct responses (Column 6).

Sch. years	Gender	n	Age range (SD)	Mean age	Prop. correct (SD)
0	F	84	4.20 – 7.18	5.60(0.80)	0.47 (0.50)
	M	96	4.19 – 7.36	5.63(0.77)	0.44 (0.50)
1	F	26	7.07 – 8.28	7.68(0.33)	0.70 (0.46)
	M	36	7.12 – 8.12	7.68(0.26)	0.68 (0.47)
2	F	53	7.88 – 10.22	8.68(0.39)	0.77 (0.42)
	M	46	8.14 – 9.37	8.74(0.28)	0.75 (0.43)
3	F	56	9.06 – 10.17	9.67(0.32)	0.82 (0.38)
	M	49	9.16 – 10.21	9.74(0.29)	0.79 (0.40)
4	F	23	10.32 – 11.98	10.94(0.37)	0.83 (0.38)
	M	40	9.30 – 11.41	10.80(0.36)	0.82 (0.38)
5	F	30	11.02 – 12.06	11.59(0.32)	0.86 (0.35)
	M	29	11.12 – 11.90	11.53(0.22)	0.84 (0.37)

The subsample of children that also completed the CPM and the ToL was composed of 237 children aged 4 to 11 years old (M = 7.89, DS = 2.04), 51% males (n = 120), attending kindergarten (n = 60, 25%) or primary school (n = 177, 75%). Table 1 of the supplementary material reports the distribution of participants across different school levels. Participants were recruited in two Italian regions: Emilia Romagna (n = 103, 43%) and Umbria (n = 134, 57%), located in the north and center Italy, respectively.

4.2 Assessment tools and administration procedure

MatriKS was administered to each child individually in a separate classroom by trained researchers. The test was presented on the PsycAssist platform (de Chiusole et al., 2024) available at

https://psycassist.fisppa.unipd.it by using a tablet (IoS operating system with a 10.9in screen). The administration of MatriKS was carried out with the tablet held horizontally.

The MatriKS test was introduced by a brief animated video (95 seconds), which presented the test, explained the structure of the stimuli and the correct way to select the response from the response list. After the video presentation, children were presented with four practice items and, then, the test started. There were no time constraints. A response was registered once the child touched one of the response options. If the child struggled to find any answer to an item, the researcher could decide to skip it. No feedback was given about the correctness of the responses provided by the children. The stimuli were presented one at a time in random order.

To investigate evidence of convergent and discriminant validity, the Colored Progressive Matrices (CPM; Raven (1954); Italian adaptation by Belacchi, Scalisi, Cannoni, and Cornoldi (2008)) and the Tower of London (ToL; Shallice (1982); Italian adaptation in Sannio Fancello, Vio, and Ciacchetti (2021)) were administered to a subsample of children, in addition to MatriKS.

CPM represent one of the most popular measures of non-verbal reasoning abilities for children and elders. It is composed of 36 items of increasing difficulty, organized into three series (i.e., A, AB, and B) having 12 items each. The task consists in identifying the piece that best complete the given matrix choosing among six alternatives. Serie A evaluates the ability to identify similarities (based on shape, dimension, direction, quantity, orientation, figure/background, density criteria), serie AB assesses the ability to detect symmetry, and serie B measures the conceptual thinking skills (i.e., detection of abstract relations according to operant-deductive logic and their retention in working memory). A score of one is assigned if the person identified the correct option. Otherwise, the score is zero. Thus, the highest total score achievable is 36.

ToL represents one of the most popular measures of planning abilities of individuals with diverse chronological ages. The materials include a wooden base with three pegs of different lengths and

three colored beads (red, green, blue). The task consists in moving the beads from a starting configuration, the same for each problem, to a drawn target position, different for each problem, in a given number of moves and respecting three rules (i.e., moving one bead at a time, placing the picked bead into a peg before picking up another one, placing no more than two beads on the middle peg and no more than one bead on the shortest peg). ToL consists of 12 items of increasing difficulty, which require an increasingly greater number of moves to be solved. For this study, a score of one was assigned if the individual reached the target position within the required number of moves at the first attempt. Thus, the highest total score achievable was 12.

The three instruments were administered individually in a quiet room at school, by trained psychologists (n = 2) or psychology interns (n = 9). Instruments were administered in a counterbalanced order, both for the typology of test (computerized vs traditional, i.e., MatriKS vs CPM and ToL) and the two traditional measures (i.e., CPM and ToL). Table 4 shows the frequencies of these administration orders. Moreover, between the administration of MatriKS and the traditional instruments (or viceversa) occurred three to four weeks.

Table 4: Frequencies of instruments administration orders.

Order	*n*	%
Typology		
Traditional measures – MatriKS	135	57
MatriKS— Traditional measures	102	43
Traditional measures		
CPM – TOL	129	54
TOL – CPM	108	46

4.3 Methods

Two items (Items 3 and 7 in Table 2) were removed from the analysis because an inconsistency was detected among the response options. The following analyses are hence based on an item pool composed of 38 items. The highest achievable observed score was hence 38.

The proportions of correct responses for female and male participants in each schooling year are illustrated in Table 3. A two-proportion z-test was used to investigate any significant difference in the proportion of correct responses between female and male participants in each schooling year (Type I error probability $\alpha = .05$). The Benjamini-Hochberg correction for multiple comparison (Benjamini & Hochberg, 1995) was applied. No significant differences were found between the proportions of correct responses of female and male participants across schooling years.

4.3.1 Classical Test Theory

In this application, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA; Joreskog, 1969) were used for investigating evidence of the construct validity of the MatriKS score interpretation.

To perform the cross-validation of the latent structure of MatriKS with EFA and CFA, the sample has been split in two sub-samples by considering the stratification of gender and schooling year of the whole sample. This was done in order to ensure gender and schooling year proportional representation in the two sub-samples.

The preliminaries for running EFA and CFA were checked with Kaiser-Meyer-Holkin (KMO, Kaiser, 1970) measure of sampling adequacy and Bartlett's test of sphericity. KMO examines the strength of the partial correlation between the items. Values greater than .80 are considered optimal, while values greater than .70 are considered acceptable for running EFA (Kaiser & Rice, 1974). The Bartlett's test of sphericity is used to test the null hypothesis that the correlation matrix is an identity matrix. To run the factor analysis the Bartlett's test needs to be below the nominal level of significance.

EFA was employed for exploring and determining the number of latent factors and the pattern of relationships between the indicators and the latent factors. CFA was employed for testing the soundness of the factorial structure and for investigating the reliability of the test.

Among other methods available for determining the dimensionality of a test (e.g., MAP, parallel analysis, eigenvalues), the scree test was employed for determining the number of latent factors to investigate with EFA. Given that MatriKS is a maximum performance test where a correct response is expected and the responses are dichotomically coded in as correct or incorrect: (i) the tetrachoric correlations between the items is usually employed as the starting point for EFA, (ii) the weighted least square mean and variance adjusted (WLSMV) estimator is used in EFA, and (iii) the diagonally weighted least squares (DWLS) estimator is used in CFA.

The models investigated with EFA and CFA were considered to have a good fit to the data when the root mean square error of approximation (RMSEA) is equal or lower than .08 (optimal ≤ .06, Hu & Bentler, 1999) and the comparative fit index (CFI) and the Tucker-Lewis Index (TLI) are equal to or greater than .95 (Hu & Bentler, 1999). The items with substantial factors loadings on the latent factor(s) (i.e., ai ≥ .30) were retained, and, in case of solutions with multiple factors the items with cross-loadings (i.e., items with substantial factor loadings on more than one factor) were discarded.

According to Hattie (1985), the following criteria were applied to rule out the over extraction of latent factors and to support the unidimensional model: (i) the first latent factor explains more than 20% of the common variance, (ii) the ratio between the eigenvalues of the first and second latent factors is greater than 3, (iii) the first and second latent factors are strongly correlated (i.e., > .50), and (iv) the indicators present substantial factor loadings (> .30) on the first latent factor.

Reliability was checked by using the Cronbach α. A value α > .85 suggests an optimal reliability, while an increase of .01 of the α value when an item is removed is sufficient for considering the item as not contributing to the internal consistency of the scale, and it is hence removed (Hattie, 1985).

Evidence of construct validity of MatriKS score interpretation based on relations to other variables was studied by computing the Pearson's correlations between the accuracy scores obtained at MatriKS and those received at measures of a similar construct (CPM; convergent evidence) or different

construct (ToL; discriminant evidence). Accuracy scores of MatriKS and CPM were computed as the proportion of correct responses (range 0 – 1), while those of the ToL were calculated as the number of problems correctly solved at the first attempt (range 0 – 12).

Given that participants were recruited in schools, some of them shared the same educational context (i.e., the specific classroom and schools attended). Thus, there was an issue about the dependency of observations. To overcome it, a categorical variable was created, in which each level represented a specific classroom and school and was assigned to all the participants who attended that particular classroom within that particular school. Then, for each test (i.e., MatriKS, CPM, and TOL) a regression was computed having this context categorical variable as the independent variable and the accuracy score of the specific measure (e.g., MatriKS, CPM, or ToL) as the dependent variable, independently for the educational level attended by participants (i.e., kindergarten, the first two years of primary school, or the last three years of primary school). In this way, nine regressions were computed overall. For each regression, the standardized residuals were saved. These residuals represented the MatriKS, CPM, and ToL accuracy scores cleaned for the effect of the specific educational context of participants.

The residuals of MatriKS, CPM, and ToL scores were transformed into normalized z scores. This is because of the differences in the score range found between the ToL residuals (−10.33, 9.67) and the two other measures, that is (−0.49, 0.33) for MatriKS and (−0.28, 0.31) for the CPM, probably due to the diverse accuracy scores initially considered. These normalized residuals were used to carry out all the analyses.

The correlation coefficients for convergent evidence (i.e., between MatriKS and CPM) were expected to be higher than those for discriminant evidence (i.e., between MatriKS and ToL). To test this hypothesis, the magnitude of the correlation coefficients for convergent evidence was compared with those for divergent evidence using the Williams t-test (Steiger, 1980).

Prior to the correlation analysis, the data underwent a comprehensive screening process, following Tabachnick and Fidell's suggestions (2013). This included checks for outliers and normality, ensuring the robustness of the data.

4.3.2 Rasch analysis

According to the Rasch model (Rasch, 1960), the probability of observing a correct response on item i by person p is dependent on both a characteristic of the respondent (as described by the ability parameter θ_p) and a characteristic of the item (as described by the difficulty parameter b_i)[1], which are taken to lie on the same latent trait.

The item response function (i.e., the probability of observing a correct response given the respondent and item characteristics) is formalized as:

$$P(x_{pi} = 1|\theta_p, b_i) = \frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)}.$$

In this application, we use the mathematical notation typical of Item Response Theory models instead of using the typical Rasch notation for describing the respondent (β_p) and item (δ_i) parameters to avoid a clash of notation with KST model parameters.

The higher the value of θ_p, the higher the latent trait level of the person, that is the higher their ability level and the higher the number of correct responses by p. The higher the value of b_i, the higher the difficulty of the item, that is the higher the latent trait level required for responding correctly to i and the higher the number of incorrect responses registered on i. A graphical representation of the item response function (Equation 6) is the item characteristic curve (ICC), which displays the changes in the probability of endorsing item i (represented on the y-axis) as a function of the latent trait level (represented on the x-axis). Generally, the ICCs of different items should be equally spread and distributed along the latent trait, this indicating that the test contains items able to target respondents with different levels of ability. However, if the aim is to develop tests focused on specific regions of

the latent trait (e.g., diagnostic tests), the majority of the ICCs should be located in the regions of interest. This would allow for precisely assessing the regions that are most crucial for discriminating between respondents. Given that MatriKS has been developed for the assessment of fluid intelligence in the clinical population (i.e., low-medium of the latent trait), the majority of the ICCs should be located in the low-medium range of the latent trait, to ensure high measurement precision and discrimination ability between the respondents with those levels.

The fit of the data to the Rasch model can be checked according to the χ^2 statistic of the model and its p-value and to the standardized root mean square residual (SRMR), which should be $\leq .08$ (Hu & Bentler, 1999). To obtain interpretable and meaningful estimates of the Rasch model, three main assumptions need to be met: (i) the items must comply to a monotonic function (i.e., the number of correct responses to each item monotonically increases as the level of the latent trait increases), (ii) the latent trait on which the person and item characteristics lie is unidimensional, and (iii) the correlation between the observed responses to pairs of items must be close to zero once the effect of the latent variable is taken into account (i.e., local independence).

Mokken (1971) suggested the so-called H coefficient to investigate the monotonicity assumption. Values of the H coefficient equal or lower than .30 suggest that the item does not comply to the monotonic function and should hence be discarded. The unidimensionality assumption can be tested by fitting a 1-factor CFA to bring evidence in favor of the unidimensionality of the latent trait. Local independence is evaluated by correlating the residuals of each pair of items, and interpreting it according the Q3 statistics (Yen, 1984). Values of Q3 greater than .30 suggest local dependence between a pair of items (La Porta et al., 2011), although other critical values could be considered (see, e.g., Davidson, Keating, & Eyres, 2004; González-de Paz et al., 2015).

The so-called outfit (outlier sensitive fit) and infit (inlier sensitive fit) statistics (Linacre, 2002; Linacre & Wright, 1994) allow for investigating whether the responses observed on each item and

from each respondent comply to what expected under the Rasch model. In this application, items with outfit and infit statistics between .50 and 2.00 were considered as well functioning items (Linacre, 2002). Specifically, outfit and infit statistics below .50 indicated overfitting items, while infit and outfit statistics over 2.00 indicated underfitting items.

The item and test information functions allow for estimating the amount of information provided by each item and by the scale as a whole, respectively. The item information function (IIF) informs about the precision with which each item measures different levels of the latent trait and, in the Rasch model, it is maximized when the difficulty of item i matches the ability of person p (i.e., $b_i = \theta_p$), while it decreases as the distance between the item difficulty and the ability level increases. The test information function is obtained by summing together the IIFs of all items. Being the sum of the IIFs, the more the locations of the items are scattered throughout the latent trait, the more the test would be informative of different levels of the latent trait.

Finally, the standard error of measurement (SEM) indicates the error of measurement of the test for different levels of the latent trait. Being the square root of the inverse of the TIF, the higher the TIF, the lower the SEM, which implies that the higher the informativeness of the test in a specific region, the lower the SEM, and the higher the measurement precision. Since the final target population of MatriKS is the clinical population (i.e., low-medium levels of the latent trait), the TIF is supposed to be higher and the SEM is supposed to be lower in these regions, thus indicating high measurement precision for low-medium levels of the latent trait.

As an external criterion, the schooling years of the respondents have been considered to check for the validity of the ability estimates provided by the Rasch model. Specifically, the ability estimates of the Rasch model have been grouped according to the schooling years of the respondents. The expectation was that respondents with lower schooling years are located at the bottom of the latent trait

(suggesting lower ability), whereas respondents with higher schooling years are located at the top of the latent trait (suggesting higher ability)

4.3.3 Knowledge structure Theory

The skill map (Q, S, μ) depicted in the last column of Table 2 was used for constructing the knowledge structure of the test under the conjunctive model. A knowledge structure K of 115 states was obtained. It is important to note that the skill map was built under the rule-to-skill correspondence assumption. This assumption might be too strong since, beyond the recognition of all the transformation rules manipulated in generating the matrix, other more intuitive mechanisms could lead to the correct response of an individual. Thus, the empirical test is a crucial phase for understanding if the rule-to-skill correspondence assumption can be retained.

The BLIM was applied to the collected data and its parameters were estimated by ML (Stefanutti & Robusto, 2009). Then, the absolute goodness-of-fit of the model was tested by using the Pearson Chi-square statistic. It is known that the approximation to the asymptotic distribution of the Chi-square statistic lacks accuracy for large and sparse data matrices. Thus, a parametric bootstrap procedure over 1,000 replications was used to compute the p-value of the Chi-square.

The identifiability of the BLIM parameters was checked by applying the empirical procedure described in Section 2. In particular, the estimation of the BLIM's parameters was repeated 1,000 times, each time starting from a different point in the interval (0, .50]. The estimates obtained in the single repetition were retained only if the termination criterion of 10^−3 was reached by the EM algorithm. Otherwise, the obtained estimates were discarded, and the algorithm started again. Then, the standard deviation of the parameter estimates and the difference between the maximum and the minimum values of the estimates were computed. If parameters are identifiable, the standard deviation of the estimates is expected to be very close to zero, otherwise it does not.

Item reliability was tested by checking if condition $\beta_q + \eta_q < 1$ holds for all items. The items that resulted not reliable were removed from the knowledge domain, the skill map, and the data. A new knowledge structure was built and the BLIM was applied again to the data. This procedure was reiterated until all the items turned out to be reliable.

Having available a model that fits the data and whose item parameter estimates are reliable, the knowledge state $\hat{K}$ and the corresponding minimum competence state $\hat{C}$ were estimated for each participant by Equation (4) and (5), respectively. On the basis of these estimates two indexes were computed. The first one is the average proportion of rules mastered by individuals in the general population. It was computed by:

$$p = \frac{\sum_{i=1}^{N} |\hat{C}_i|}{|S|N},$$

where N is the sample size and $\hat{C}_i$ is the modal competence state estimated for an individual i (given a set X, the notation |X| stands for its cardinality). This index was computed separately for the six different schooling years 0 to 5.

The second index is the rule probability P_s, that, for each rule $s \in S$, was computed as

$$P_s = \sum_{C \in \mathcal{C}_s} \pi_{p(C)},$$

where C_s is the collection of all the minimum competence states containing rule s, and p(C) is the knowledge state obtained by applying the problem function in Equation (1). Therefore, P_s is the marginal probability that an individual, in the general population, masters rule s. This index was estimated by maximum likelihood separately for each rule and each of the six schooling years 0 to 5. To this aim, an average posterior probability distribution $\bar{\pi}$ was estimated for each schooling year by simply computing the mean of the estimated posterior probability distributions of the individuals

belonging to the same schooling year. The estimated value $\hat{\pi}$ replaced that of π in Equation (8), to obtain a maximum likelihood estimation of P_s.

What is expected is that both $\bar{p}$ and P_s are monotonically increasing in the schooling years.

4.4 Results

4.4.1 Classical Test Theory

EFA and CFA have been applied with the lavaan package (Rosseel, 2012), while Cronbach α has been computed with the psych package (Revelle, 2023) in R (R Core Team, 2023). The lavaan package has been used also for fitting the unidimensional model with CFA in the Rasch analysis. The first sub-sample (n = 278) has been used for running the scree test and EFA. The second sub-sample (n = 290) has been used for carrying out the CFA and investigating the reliability of MatriKS. No significant differences were found in the proportion of correct responses between the sub-samples and their stratifications (all $p > .05$).

The Bartlett's test of sphericity ($\chi^2 = 48{,}795$, $df = 703$, $p < .001$) indicated that the tetrachoric correlation matrix was adequate for running both EFA and CFA. Moreover, KMO values ranged between .78 and .95.

The scree plot indicated the extraction of three latent factors. However, while the second and third factors appeared to be closely clustered together and proximate to the elbow, the first latent factor was notably distant from the elbow and the other two latent factors. Parallel analysis suggested the extraction of 8 factors, although only the first three observed factors were above the simulated latent factors. Given that: (i) the first latent factor explained more than the 20% of the common variance (i.e., 50%), (ii) the ratio between the eigenvalues of the first two factors was greater than 3 (i.e., 6.52), (iii) the correlation between the first two latent factor was strong (i.e., r = .60), and (iv) all items but one (item 18) presented substantial factor loadings on the first latent factor ($a_i \geq 30$), the unidimensional model was preferred. RMSEA and CFI suggested a good fit of the unidimensional

model to the data (0.03 and .97, respectively). After removing item 18, the Bartlett's test of sphericity (χ^2 = 44,627, df = 630, p < .001) indicated that the tetrachoric correlation matrix was adequate for running both EFA and CFA. Moreover, KMO values ranged between .88 and .96. Discarding item 18 led to a slight increase in the proportion of explained common variance (51%), while RMSEA and CFI remained unaltered. All items showed substantial and significant factor loadings, ranging from .46 to .88.

The factor solution found with EFA was further investigated with CFA, where the 1-factor model was fitted to the data excluding item 18. CFI, TLI, and RMSEA suggested a good fit of the 1-factor CFA model to the data (.96, .96, and .03, respectively). All items presented substantial and significant factor loadings (reported in Table 5 along with their respective standard errors) on the latent factor.

Table 5: Standardized factor loadings of the unidimensional model fitted on MatriKS.

Item	a_i	SE	Item	a_i	SE
1	0.73	0.07	23	0.68	0.06
2	0.84	0.05	24	0.72	0.05
4	0.59	0.08	25	0.58	0.06
5	0.45	0.07	26	0.66	0.05
6	0.74	0.05	27	0.69	0.05
8	0.78	0.04	28	0.71	0.05
9	0.56	0.06	29	0.76	0.06
10	0.65	0.06	30	0.52	0.06
11	0.61	0.06	1	0.75	0.05
12	0.69	0.05	32	0.84	0.03
13	0.56	0.06	33	0.73	0.07
14	0.54	0.07	34	0.43	0.07
15	0.74	0.05	35	0.83	0.06
16	0.53	0.08	36	0.65	0.05
17	0.87	0.04	37	0.73	0.05
19	0.71	0.05	38	0.75	0.05
20	0.66	0.06	39	0.43	0.07
21	0.52	0.08	40	0.63	0.06
22	0.62	0.07			

Note: a_i: Standardized factor loadings of the items on the latent factor, SE: Standard errors of the factor loadings.

Cronbach α = .91 suggested a strong reliability of MatriKS. Moreover, the values of the α without each of the item remained unvaried, indicating that none of the item needs to be removed from the test.

The preliminary screening process for convergence and divergence validity analyses did not find univariate or multivariate outliers, and the univariate and multivariate normality of the score distributions was assured. The descriptive statistics were for all three measures: M = 0, SD = 1, and ranged between (−2.79, 2.79). The magnitude of the correlation coefficient of convergent evidence between MatriKS and CPM was equal to .46 ($p < .01$) and was statistically significantly higher than that of discriminant evidence between MatriKS and ToL ($r = .06$; $p = .38$), as indicated by the results of the Williams t-test ($t(234) = 5.05$; $p < .001$).

4.4.2 Rasch analysis

The Rasch model was fitted to the data by using R (R Core Team, 2023) with the TAM package (Robitzsch, Kiefer, & Wu, 2022), while monotonicity has been checked with the mokken package (Mokken, 1971).

Four items (Items 5, 18, 34, 39) showed an H coefficient below .30 (range: .12−.29), and were hence excluded from the item pool. After removing these non-monotonic items, the 1-factor CFA model was fitted on the remaining 34 items, and it adequately fitted the data according to CFI, TLI, and RMSEA (.96, .96, and .04, respectively). All items presented substantial and significant factor loadings on the latent factor (a_i range: 0.53 to 0.83).

The SRMR of the Rasch model suggested a good fit of the data to the model, although the p-value associated to χ^2 statistic was significant. Nonetheless, given the sensitiveness of the χ^2 to the large sample size employed in this study, this result might reflect a false positive.

No items showed local dependence, but outfit statistics pinpointed the underfit of Item 21 (outfit = 2.08) that was hence removed from the item pool. After removing the item, the same analyses have

been performed again. The 1-factor CFA model adequately fitted the data and the fit of the data to the Rasch model remained the same. No items showed local dependence or underfit.

The item characteristics curves (ICCs) are illustrated in Figure 4. The probability of giving the correct response is represented on the y-axis, while the levels of the latent trait are represented on the x-axis. The dotted light gray line, the solid dark gray line, and the dashed black line represent the ICCs of the monothematic, 2 × 2, and 3 × 3 matrices, respectively. The leftmost curves represent the response functions of the easiest items.

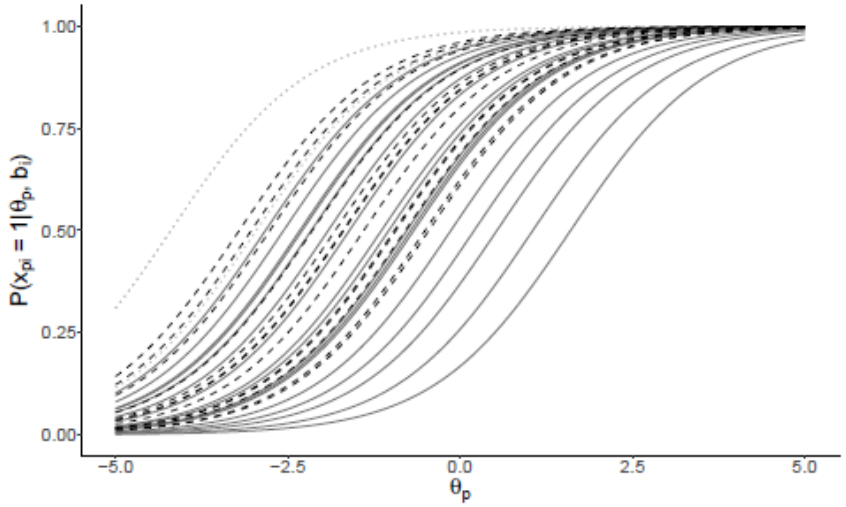


Figure 4: Item Characteristics Curves of the matrices in MatriKS. The latent trait θ_p is on the x-axis. The probability of observing a correct response conditioned to the ability of the respondent and the difficulty of the item is reported on the y-axis. The dotted light gray line represents the Monothematic matrices, the solid gray line represents the 2 × 2 matrices, and the dashed black line represents the 3 × 3 items.

The response functions of two of the monothematic matrices (Items 1 and 2) are represented by the dotted overlapping leftmost ICCs. Interestingly, the outfit statistics indicated these items as overfitting items, which means that these items were so easy that a correct response to them was guaranteed. The 3×3 matrices showed medium to low difficulty levels, while the difficulty levels of the 2×2 matrices were most scattered along the latent trait. The most difficult item was a 2 × 2 matrix (the rightmost solid dark gray line), which was an item that required the mastering of pre-inferential rules for being

solved. The difficulty estimate of each item along with its standard errors are reported in Table 2 of the supplementary materials. The TIF (solid line) along with the SEM (dotted line) of MatriKS are reported in Figure 5.

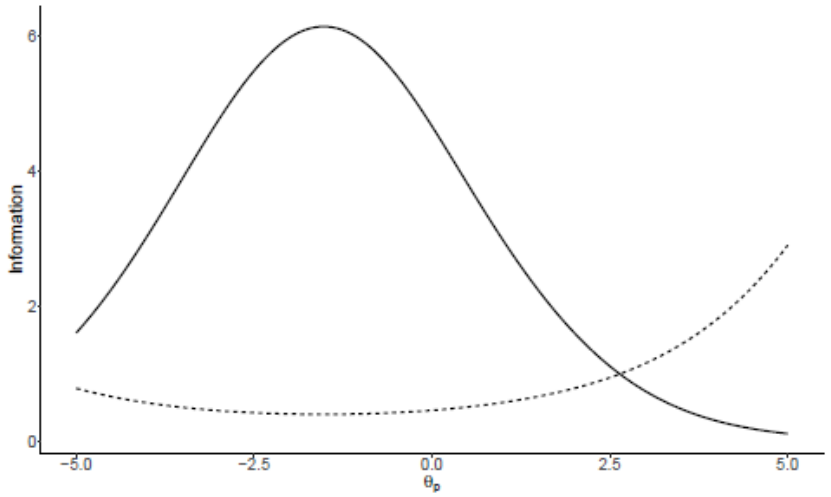


Figure 5: Test information function (solid line) and standard error (dotted line) of MatriKS.

MatriKS resulted highly informative of low to medium levels of the latent trait. Specifically, the TIF is maximized for ability levels around −2, and the test is generally most informative around these levels. The TIF decreases and the SEM increases for higher levels of the latent trait, such that MatriKS might not be precise in the assessment of latent trait levels above approximately 1.5. Taken together, the TIF and the SEM suggest that MatriKS is most informative for latent trait levels that are below the mean, for which it provides the most precise assessments. Nonetheless, this is in line with the aim with which MatriKS has been developed, since it will be used for the assessment of fluid intelligence in clinical populations. The Wright Map, illustrating both respondent abilities and item difficulties, is provided and discussed in Figure 1 of the supplementary materials.

The relationship between the ability estimates of the Rasch model, the schooling years, and the gender of the respondents has been considered. Regardless of gender, the ability levels should increase as the number of schooling years increases. The average ability levels θ_p (and standard deviations) of female (circles over solid line in Figure 6) and male (triangles over dotted line in Figure 6) children have been computed for each schooling level, and the results are illustrated in Figure 6. The light gray line represents the average level of ability.

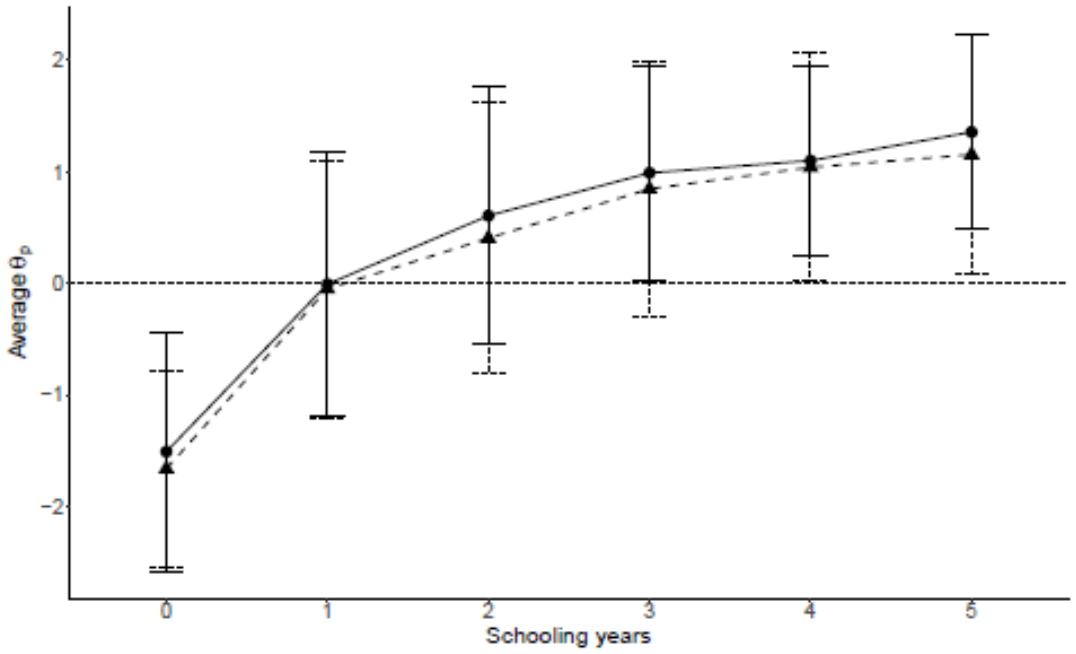


Figure 6: Average ability levels θ_p and standard deviation (x-axis) of female (circles over solid line) and male children (triangles over dashed line) in each schooling year. The horizontal dotted line at 0 indicates the average level of ability θ.

The ability levels estimated by the Rasch model increase as the schooling years of the children increase. On average, children with 0 schooling years show ability levels that are about one standard deviation and a half below the mean, while children with one or more schooling show ability estimates in line with the mean or above the mean, respectively. There are no differences in ability levels between males and females across schooling years.

4.4.3 Knowledge structure Theory

For running all the data analyses, the MATLAB toolbox "kst" (Brancaccio et al., 2024) was used. It is available from the MATLAB file exchange at https://it.mathworks.com/matlabcentral/fileexchange/157751-kst-toolbox.

The knowledge structure obtained with the skill map depicted in the last column of Table 2 is characterized by 115 knowledge states. The BLIM based on this structure applied to the data exhibited a non significant bootstrap p-value = .15. Thus, the model fitted the data well. Moreover, all items turned out to be reliable.

Concerning the identifiability of the model, the standard deviation of both the β_q and η_q estimates turned out to be very close to zero. The highest difference between the maximum and the minimum estimates was 1.6×10^−3 for the β_q and 2.7×10^−3 for the η_q. These values support the identifiability of the model parameters.

It is worth mentioning that β_q and η_q estimates are conditional probabilities and, as such, they are affected by the frequency that the item responses have in the sample (i.e., the item probability). In particular, if the frequency of observing a correct response to an item is very high (item probability close to 1), the conditional probability η_q can be overestimated. Similarly, if the frequency of observing a correct response is very small (item probability close to 0), the conditional probability β_q can be overestimated. This is an issue of ML estimation that arises when there is not enough information in the sample for assuring an unbiased estimate of the parameter (see, e.g., Stefanutti et al., 2020; De Chiusole, Stefanutti, Anselmi, & Robusto, 2013). A way for overcoming this problem is to estimate joint probabilities instead of conditional probabilities. For each item q, the joint probability of β_q was computed by multiplying β_q times π_q, whereas, the joint probability of η_q was computed by multiplying η_q times $1 - \pi_q$.

Figure 7 shows the β_q (top panel) and η_q (bottom panel) parameter estimates obtained by applying the BLIM with K2. In both panels, the conditional (blue circles) and the joint (magenta circles) probabilities are displayed. Labels of the x-axis refer to the ID number of the matrices, as defined in Table 2. The dashed line is for reference.

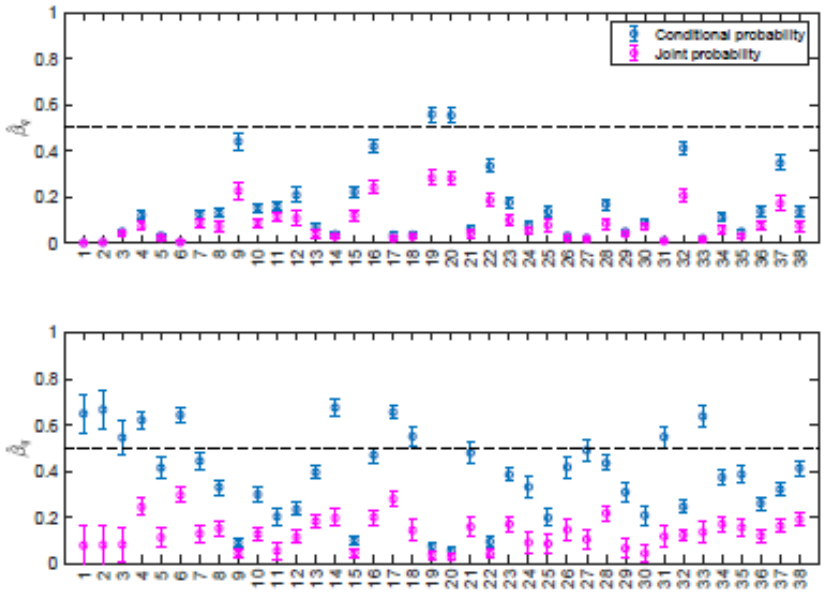


Figure 7: The BLIM's item error parameters estimated on the data. Top panel refers to β_q parameters, whereas bottom panel to η_q. In each panel, the conditional (blue) and the joint (magenta) probability is displayed. Dashed lines are for reference.

The conditional careless error probabilities are, in general, quite small, whereas those of the lucky guesses are quite high. As mentioned above, high values for the lucky guess estimates can be due to high item probabilities. This was the case, because the average item probability is .63, with a minimum value of .50 and a maximum of .88. The joint values of both parameters were quite small: the average value of the joint β_q was .08 (sd = .08) and that of η_q was .13 (sd = .07). Overall, these results indicate that the knowledge structure of MatriKS is plausible and can be used to assess the skills related to fluid intelligence.

The performance of individuals in the general population is described here by means of the average proportion of rule mastered $\bar{p}$ and the rules probability P_s indexes. For each schooling year, Table 6 shows the average proportion of rules mastered $\bar{p}$, and the corresponding standard deviation, estimated for the general population.

Table 6: Average values and standard deviations of the proportion of rule mastered index estimated in the general population for schooling years 0 to 5.

Schooling year	Mean	Sd	Schooling year	Mean	Sd
0	0.32	0.33	3	0.91	0.19
1	0.78	0.26	4	0.95	0.12
2	0.86	0.23	5	0.95	0.14

The average value of this index increases monotonically across the schooling years, while its standard deviation decreases. This suggests that fluid intelligence in the general population improves as the cognitive system of children develops, and that variability among individuals decreases over time. These results are consistent with the relevant literature (see, e.g., Horn, 2007; Schroeders, Schipolowski, & Wilhelm, 2015). Additionally, the average proportion of rules mastered increases rapidly until three schooling years and then approaches a value of 95%. This finding indicates that MatriKS is an adequate tool for assessing fluid intelligence in individuals aged 4 to 11 within the general population.

Figure 8 shows the results about the rule probability P_s index, computed for all the 11 transformation rules. The top panels refer to configuration elaboration (on the left) and visuospatial (on the right) transformation rules. The bottom panels refer to pre-inference (on the left) and directional logic (on the right) rules.

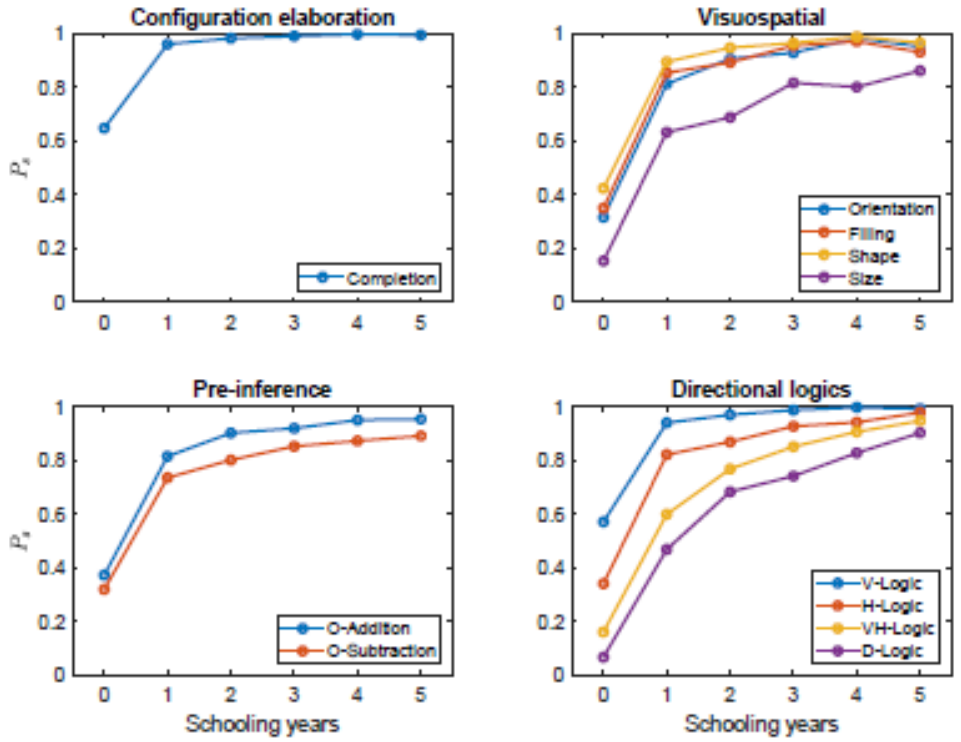


Figure 8: Rule mastering proportion as a function of the schooling years. The top panels refer to configuration elaboration (on the left) and visuospatial (on the right) transformation rules. In contrast, the bottom panels refer to pre-inference (on the left) and directional logic (on the right) rules.

The monotonic increase in schooling years was observed for almost all rules. There were very few exceptions (e.g., for rule Filling at schooling year 5) but by a small amount (the biggest drop was .08). In contrast, standard deviations of this index were monotonically decreasing in the schooling year, with very few exceptions (see Table 7). This means that, in the general population, as the cognitive system of children develops the variability of their fluid intelligence decreases.

Table 7: Standard deviations of rule probability index estimated in the general population for schooling years 0 to 5.

Schooling year	C	O	F	Sh	S	V	H	VH	D	OA	OS
0	0.48	0.47	0.48	0.49	0.36	0.49	0.47	0.37	0.25	0.48	0.47
1	0.19	0.39	0.35	0.31	0.48	0.24	0.38	0.49	0.50	0.39	0.44
2	0.13	0.29	0.31	0.22	0.46	0.17	0.34	0.42	0.47	0.30	0.40
3	0.09	0.26	0.21	0.18	0.39	0.12	0.26	0.36	0.44	0.27	0.36
4	0.04	0.13	0.17	0.10	0.40	0.06	0.24	0.29	0.38	0.22	0.33
5	0.07	0.21	0.25	0.18	0.34	0.09	0.15	0.23	0.30	0.21	0.31

Note: C = Completion; O = Orientation; F = Filling; Sh = Shape; S = Size; V= Vertical Logic; H = Horizontal Logic; VH = Vertical & Horizontal Logic; D = Diagonal Logic; OA = Object Addition; OS = Object Subtraction.

Configuration elaboration transformation rule (top left panel) was the easiest rule. Indeed, its probability was higher than 60% also in preschool children. Among visuospatial rules (top right panel), Shape appeared to be the easiest one, whereas Size appeared to be the most difficult one (less than 20% of preschool children mastered this rule).

Interestingly, Objective addition and subtraction pre-inference rules (bottom left panel) showed very similar probabilities of being mastered at all schooling years, with Objective addition rule being slightly easier than the Objective subtraction rule.

Finally, the bottom right panel of the figure displays the results of the four Directional logic transformation rules. Among these rules, it seems that a dominance relation exists, with the following order (from the easiest to the most difficult): Vertical only, Horizontal only, Vertical and Horizontal, Diagonal. This last, in particular, is the most difficult rule, with a probability less than 10% to be mastered by preschool children.

4.5 Brief comparison among the results obtained with the three approaches

The three measurement approaches CTT, Rasch model, and KST employed for the multi-method validation of MatriKS showed a good fit to the data, bringing evidence in favor of the good functioning of the test under different perspectives. Nonetheless, the three approaches led to different

selection of items. Specifically, Rasch validation led to the deletion of five items from the item pool (i.e., Items 5, 18, 21, 34, 39), CTT validation discarded one item (Item 18), while KST validation allowed for retaining all the items. Item 18 was discarded by both Rasch and CTT. Interestingly, the other items discarded by Rasch are the items that present the four lowest factor loadings on the latent factor in CTT (.43, .43, .45 and .52 for items 34, 39, 5 and 21, respectively). In other words, the items discarded by the Rasch validation are the items that present the lowest proportion of variance explained by the latent factor according to CTT.

It is worth mentioning that the subset of items commonly validated comprises 33 out of 38 items. Thus, it is possible to delineate the assessment outcomes of the individuals by using the three approaches, keeping in mind that the feedback is here provided with a different number of items. However, this aspect does not threaten the quality of the comparison, which aims to illustrate the diversity of the feedback supplied by the three approaches.

Indeed, each assessment method offers a unique perspective, which provides varying levels of detail on individual performance and sheds light on different aspects of the evaluation, from general trends to more granular information. While CTT allows for ordering the respondents according to their scores (i.e., the number of items endorsed), the Rasch model allows for locating the respondent in a specific point of the latent trait (i.e., the ability estimate) and ordering them according to their ability. Since the latent trait is measured as an interval scale, it is possible to interpret the distance of each respondent from the mean in terms of standard deviations, as well as the distances between the respondents. Finally, KST allows for having the most detailed picture on the respondents, as it focuses not just in how much a person knows but specifically on which items they are able to solve and which transformation rules they master. Thus, it is possible that two individuals having the same CTT score and the same level of ability according to the Rasch model, have different profiles in terms of the transformation rules they master.

To better explain these aspects, an example follows where three real respondents taken from the sample, having different age and schooling years, are compared from the point of view of each measurement approach. The three respondents are here named as John, Sophie, and Jane. John and Sophie have 0 schooling years and are 5 and 7 years old, respectively. Jane has 5 schooling years and is almost 12 years old.

According to CTT, John and Sophie obtained a total score of 18 out of 37 while Jane a total score of 33 out of 37, leading to the conclusion that Jane performed better than both respondent John and Sophie. By standardizing the scores and computing the IQ points according to the schooling years of the respondents, John, Sophie and Jane show a performance that is in line with the average performance of the group to which they belong ($z = 0.20$ and IQ points 103 for John ans Sophie and $z = 0.33$ and IQ points 105 for Jane).

The Rasch model allows for locating the respondents along the latent trait, such that it is possible to interpret their performance in light of their latent trait level and their distance from the mean. To best interpret the ability levels of John, Sophie and Jane, the average ability levels for their group of schooling years is computed. In the population, the average ability levels for respondents with 0 and 5 schooling years are -1.59 and 1.26, respectively. John has an ability level $\theta\_John = -1.19$, while Sophie has an ability level of $\theta\_Sophie = -1.34$. Indeed, given the item selection provided by the Rasch validation, John responded correctly to one item more than Sophie, from which the difference in their ability levels. Nonetheless, both John and Sophie show ability levels that are above the mean of their group. Jane, with an ability level of $\theta\_Jane = 1.35$, is above the mean of her own group. These results are in line with those provided by CTT. Nonetheless, it is worth mentioning that the measures provided by the Rasch validation allow for distinguishing between John and Sophie, while those provided by CTT does not.

Concerning KST, it provides detailed feedback on three key aspects: (i) the cardinality of the individual's knowledge state; (ii) the number of careless errors and lucky guesses; and (iii) how many and which transformation rules belong to the competence state of the individual.

Continuing the running example from the KST perspective, John masters 10 items, commits one careless error, and makes 9 lucky guesses. In contrast, Jane masters 38 items, makes 4 careless errors and 0 lucky guesses. According to CTT and Rasch, Jane outperforms John based on the number of items solved, but the extent of this difference varies between the two approaches. Interpreting the observed scores, Jane solved 16 problems more than John, while interpreting knowledge states, Jane masters 28 problems more than John. This difference is explained by the fact that KST provides additionally insights into the error probabilities of each individual. Based on their knowledge states, John made more careless errors than Jane, while Jane made more lucky guesses than John. This nuanced information highlights how KST can offer a more comprehensive evaluation of an individual's performance by also considering their error tendencies, something not typically addressed by CTT and Rasch.

Finally, KST provides detailed insights into both the number and the specific types of transformation rules that each individual masters. This information is crucial for understanding the depth and breadth of an individual's cognitive abilities. Jane masters all 11 transformation rules, demonstrating a comprehensive grasp of the entire set of cognitive tasks. These rules include complex transformations and logical operations, indicating a high level of cognitive flexibility and problem-solving ability.

On the other hand, John masters only 5 of the 11 transformation rules. The specific rules mastered by John are Completion, Shape, V-Logic, O-Addition, and O-Subtraction. This indicates that John has a partial understanding of visuospatial transformation rules, and, moreover, lacks mastery in the more advanced directional logic rules. Specifically, John struggles with Orientation, Filling, Size, H-Logic,

VH-Logic, and D-Logic, which are critical for solving problems that require understanding and manipulating spatial relationships in different directions.

Indeed, the difference in schooling years and age between John and Jane could reasonably account for variations in their rule mastery profiles. However, what is particularly noteworthy within the framework of KST is that also individuals with identical observed scores and ability level may exhibit distinct knowledge states and rule profiles.

For example, let us consider Sophie, another respondent of the sample who shares the same observed score of 18 with John. Despite this similarity, Sophie exhibits a different knowledge state and rule profile. She has mastered 15 matrices and 7 rules, including Completion, Orientation, Filling, Shape, V-Logic, O-Addition, and O-Subtraction. This indicates that Sophie has a strong understanding of visuospatial transformation rules, more extensive than that of John, but she, like John, lacks mastery in directional logic rules.

This example underscores a crucial aspect of KST: it accounts for the complexity and diversity of individual cognitive profiles. Even with equivalent performance on a given assessment, individuals may experience and interact with the world in different ways, resulting in distinct profiles of rule mastery. In this case, while both John and Sophie achieve the same observed score, their underlying cognitive profiles and areas of strength and weakness differ.

5. Discussion and Conclusions

This study underscores the robustness and versatility of MatriKS, an assessment tool of fluid intelligence for children in the general population aged 4 to 11. By employing a multi-method validation approach, incorporating CTT, the Rasch model, and KST, MatriKS displayed a good performance across all three assessment approaches. CTT confirms the reliability and validity of the test, the Rasch model ensures precise measurement across different ability levels, and KST provides nuanced insights into individual cognitive profiles and cognitive trajectories of fluid intelligence.

Furthermore, the obtained results – such as the consistent increase in average fluid intelligence scores among children aged 4 to 11 and the decrease in score variability across schooling years – are consistent with established literature (see, e.g., Horn, 2007; Schroeders et al., 2015). These findings provide compelling evidence supporting the external validity of the test. Moreover, the multi-method validation approach sets a benchmark for the development of psychological and neuropsychological tests. While this study establishes the validity and robustness of MatriKS, several avenues for future research and development remain to be explored.

First, future studies should explore the application of MatriKS in clinical populations. This would involve testing its reliability and validity among children with various cognitive, neurodevelopmental, or learning disabilities. Understanding how MatriKS performs in these groups could provide valuable insights into its diagnostic utility and potential modifications needed to accommodate specific clinical needs.

Second, an adaptive version of MatriKS could be developed, which adjust the administration of subsequent questions based on the responses to the previous ones. This approach reduces time and frustration for the test-takers. Moreover, adaptivity allows the clinicians to administer the test to individuals who, otherwise, could be excluded due to their clinical conditions.

Finally, exploring how MatriKS can be integrated into training programs to monitor and support children with various cognitive disabilities represents another valuable direction. The multi-method feedback provided by MatriKS, which identifies both strengths and areas for improvement, makes it a robust tool for both assessment and training design. In particular, the KST-like fined grained feedback provided by MatriKS represents a valuable tool for supporting the clinician in developing personalized paths of interventions. The ultimate goal is to provide each child with tailored support to their cognitive development.

Study 4

Two Markov solution process models for the assessment of planning in problem-solving

Brancaccio, A., de Chiusole, D., Epifania, O. M., Anselmi, P., Spinoso, M., Mazzoni, N., Bacherini, A., Orsoni, M., Giovagnoli, S., Pierluigi, I., Benassi, M., Balboni, G., & Stefanutti, L. (2025, in). Two Markov solution process models for the assessment of planning in problem-solving. *Psychometrika, 1-31.*https://doi.org/10.1017/psy.2025.10042

Abstract

Tower tasks are popular tools used to measure planning skills. The sequences of moves undertaken by the respondents in solving tower tasks might provide important and useful information to shed light on their planning skills. The paper focuses on the distinction between a situation where planning occurs before action (pre-planning) from one where planning and action are interlaced all along the execution of the task (interim-planning). While the model for pre-planning was already developed by Stefanutti, de Chiusole, and Brancaccio (2021), an alternative model for the interim-planning is proposed. The two models are compared with one another in an empirical study. In accordance with the literature on the development of planning skills, the pre-planning model better fits data collected on individuals aged 14 on, while the interim-planning model displays a better fit with data collected on individuals aged 4 to 8. This result is further corroborated by the analysis of the time performance.
Keywords: procedural knowledge space theory, problem space, tower of London test, Markov models, problem-solving modeling.

1. Introduction

Planning is the ability to identify and organize the sequence of actions necessary to carry out goal-directed behaviors (Cristofori, Cohen-Zimerman, & Grafman, 2019; Lezak, Howieson, & Loring, 2004). In clinical and experimental neuropsychology, tower tasks such as the Tower of London test (ToL test; Shallice, 1982) and its variants (Berg & Byrd, 2002; Unterrainer, Rahm, Halsband, & Kaller, 2005) are popular tools used to measure this ability (Anderson & Reidy, 2012; McCormack & Atance, 2011a; Sullivan, Riccio, & Castillo, 2009). These tasks require planning in terms of means-ends analysis to solve a series of increasingly challenging sequential problems (Krikorian, Bartok, & Gay, 1994).

Despite its popularity, fundamental questions about what exactly the ToL measures remain unsolved (Georgiou, Li, & Das, 2017). For example, Leontjev (1978) proposed three levels of planning (i.e., activity, action, and operations) but it is unclear which one the ToL assesses. Some argue it captures high-level cognitive control (e.g., Shallice, 1982; Morris, Ahmed, Syed, & Toone, 1993), while others suggest it reflects lower-level procedural planning (e.g., Ward & Allport, 1997) or working memory demands (e.g., Phillips et al., 1999). Additionally, different scoring methods (e.g., first move time, total correct, etc.) may not measure the same construct, complicating interpretation. Indeed, a large corpus of literature analyzed the performance at tower tasks considering the accuracy of the performance (i.e., the number of problems correctly solved, either at the first attempt or with multiple attempts), while fewer studies have focused on the time performance (Asato, Sweeney, & Luna, 2006a; Luna, Garver, Urban, Lazar, & Sweeney, 2004), distinguishing between the time spent in pre-planning the solution of the task (i.e., the time elapsed between the presentation of the problem and the first move) and the execution time (i.e., the time elapsed between the first move and the completion of the task). This lack of clarity affects clinical and developmental research, highlighting the need for studies that distinguish planning from other executive functions.

A road for clarifying the relationship between ToL performance and distinct levels of planning could be derived from the integrating process-tracing methods such as eye-tracking, response-time modeling, or model-based approaches that could provide deeper insights into planning processes. For example, Mitsopoulos, Mareschal, and Cooper (2015) applied reinforcement learning models to ToL performance, highlighting how different problem-solving strategies emerge over time. Similarly, Köstering, McKinlay, Stahl, and Kaller (2012) used a latent-class approach to identify distinct patterns of planning impairment in clinical populations. These studies suggested the need for a more nuanced understanding of planning, moving beyond simple accuracy-based measures to computational and latent-variable analyses that capture the dynamics of decision-making and strategy selection. Similarly to what is proposed here, these works have the novelty of considering the so-called "problem space" of the ToL (e.g., Berg & Byrd, 2002; Stefanutti, 2019). The problem space refers to the complete set of possible states, transitions, and solution paths that define the solution process of the problems. Using the formal modeling of the problem space in the ToL task might offer significant advantages in assessing individuals' planning abilities. Indeed, unlike traditional scoring methods that focus on the number of moves or rule violations, a problem-space-based model allows for a deeper understanding of an individual's planning abilities. Indeed, the solving of the ToL test problems involves the planning of sequences of moves. Recording the entire solution process might be difficult when administering the physical version of the test, but not when a computerized version is used. The analysis of the sequences of moves might provide insights into the cognitive processes involved in the resolution of tower tasks. To the best of our knowledge, only two studies considered the specific sequences of moves. Specifically, Stefanutti et al. (2021) proposed a mathematical framework based on procedural knowledge space theory to model the solution process behind tower tasks. A critical assumption of this model, named pre-planning assumption, is coherent with the idea that the problem solver plans the entire sequence of moves in advance. Although this assumption is

plausible in certain populations (e.g., young adults), it might not be in others (e.g., children or clinical populations). In this regard, Brancaccio, de Chiusole, and Stefanutti (2023) proposed a model based on an alternative assumption, named interim-planning, which is coherent with the idea that the planning process might be fragmented throughout the execution of the problem. While a model based on the pre-planning assumption has already been empirically validated, a model based on the interim-planning assumption has not.

The aims of this study are to (i) provide the algorithms for the maximum likelihood parameter estimation of the interim-planning model, (ii) empirically validate it, (iii) compare the pre-planning and interim-planning models in populations of different ages, and (iv) externally validate these models considering the time performance of the respondents.

The manuscript is organized as follows: Section 2 provides a brief literature review on the measurement of planning skills through tower tasks (Section 2.1) as well as an introduction to knowledge space theory (Section 2.2) and procedural knowledge space theory (Section 2.3). Section 2.5 thoroughly presents the Markov solution process models, along with the two assumptions consistent with the pre-planning and the interim-planning strategies. Section 3 illustrates a simulation study carried out for investigating the parameter recovery of the model based on the interim-planning assumption. Then, Section 4 presents the results of an empirical study where the data of a tower task are analyzed with the two models, along with the analysis of the response times. Finally, Section 6 concludes the argumentation.

2. Backgrounds

2.1 Assessment of planning with the Tower of London Test

The original version of the ToL test (Shallice, 1982) consists of three differently colored beads placed on three vertical pegs of different heights that may hold at maximum either one, two or three beads, respectively. The task requires to transform the start state into a goal state with a sequence of actions,

each of which consists of moving one bead from one peg to another, in accordance with specific rules and given the number of moves. A problem is correctly solved when the goal state is achieved within the minimum number of moves.

During the execution of the ToL task many indices can be recorded - such as accuracy, solution time, number of moves, sequences of moves, and rule violations. The most frequently used behavioral measures are accuracy, pre-planning time, and execution time, (e.g., Anderson, Anderson, & Lajoie, 1996; Culbertson & Zillmer, 1998; Krikorian et al., 1994; Unterrainer, Rahm, Leonhart, Ruff, & Halsband, 2003).

Evidence from developmental studies using accuracy as a performance indicator shows a gradual growth in planning skills from childhood to adolescence and early adulthood (Albert & Steinberg, 2011; Asato, Sweeney, & Luna, 2006b; Crone & Steinbeis, 2017; De Luca et al., 2003; Korkman, Kemp, & Kirk, 2001; Luciana & Nelson, 1998; Luciana, Collins, Olson, & Schissel, 2009). Specifically, traditional and computerized versions of ToL test indicate that planning ability improves with age, reaching adult-like levels by 14 years (Krikorian et al., 1994; Malloy-Diniz et al., 2008; Luciana & Nelson, 1998, 2002; De Luca et al., 2003). This trend is consistent with findings using the Tower of Hanoi, which show progressive planning improvement during childhood (Díaz et al., 2012). Findings on age-related changes in planning time are mixed. Some studies report increased planning time with age, particularly in complex tasks (Asato et al., 2006b; Filippi, Ceccolini, Periche-Tomas, & Bright, 2020; Steinberg et al., 2008), while others show a decrease during childhood and early adolescence (Huizinga, Dolan, & Van der Molen, 2006; Injoque-Ricle, Barreyro, Calero, & Burin, 2014). Task difficulty appears to modulate these effects across age groups.

The aforementioned studies used pre-planning time as an index of planning skills. However, it is conceivable that, in some cases, planning during the ToL task does not end when the first move is made, but extends throughout the task. Indeed, the execution phase likely encompasses not only the

motor actions of bead movement, but also the cognitive processes involved in ongoing online planning and error correction (Berg & Byrd, 2002; Luciana et al., 2009). Notably, to our knowledge, there is a paucity of research that attempts to isolate, differentiate or quantify, within the execution phase, the time spent on movement and the time spent thinking. Indeed, there have been attempts to reduce the planning before the first move (Owen et al., 1995; Phillips, Wynn, Gilhooly, Della Sala, & Logie, 2001; Ward & Allport, 1997), however no study has measured planning time within the execution phase. In the traditional paper-and-pencil versions of the ToL, the distinction between movement and planning time is not typically provided, and would pose challenge for the experimenter to measure it. Nevertheless, computerized versions offer the possibility to dissect these processes easier, potentially revealing valuable information and insights into the diverse strategies individuals employ to complete the task.

2.2 Knowledge Space Theory

Knowledge Space Theory (KST; Doignon & Falmagne, 1985, 1999; Falmagne & Doignon, 2011) is a mathematical framework developed for the efficient and accurate assessment of knowledge. At its core is the concept of a knowledge state, denoted as a set $K \subseteq Q$, representing the problems that an individual masters within a specific collection of problems Q named the knowledge domain. The entire collection of possible knowledge states for a population is a knowledge structure K. When a knowledge structure is closed under union[1] it is called a knowledge space.

Various methods have been developed over the years to construct a knowledge structure. These methods include exploiting the relation between the problems (e.g., Dowling, 1993; Koppen, 1993; Stefanutti & Koppen, 2003), assuming a competence level along with a set of underlying skills (e.g., Heller, Augustin, Hockemeyer, Stefanutti, & Albert, 2013; Heller, Ünlü, & Albert, 2013; Korossy, 1999), employing data-driven procedures (e.g., de Chiusole, Stefanutti, & Spoto, 2017; de Chiusole, Spoto, & Stefanutti, 2020; Spoto, Stefanutti, & Vidotto, 2016), or deriving a knowledge space from

a problem space underlying the tasks (Stefanutti, 2019). This last approach is the one adopted in this article.

The most common probabilistic model in KST is the Basic Local Independence Model (BLIM; Falmagne & Doignon, 1988). The BLIM is a restricted latent class model where the latent classes represent the knowledge states. In this model, what is observed is a response pattern, denoted as $R \subseteq Q$, which indicates the subset of items correctly solved by each individual. Given a population of individuals, a probability distribution $\pi : K \rightarrow (0, 1)$ on the knowledge states is assumed. Moreover, two parameters estimated for each item $q \in Q$ are considered: the conditional probability of observing a wrong answer for item q given that $q \in K$, denoted βq; the conditional probability of observing a correct answer for item q given that $q \notin K$, denoted ηq. Both parameters have a typical interpretation in the BLIM: βq is a careless error probability for the item q and ηq is a lucky guess probability for the same item.

In the BLIM, the restriction $\beta q + \eta q < 1$ for each item is a required property for obtaining a consistent assessment of the knowledge state, that is, for identifying the knowledge state of an individual based on the observed item responses of this individual. Specifically, the probability of making a careless error on item q should be lower than the probability of failing q due to lack of mastery. On the other hand, the probability of correctly solving an item because it is mastered must exceed that of guessing it. The above restriction is called the "monotonicity condition".

The βq, ηq, and πK parameters of the BLIM can be estimated by maximum likelihood (ML) via the expectation-maximization algorithm (Stefanutti & Robusto, 2009), by minimum discrepancy (MD, Heller & Wickelmaier, 2013) or by an estimation procedure that is a mix of the two (MDML, Heller & Wickelmaier, 2013). Methods are also available for computing ML estimates of the model parameters in the presence of missing data (Anselmi, Robusto, Stefanutti, & de Chiusole, 2016; de Chiusole, Stefanutti, Anselmi, & Robusto, 2015). Once the BLIM based on a specific knowledge

structure K has been estimated and validated on an appropriate sample of individuals, the model parameter estimates can be used to identify the knowledge state $K \in \mathcal{K}$ of each individual based on their item responses. This allows for the assessment of individual differences. Since, in general, knowledge states are partially ordered, the assessment is multidimensional.

[1]K is closed under union if for every subset $\mathcal{F} \subseteq \mathcal{K}$, the union of all the elements in $\mathcal{F}$ is still in K.

2.3 Procedural Knowledge Space Theory

The notion of problem space introduced by Newell and Simon (1972) is a commonly used concept in the study of human problem-solving. The connection between KST and the problem space theory of Newell and Simon (1972) was first pointed out by Stefanutti and Albert (2003). The theoretical foundations were established by Stefanutti (2019), and take the name of Procedural Knowledge Space Theory (PKST). The remaining of the section delves into the fundamental concepts of PKST.

Let Ω be a set of operations. A string of length n of elements in Ω is a sequence $\omega_1\omega_2 \cdots \omega_n \in \Omega^n$. The concatenation between two strings $\omega_1\omega_2 \cdots \omega_m$ and $\omega_{m+1}\omega_{m+2} \cdots \omega_n$, is the string $\omega_1\omega_2 \cdots \omega_m\omega_{m+1}\omega_{m+2} \cdots \omega_n$. The collection of all the strings of operations of finite length is

$$\Omega^* = \bigcup_{n \in \mathbb{Z}^+} \Omega^n,$$

where $\mathbb{Z}_+$ is the set of non-negative integer numbers. The empty sequence ϵ belongs to Ω^* (i.e., $\epsilon \in \Omega^*$).

In PKST, a problem space is defined as a triple $P = (S, \Omega, \cdot)$, where S is a nonempty set of problem states, Ω is a nonempty set of operations, and $\cdot : S \times \Omega^* \rightarrow S$ is an operator satisfying, for all $s \in S$ and $\pi, \sigma \in \Omega^*$, the two axioms:

(P1) $s \cdot \epsilon = s$,

(P2) $(s \cdot \pi) \cdot \sigma = s \cdot \pi\sigma$.

The operator · is named operation application. Problem states and operations are primitive concepts in PKST.

A problem within the problem space P is a pair (s, t) where s, t ∈ S and there exists a string π ∈ Ω* such that s · π = t. The collection of all the problems in P is thus

$$Q = \{(s, t) \in S \times S : s \neq t \text{ and } s \cdot \pi = t \text{ for some } \pi \in \Omega^*\}.$$

Notably, the set Q corresponds to the domain of knowledge in KST. In the problem space P, a solution path is any pair sπ ∈ Π = S × (Ω* \ {ϵ}). In addition, a solution path sπ ∈ Π is said to solve the problem (s, t) ∈ Q if s · π = t. It is worth mentioning that the same problem may be solved by multiple solution paths. A solution path sπ is a subpath of another solution path tσ (denoted by sπ ⊑ tσ) if there are α, β ∈ Ω* such that σ = απβ and t · α = s. The relation ⊑, called sub-path relation is a partial order for the set Π. A subset C ⊆ Π is said to respect path inclusion if the condition

$$s\pi \sqsubseteq t\sigma,\ t\sigma \in C \Longrightarrow s\pi \in C$$

is respected for all sπ, tσ ∈ Π. A subset of solution paths respecting path inclusion is named a competence state of the problem space P. The collection C of all the competence states is the competence space.

The function τ : Q → 2^Π{∅} maps each problem (s, t) ∈ Q to the collection τ(s, t) of all solution paths that solve it.

The collection of all the problems in Q that can be solved by an individual whose competence state is C ∈ C is

$$p(C) = \{(s, t) \in Q : \tau(s, t) \cap C \neq \emptyset\}.$$

This collection p(C) is named the knowledge state delineated by competence state C. It is evident that a problem (s, t) belongs to p(C) if and only if C includes at least one solution path solving (s, t). The collection K of all the knowledge states is the knowledge space derived from the problem space P. Stefanutti (2019) introduced an algorithm for the automatic derivation of a knowledge space from a

problem space. Like KST, also PKST classifies individuals based on their knowledge states, allowing for the assessment of individual differences in problem-solving skills.

A particular kind of problem space called "goal space" has been introduced in Stefanutti et al. (2021) and it is central in this work. In a goal space, each step of the solution process of a problem is classified as "correct-so-far" or "incorrect". A goal space is defined as a problem space $(S, \Omega, \cdot)$ with two distinct states $f, g \in S$ such that

(GS1) for all $\omega \in \Omega$, $f \cdot \omega = f$ and $g \cdot \omega = g$;

(GS2) for each $s \in S \setminus \{f\}$ there is a string $\pi \in \Omega^*$ such that $s \cdot \pi = g$.

The problem states f and g are called the failure and the goal states, respectively, and a goal space is denoted by the 5-tuple $(S, f, g, \Omega, \cdot)$. A solution process is deemed "correct" when it reaches g and "incorrect" when it reaches f.

According to property (GS1), both the failure and the goal states are considered final states, meaning that any operation applied to either of them is ineffective. Property (GS2) stipulates that the goal state g is reachable from any problem state except f. Consequently, any pair (s, g) with $s \in S \setminus \{f, g\}$ constitutes a problem within the goal space.

2.4 An Example with the Tower of London

In this section, an example based on the problem space of the Tower of London (ToL) test is presented to provide a practical illustration of the theoretical concepts introduced in the previous section. In the ToL test a problem state consists of three beads of different colors placed on three pegs of different heights mounted on a wooden support. The three beads can be moved, one by one, from one peg to another. The number of possible problem states in the ToL test problem space is 36, and they coincide with the configurations of the beads in the pegs. Figure 1 depicts the 36 problem states of the ToL.

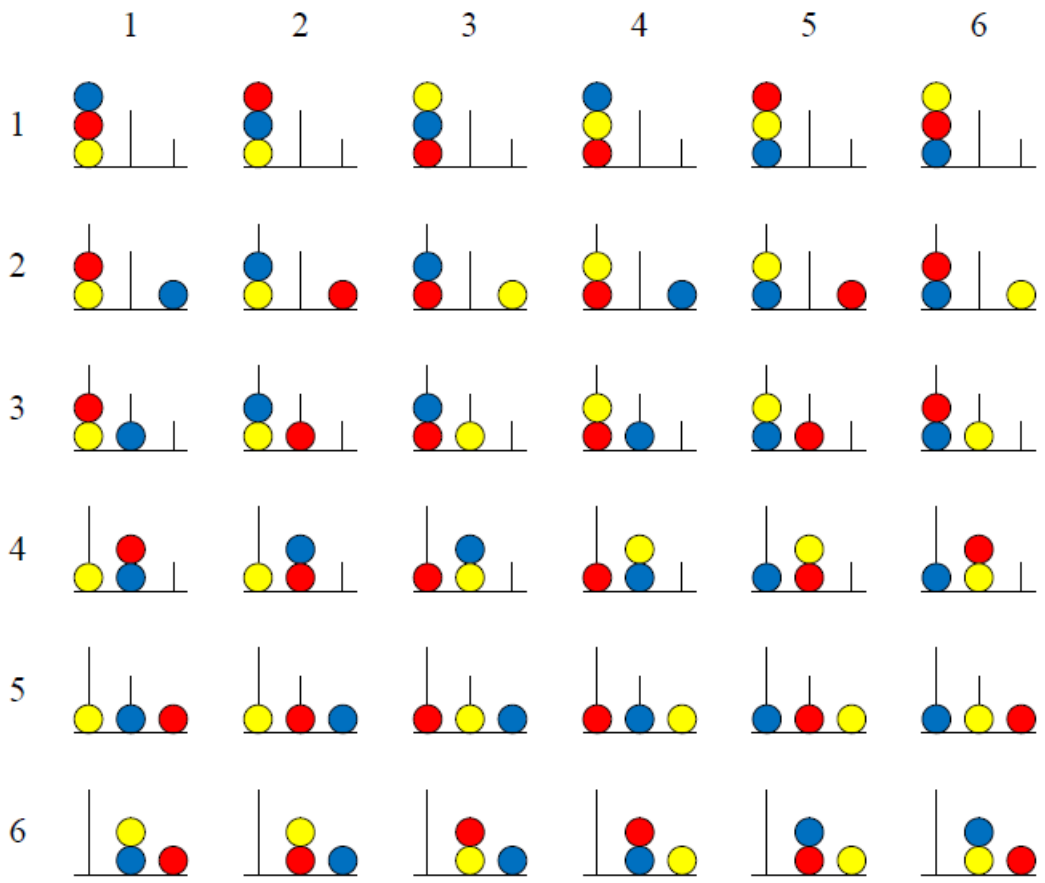


Figure 1: The 6 × 6 diffent problem states of the Tower of London test.

Each of them is labeled by a pair xy of numbers, where x stands for one of the six spatial arrangements (rows in the figure), whereas y stands for one of the color permutations (columns in the figure). An operation consists of moving a ball from one of the three pegs to another. Naming the three pegs as left, center, and right, six operations are entailed: left to center (a), center to left ($\bar{a}$), center to right (b), right to center ($\bar{b}$), left to right (c), and right to left ($\bar{c}$). Therefore, in the ToL, the set of operations is $\Omega_1 = \{a, b, c, \bar{a}, \bar{b}, \bar{c}\}$.

Figure 2 depicts the directed graph of a portion of the ToL test problem space in which the collection of problem states is $S_1 = \{s_1, s_2, s_3, s_4, s_5, s_6\}$. Each problem state corresponds to a label based on the coding used in Figure 1. Specifically, $s_1 = 32$, $s_2 = 22$, $s_3 = 52$, $s_4 = 51$, $s_5 = 21$ and, $s_6 = 31$. The triple $P_1 = (S_1, \Omega_1, \cdot)$ is the problem space for this portion of the ToL test problem space. We can trivially

derive a goal space $P_1g = (S_1g, f, g, \Omega_1, \cdot)$ where $S_1g = S_1 \cup \{f\}$, and $g = s_6$. In addition, all problem states different from the goal are connected to the failure state f by some operations in Ω_1.

In the example, only the subset of problems from goal space P_1g which have s_6 as goal state is considered, namely $Q_1 = \{(s_1, s_6), (s_2, s_6), (s_3, s_6), (s_4, s_6), (s_5, s_6)\}$. Since all the problems in Q_1 have form (s_i, s_6), for lightening the notation, each of them is just represented by the initial state $s_i \in S_1g$. To solve a problem, one needs to know at least one of the solution paths of that problem. For instance, problem s_1 has two possible solution paths, namely $s_1c\bar{a}\bar{b}$ and $s_1ba\bar{c}$. The set of all solution paths that solves anyone of the problems in Q_1 is

$$\Pi_1 = \{s_1c\bar{a}\bar{b}, s_1ba\bar{c}, s_3\bar{a}\bar{b}, s_2a\bar{c}, s_5\bar{b}, s_4\bar{c}\}$$

The mapping τ_1 for the subset Π_1 is constructed instead of deriving the mapping τ for the whole set Π of solution paths. Thus, the mapping τ_1 is defined as follows:

$$\tau_1(s_1) = \{s_1c\bar{a}\bar{b}, s_1ba\bar{c}\},$$

$$\tau_1(s_2) = \{s_2a\bar{c}\},$$

$$\tau_1(s_3) = \{s_3\bar{a}\bar{b}\},$$

$$\tau_1(s_4) = \{s_4\bar{c}\},$$

$$\tau_1(s_5) = \{s_5\bar{b}\}.$$

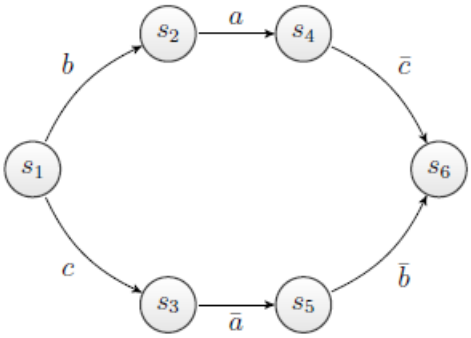


Figure 2: Directed graph of a portion of the ToL test problem space.

The collection of solution paths that respects path inclusion in $\Pi_i g$ is shown in Table 1 Column 1.

Table 1: The 16 competence states obtained from the goal space $\Pi_i g$ and the knowledge states delineated from them are presented in the first and the second columns respectively.

$\mathcal{C}_1$	$\mathcal{K}_1$
$\emptyset$	$\emptyset$
$\{s_4\bar{c}\}$	$\{s_4\}$
$\{s_5\bar{b}\}$	$\{s_5\}$
$\{s_5\bar{b}, s_4\bar{c}\}$	$\{s_5, s_4\}$
$\{s_3\bar{a}\bar{b}, s_5\bar{b}\}$	$\{s_5, s_3\}$
$\{s_2a\bar{c}, s_4\bar{c}\}$	$\{s_2, s_4\}$
$\{s_3\bar{a}\bar{b}, s_5\bar{b}, s_4\bar{c}\}$	$\{s_5, s_4, s_3\}$
$\{s_2a\bar{c}, s_5\bar{b}, s_4\bar{c}\}$	$\{s_5, s_2, s_4\}$
$\{s_3\bar{a}\bar{b}, s_2a\bar{c}, s_5\bar{b}, s_4\bar{c}\}$	$\{s_5, s_2, s_4, s_3\}$
$\{s_1ba\bar{c}, s_2a\bar{c}, s_4\bar{c}\}$	$\{s_2, s_4, s_1\}$
$\{s_1c\bar{a}\bar{b}, s_3\bar{a}\bar{b}, s_5\bar{b}\}$	$\{s_5, s_3, s_1\}$
$\{s_1ba\bar{c}, s_2a\bar{c}, s_5\bar{b}, s_4\bar{c}\}$	$\{s_5, s_2, s_4, s_1\}$
$\{s_1c\bar{a}\bar{b}, s_3\bar{a}\bar{b}, s_5\bar{b}, s_4\bar{c}\}$	$\{s_5, s_2, s_4, s_1\}$
$\{s_1ba\bar{c}, s_3\bar{a}\bar{b}, s_2a\bar{c}, s_5\bar{b}, s_4\bar{c}\}$	$\{s_5, s_2, s_4, s_3, s_1\}$
$\{s_1c\bar{a}\bar{b}, s_3\bar{a}\bar{b}, s_2a\bar{c}, s_5\bar{b}, s_4\bar{c}\}$	$\{s_5, s_2, s_4, s_3, s_1\}$
$\{s_1c\bar{a}\bar{b}, s_1ba\bar{c}, s_2a\bar{c}, s_3\bar{a}\bar{b}, s_4\bar{c}, s_5\bar{b}\}$	$\{s_5, s_4, s_3, s_2, s_1\}$

Let consider the competence states C = $\{s_3\bar{a}\bar{b}, s_5\bar{b}, s_4\bar{c}\} \in C_1$, since $s_5\bar{b} \sqsubseteq s_3\bar{a}\bar{b}$ and both $s_5\bar{b}$, and $s_4\bar{c}$ do not have any subpath indeed C respects path inclusion. The knowledge state delineated by C is obtained by a simple application of the problem function:

$$p(\{s_3\bar{a}\bar{b}, s_5\bar{b}, s_4\bar{c}\}) = \{s_3, s_5, s_4\}.$$

On the whole, applying the problem function p to each of the competence states yields the knowledge states in Column 2 of Table 1. It is worth noting that certain knowledge states, specifically $\{s_5, s_4, s_3, s_2, s_1\}$, are repeated in the table. This repetition arises because the problem function p is not injective.

It is worth recalling that each problem in this example is expressed only by its initial state. In general, problems in the knowledge state would be expressed as ordered pairs of initial and final states. In this example, since P_1g is a goal space, every knowledge state can be obtained from some competence state by just dropping the sequence of operations and retaining the initial state for each of the solution paths in the competence state.

2.5 Markov solution process models

A sequence $(s_0, s_1\ldots, s_m)$ of moves that, in a goal space, leads from the initial state $s = s_0$ to the goal state $g = s_m$ of a problem (s, g) is named *a problem solution process*. Therefore, it is assumed that all the states of a problem space that are traversed by a problem solver are observable. It follows that what is observed for every single problem is not only the correctness of the problem solution, but the entire solution process.

Stefanutti et al. (2021) developed a model, named the Markov solution process model (MSPM), where the problem solution process is regarded as a random process $S = \{S_k : k \in \mathbb{Z}\}$ in which every S_n is a random variable whose realizations are problem states in S. The MSPM postulates a dependence of the problem solution process on the latent knowledge state of the problem solver, and assumes that, given the state of this last, the solution process is Markovian. More precisely, if Q is the set of all the problems having state g as the goal, (Q, K) is a knowledge structure, $K \in K$ is the problem solver's knowledge state, and (s_0, g) is the problem to be solved, then the probability of observing problem state s_n at the n-th step of the solution process, given the previously traversed states, and the knowledge state K, satisfies the Markov property

$$P(\mathbf{S}_n = s_n|\mathbf{S}_{n-1} = s_{n-1}, \mathbf{S}_{n-2} = s_{n-2}, \ldots, \mathbf{S}_0 = s_0; K) = P(\mathbf{S}_n = s_n|\mathbf{S}_{n-1} = s_{n-1}; \mathbf{S}_0 = s_0, K).$$

In the sequel, the probability $P(S_n = s_n|S_{n-1} = s_{n-1}; S_0 = s_0, K)$ is referred to as a transition probability and, to lighten the notation, it is denoted $P(s_n|s_{n-1}, s_0, K)$. It immediately follows from (1) that the

conditional probability of an entire solution process ($s_0, s_1, \ldots, s_m$), given the knowledge state K can be factored as

$$P(s_0, s_1, \ldots, s_m|K) = \prod_{k=0}^{m-1} P(s_{k+1}|s_k, s_0, K).$$

Indicating with PK : K → (0, 1) the probability distribution on the knowledge states in K, the conditional marginal probability of observing ($s_0, \ldots, s_m$) as a solution process for problem (s_0, g), is given by

$$P(s_0, s_1, \ldots, s_n) = \sum_{K \in \mathcal{K}} \prod_{k=0}^{n-1} P(s_{k+1}|s_k, s_0, K) P_{\mathcal{K}}(K).$$

This last is the general equation of the MSPM. With no further assumptions, the number of transition probability parameters $P(s_{k+1}|s_k, s_0, K)$ in the model may be huge, but this number can be drastically reduced. In the first place, as a necessary assumption, a transition from a state s_k to another state s_{k+1} of the goal space can only occur (with a positive probability) if there is an operation $\omega \in \Omega$ that transforms s_k into s_{k+1}. Formally, call a directed edge of the goal space any ordered pair (s, t) of problem states in S for which there is a single operation $\omega \in \Omega$ transforming s into t (i.e., such that $s \cdot \omega = t$). Then $P(s_{k+1}|s_k, s_0, K) > 0$ only if (s_k, s_{k+1}) is a directed edge. Because of this constraint, many transition probabilities will be zero.

Moreover, empirically sound assumptions that allow for further reducing the total number of distinct transition probabilities that need to be estimated can be introduced. Two of these assumptions are investigated in this article. In what follows, these assumptions are labeled (A1) and (A2). The former one was proposed and empirically tested by Stefanutti et al. (2021), whereas the latter one was proposed by Brancaccio et al. (2023), but it was never tested in empirical studies.

According to the (A1) assumption, the transition probability from problem state i to problem state j is assumed to depend on the knowledge state K of the problem solver through the following simple mechanism:

(A1) Given any problem (s_0, g), any directed edge (i, j), and any knowledge state K ∈ K,

$$P(j|i, s_0, K) = \begin{cases} \beta_{ij} & \text{if } (s_0, g) \in K, \\ \eta_{ij} & \text{if } (s_0, g) \in Q \setminus K. \end{cases}$$

where β_{ij} and η_{ij} are real parameters of the single directed edge (i, j).

As such, the transition probability from i to j depends both on the knowledge state K of the problem solver and on the initial state s_0 of the problem.

The alternative assumption (A2) establishes that a transition from the current state i to an adjacent state j still depends on the knowledge state K, but it is independent of the initial problem state s_0. Precisely:

(A2) For every problem (s_0, g) ∈ Q, every directed edge (i, j) and every knowledge state K ∈ K, the probability of a transition from i to j is

$$P(j|i, s_0, K) = P(j|i, K) = \begin{cases} \beta_{ij} & \text{if } (i, g) \in K, \\ \eta_{ij} & \text{if } (i, g) \notin K. \end{cases}$$

An observation concerning the two parameter types β_{ij} and η_{ij}, which holds under both assumptions (A1) and (A2) is as follows. If i is any problem state different from both g and f, and (i, f) is a directed edge, then the parameter $\beta_i f$ is the probability to observe a transition to the failure state: A kind of careless error. Conversely, if $j \neq f$, and (i, j) is a directed edge, then η_{ij} is the probability of a correct transition towards the goal: A kind of guessing.

Table 2 presents a classification of possible interpretations for all transition parameters based on Assumption (A1).

Table 2. Interpretation of β and η parameters on the basis of problem mastery and observed transition under (A1) assumption

Observed transition	$(s_0, g) \in K$: Mastered	$(s_0, g) \notin K$: Not mastered
(i,j): Correct	β_{ij}: Non-careless error	η_{ij}: Lucky guess
(i,j): Incorrect	β_{if}: Careless error	η_{if}: Non-lucky guess

Table 2 refers to the (A1) assumption, where the parameters' interpretation is based on the relationship between the observed transition (i, j) and the individual's mastery of the initial problem (s_0, g). Columns represent whether problem (s_0, g) is mastered ((s_0, g) ∈ K) or not ((s_0, g) ∉ K), while rows represent the observed transition during a single step of the solution paths, which can be either correct (i, j) or incorrect (i, f).

As the table shows, a correct transition performed by an individual who masters (s_0, g) is classified as a *non-careless err*or. Hence, β_{ij} is the probability of a non-careless error. Conversely, a correct transition performed by an individual who does not master the problem is classified as a *lucky guess*. Hence, η_{ij} is the probability of a lucky guess, implying that the correct response occurred by chance. An incorrect transition (i,*f*) made by an individual who masters the problem (s_0, g) is considered a *careless error*. Hence, β_{if} is the probability of a careless error. Finally, an incorrect transition made by an individual who does not master the problem is classified as a *non-lucky guess*. Hence, η_{if} is the probability of a non-lucky guess.

In Assumption (A2), mastery is assumed for the problem at the beginning of the observed transition, (i, g), rather than for the problem on which the entire solution process started, (s_0, g). Therefore, under

(A2), the η and β parameters receive an analogous interpretation, which is easily obtainable by replacing (s_0, g) with (i, g) in the heading of Table 2.

A monotonicity condition similar to the one introduced in Section 2.2, is required in the MSP models: If (i, j) is a directed edge with $i \neq f$, then η_{ij} is regarded as a guessing, whereas β_{ij} is regarded as a non-careless error, and the inequality $\eta_{ij} < \beta_{ij}$ is expected. The monotonicity condition can be formally stated as follows: given any two problem states $i, j \in S$,

$$\text{if } t \neq f \text{ then } \eta_{ij} < \beta_{ij}.$$

In general, under both assumptions, the transition probability of a directed edge (i, j) is either a β_{ij} or an η_{ij}. The only difference between assumptions (A1) and (A2) is that in the later one this transition probability depends on the current state i and not on the initial state s_0. Stated differently, while in (A1) assumption the transition probabilities of the solution process are globally determined by the overall problem (s_0, g) that has to be solved, in (A2) assumption, they are locally determined by the particular sub-problem (i, g) that still need to be solved in order to reach the goal. This distinction between the two processes has several implications, leading to different predictions in the models that are based on one assumption or the other.

As an example, the two diagrams displayed in Figure 3 show two directed graphs of the same goal space, with set of states $S = \{s_1, s_2, s_3, s_4, s_5, f, g\}$. In each graph, labeled circles represent the states of the problem space, whereas labeled edges represent moves, labeled with the names of the transition probability parameters of the MSPM. The scenario depicted in the figure reflects the case in which the problem solver's knowledge state is $K = \{(s_1, g), (s_2, g), (s_4, g)\}$, and the problem to be solved is (s_1, g).

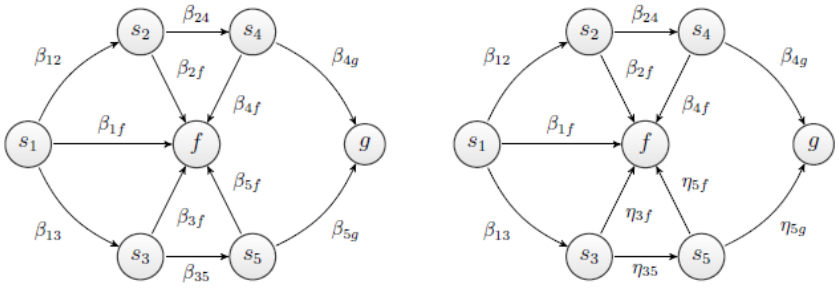


Figure 3: Goal spaces of Example 1, where problem (s_1, g) has two 3-move solution paths. On the left side the edges refer to transitions in the MSPM1 (pre-planning assumption), whereas on the right side they refer to transitions in the MSPM2 (interim planning assumption).

The left-hand side diagram is based on Assumption (A1), whereas the right-hand side diagram is based on Assumption (A2). Since the problem to be solved belongs to the knowledge state K of the problem solver, according to (A1) (left-hand diagram), all of the transition probabilities are of type β_{st}. This is not the case for (A2) (right-hand diagram). For instance, while the transition from s_3 to s_5 has probability β_{35} for (A1), it has probability η_{35} for (A2). This happens because problem (s_3, g) is not in K. Because of these differences, the same solution path may have different probabilities in the two models. For instance, the solution path passing through problem states s_1, s_3, s_5, g has probability $\beta_{13}\beta_{35}\beta_{5}g$ according to (A1), whereas its probability is $\beta_{13}\eta_{35}\eta_{5}g$ according to (A2).

Another example, similar to the previous one, is based on the two diagrams displayed in Figure 4. The scenario depicted in the figure reflects the case in which the problem solver's knowledge state is K = {(s_2, g), (s_3, g)}, and the problem to be solved is (s_1, g). This last does not belong to the state K of the problem solver. Assume that the problem is solved correctly. According to (A1) assumption, the sequences of moves will be a series of guesses and the probability of solving the problem is $\eta_{12}\eta_{23}\eta_{3}g$ (left-hand side diagram of Figure 4). Conversely, according to (A2) assumption, only the first move will be a guess because the remaining problems (s_2, g) and (s_3, g) belong to the state K of

the problem solver. Therefore, the solution path leading to the goal has a probability of $\eta_{12}\beta_{23}\beta_{3}g$ (right-hand side diagram of Figure 4).

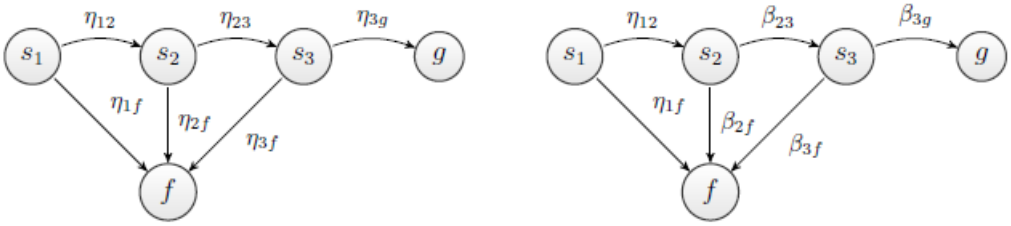


Figure 4: Goal spaces of Example 2, where problem (s_1, g) has a single 3-move solution paths. On the left side the edges refer to transitions in the MSPM1 (pre-planning assumption), whereas on the right side they refer to transitions in the MSPM2 (interim planning assumption).

In task such as the ToL test, problem solvers are required to plan in advance the solution path of each problem. If the problem solver is able to plan the solution path (i.e., the problem belongs to the state of the problem solver), in this example MSPM1 and MSPM2 will converge on the same predictions, which means that both are able to model preplanning strategies. This is not the case if the problem does not belong to the state of the problem solver (i.e., they are not able to plan in advance the entire solution path). Unlike MSPM1, MSPM2 disentangles knowledge states on the basis of which at least some of the subproblems can be solved from knowledge states on the basis of which none of them can be solved. Therefore, if interim-planning takes place (i.e., multiple planning instances occur at various steps throughout the solution process), only MSPM2 can model it.

2.6 Parameter estimation

The parameters of the MSPMs can be estimated by maximum likelihood via an adaptation of the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977). The details concerning the derivation of the estimation equations are provided in the Appendix section of the published paper. It should be observed that (unconstrained) maximization of the model's likelihood does not guarantee that the monotonicity conditions of the form given in (2) are satisfied. If they are violated for one or

more pairs of transition probabilities η_{ij} and β_{ij} (with $j \neq f$), then the interpretation of these last parameters as "lucky guess" and "non-careless error" is not credible anymore, and the interpretation of the whole model may become problematic. Furthermore, if the model parameters are applied for carrying out adaptive assessment, convergence to the true state of the problem solver is not guaranteed when the monotonicity conditions (2) are violated (see, in this respect, Brancaccio et al., 2023). For these reasons, a constrained maximum likelihood procedure was derived for both MSPM1 and MSPM2 models, where the constraints have the form of inequalities $\eta_{ij} < \beta_{ij}$.

The unconstrained estimation of the MSPM1 was derived by Stefanutti et al. (2021) and the MATLAB code can be found at https://osf.io/qa8mg/?view_only=8b4e148300de40a6941df4a102067fc1/.

Concerning the unconstrained estimation of the MSPM2 and the constrained estimation of both the MSPM1 and MSPM2, the corresponding algorithms are available at: https://osf.io/m658x/?view_only=25c8e23729554a1a84e74ed8e50d484b.

Once estimated and validated on an appropriate sample of individuals, both the MSPM1 and MSPM2 allow for the identification of the knowledge state of an individual based on their observed response pattern. For example, the posterior probability of a knowledge state K given an observed jump matrix J can be computed using Bayes' rule:

$$P(K|J) = \frac{P(J|K)P(K)}{\sum_{K' \in \mathcal{K}} P(J|K')P(K')}$$

The knowledge state that maximizes P(K|J) is taken as the most plausible estimate for the individual who produced the response pattern represented by J. This enables the assessment of individual differences. However, since P(J|K) takes on different forms under assumptions (A1) and (A2), the two models might assign different knowledge states to the same individual. Consequently, it is important to identify the best fitting model for an accurate assessment (Brancaccio et al., 2023).

3 Simulation study

The simulation study presented in this section has two aims. The first one is to test the parameter recovery of the MSPM2. It is worth mentioning that the parameter recovery of MSPM1 has been tested in Stefanutti et al. (2021). The second aim is to test the capability of model selection criteria to discriminate between MSPM1 and MSPM2. To pursue these aims, a simulation study is presented using the problem space of the ToL test.

3.1 Simulation design and data set generation

The present simulation study is based on the goal space $P = (S, f, g, \Omega, \cdot)$ obtained from the ToL problem space. It was defined according to two criteria: (i) the goal state is problem state 11; (ii) the solution path length ranges from 1 to 5 moves. The resulting goal space P contains 20 problem states plus the goal and the failure states. Thus, 20 problems can be derived from P, which are all the pairs (s, g) with $s \in S \setminus \{g, f\}$. Knowledge space K derived from P contained 1,573 knowledge states.

The data were generated under 12 different conditions displayed in Table 3. The 12 conditions result from the combination of the two assumptions (A1) and (A2) used to simulate the data, three sample sizes (300, 1,500, and 6,000), and two error levels. The three sample sizes were chosen to represent situations with small, medium, and large data sets, respectively, compared to 1,573, which is the cardinality of the structure. The two error levels were selected to represent low and medium-high error conditions.

Table 3: Design of the simulation study used for generating the data. Column one displays the condition number, column two displays the assumption underlying the generative model, columns three and four show the error interval used for generating the data, and column five displays the sample size (see Section 3.1).

Condition number	Assumption	β_{if}	η_{if}	Sample size
1	(A1)	(0,.1]	[.9,1)	300
2	(A1)	(0,.3]	[.7,1)	300
3	(A1)	(0,.1]	[.9,1)	1,500
4	(A1)	(0,.3]	[.7,1)	1,500
5	(A1)	(0,.1]	[.9,1)	6,000
6	(A1)	(0,.3]	[.7,1)	6,000
7	(A2)	(0,.1]	[.9,1)	300
8	(A2)	(0,.3]	[.7,1)	300
9	(A2)	(0,.1]	[.9,1)	1,500
10	(A2)	(0,.3]	[.7,1)	1,500
11	(A2)	(0,.1]	[.9,1)	6,000
12	(A2)	(0,.3]	[.7,1)	6,000

The error level in the data was manipulated through the β_{ij} and η_{ij} transition probabilities of the model. In particular, for each $i \in S \setminus \{f, g\}$, the $\beta_i f$ parameters were drawn at random from a uniform distribution with intervals (0, .1] and (0, .3] depending on the conditions. On the other hand, the $\eta_i f$ parameters were randomly drawn from a uniform distribution in the intervals [.7, 1) and [.9, 1). Different intervals were chosen for the $\beta_i f$ and $\eta_i f$ parameters because they have different interpretations. In particular, the $\beta_i f$ are interpreted as "careless errors" and, thus, they are expected to be small, whereas the $\eta_i f$ are interpreted as "non-lucky guesses" and, hence, they are expected to be close to one. The remaining β_{ij} and η_{ij} were drawn at random from a uniform distribution and then normalized to sum up to $1-\beta_i f$ and $1-\eta_i f$, respectively. For these reasons, conditions where $\beta_i f \in (0, .1]$ and $\eta_i f \in [.9, 1)$ represent a low error scenario, whereas conditions where $\beta_i f \in (0, .3]$ and $\eta_i f \in [.7, 1)$ represent a medium-high error scenario. Notice that, in each scenario, each pair (β_{ij}, η_{ij}) of parameters with $j \neq f$ also satisfies the inequality $\beta_{ij} - \eta_{ij} \geq 0$.

The data were simulated under the two different assumptions (A1) and (A2). Specifically, for each subject with knowledge state $K \in \mathcal{K}$ and for each problem (s, g), a sequence of problem states terminating with either the failure state or the goal state is generated using the β_{ij} and η_{ij} parameters. The procedure for simulating each response pattern follows the method described in Stefanutti et al. (2021, Fig. 5). The only variation between (A1) and (A2) is that, in the former case, the transition probabilities used to generate the sequence are all β_{ij} or η_{ij} if problem (s, g), respectively, belongs or does not belong to $\mathcal{K}$, whereas in (A2), each transition probability is a β_{ij} or η_{ij} if problem (i, g) belongs or does not belong to $\mathcal{K}$.

For each of the 12 conditions, 500 data sets were generated, obtaining a total of $12 \times 500 = 6{,}000$ simulated data sets. Every single data set consists of a given number of response patterns, each of which is a collection of jump matrices $J = \{J_q : q \in Q\}$, one for each of the problems in Q.

3.2 Methods

The identifiability of the MSPM1 and the MSPM2 has been tested using an informal method consisting of the following two steps: (i) The parameters of each model were re-estimated $n \geq 50$ times on the same data set, each time starting from a different initial set of parameters values, until 50 different estimations with equal likelihood were obtained; (ii) The standard deviations of the parameter estimates obtained in the 50 estimations were computed. It is expected that the standard deviation of the estimates is zero up to round-off when parameters are identifiable. Typically, this procedure is applied in those situations in which a formal test of (local) identifiability is not available (see e.g, Stefanutti, de Chiusole, Anselmi, & Spoto, 2020; Stefanutti et al., 2021).

In each of the 12 conditions, 500 data sets were simulated and both the MSPM1 and MSPM2 were fitted to the data. The goodness of the parameter recovery of the MSPM2 was tested in those conditions in which (A2) was the assumption under which the data sets were generated. In particular,

the bias and the variance of the parameter estimates across the 500 simulated data sets were computed. The estimated bias for each β_{ij} was computed as follows:

$$bias(\hat{\beta}_{ij}) = \bar{\beta}_{ij} - \beta_{ij},$$

where β_{ij} is the parameter value used for generating the data and $\bar{\beta}_{ij}$ is the arithmetic mean of the 500 estimates of the β_{ij} parameters. The standard deviation of the estimate $\hat{\beta}_{ij}$ was obtained as follows:

$$\sigma(\hat{\beta}_{ij}) = \sqrt{\frac{1}{500}\sum_{k=1}^{500}(\hat{\beta}_{ij}^{(k)} - \bar{\beta}_{ij})^2},$$

where $\hat{\beta}_{ij}^{(k)}$ is the estimated value of the transition probability for the k-th simulated sample. It is worth noticing that $\sigma(\hat{\beta}_{ij})$ is also the standard deviation around bias($\hat{\beta}_{ij}$). The bias and variance for $\hat{\eta}_{ij}$ were computed in the same way.

Moreover, as an index of the expected amount of information in the data to estimate the β_{ij} parameters, the "true" marginal problem probability was computed. Let $K_q = \{K \in \mathcal{K} : q \in K\}$ be the collection of all the knowledge states containing problem q. Then, for each $q = (i, g) \in Q$, the "true" marginal problem probability P(K_q) is given by

$$P(\mathcal{K}_q) = \sum_{K \in \mathcal{K}_q} P(K).$$

Similarly, the expected amount of information in the data to estimate the η_{ij} parameters is given by the probability

$$P(\mathcal{K}_{\bar{q}}) = 1 - P(\mathcal{K}_q).$$

For simplicity, both P(K_q) and P($\bar{K}_q$) are called "true" marginal probability of the problem all along the manuscript, because it is the probability of the states used to generate the data sets.

To test the capability of the model selection criteria to identify the assumption under which the data were generated, the Akaike information criteria (AIC; Akaike, 1973) computed for the two estimated models were compared to one another. The proportion p of times over the 500 replications in which the AIC selects the model that is consistent with the generative assumption was computed in all the conditions. For instance, a p = 1 in Condition 1 means that the AIC selects the MSPM1 for all the 500 data sets, whereas a p = 1 in Condition 7 means that the MSPM2 is selected for all the 500 data sets.

4. Results

In the MSPM1 the two parameters $\eta_{21,g}$ and $\eta_{21,f}$ may have identifiability problems, because an application of the two-step method illustrated at the beginning of Section 3.2 resulted in a standard deviation of .13. Concerning the MSPM2, no identifiability issues have been observed.

The following describes the parameter recovery of the MSPM2, which was estimated on the data sets simulated under assumption (A2). For the parameter recovery of the MSPM1 the reader can refer to Stefanutti et al. (2021).

Figure 5 displays the bias of the MSPM2 parameter estimates. The top panels correspond to the β_{ij} transition probabilities, while the bottom panels correspond to the η_{ij} transition probabilities. The left and right panels represent low and medium-high error conditions, respectively. In each panel, the pairs (i, j) representing the parameter estimates $\hat{\beta}_{ij}$ or $\hat{\eta}_{ij}$, with $i \in S$ and $j \in E(i)$, are along the vertical axis, while the bias of the parameter estimates is measured along the horizontal axis. For convenience, only the free transition probabilities are shown (e.g., if the constraint $\hat{\eta}_{i,j} + \hat{\eta}_{i,f} = 1$ holds then only $\hat{\eta}_{i,j}$ is displayed). In the condition with N = 300 and low error (Condition 7), the bias appears to be negligible for all parameters β_{ij} and most parameters η_{ij}. The only exceptions are parameters $\eta_{31,g}$ and $\eta_{21,g}$, which exhibit biases as large as .58 and .59. These biases might stem from insufficient information in the data, as explained below. Figure 6 displays the bias of the parameter estimates (vertical axis) as a function of the "true" marginal problem probability (horizontal axis). As in Figure

5, the top and bottom panels correspond to parameters β_{ij} and η_{ij} respectively, while the left and right panels represent low and medium-high error conditions.

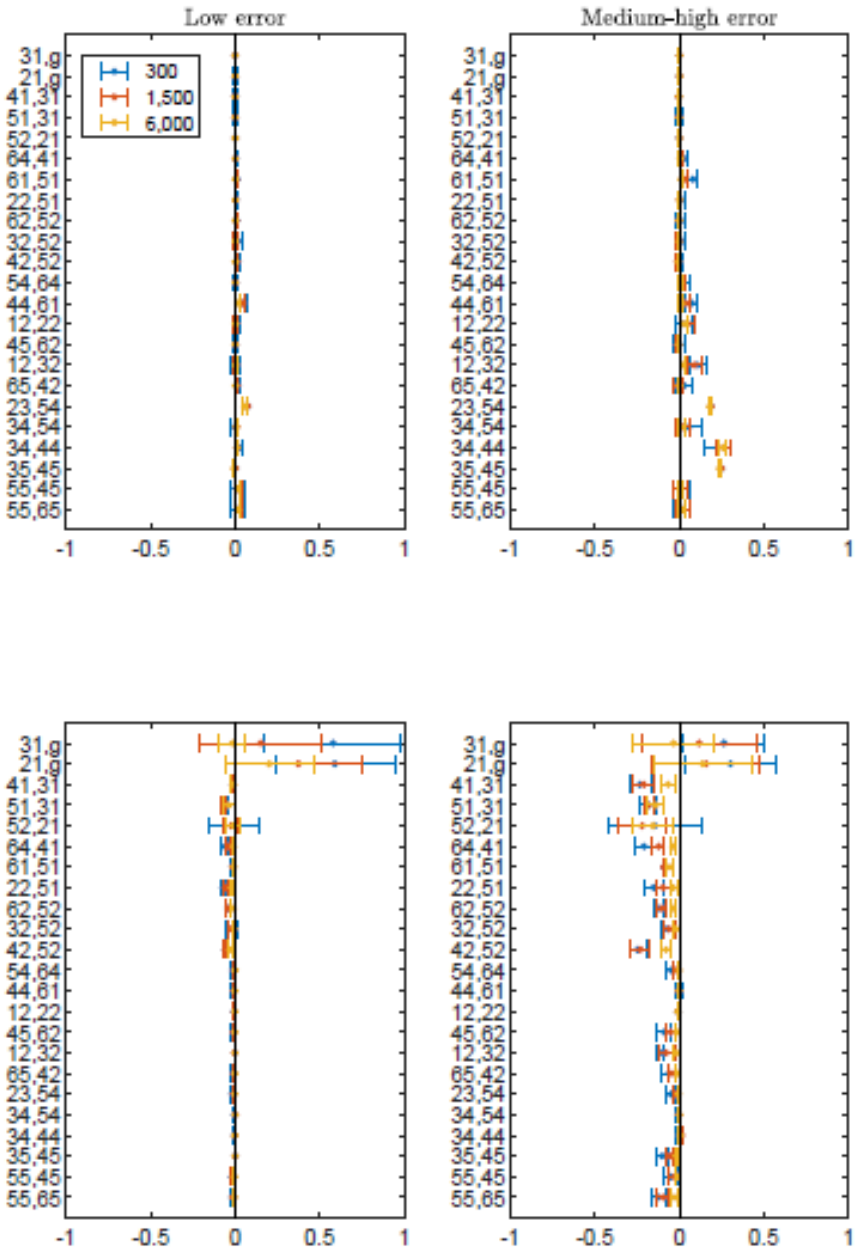


Figure 5: Parameter recovery of the MSPM2's β_{ij} (top panels) and η_{ij} parameters (bottom panels) across three sample sizes: 300 (blue), 1,500 (red), and 6,000 (yellow). In each panel, the pairs (i, j) representing the parameter estimates $\hat{\beta}_{ij}$ or $\hat{\eta}_{ij}$, with $i \in S$ and $j \in E(i)$, are along the vertical axis, while the bias of the parameter estimates is measured along the horizontal axis. The left and right panels represent low and medium-high error conditions, respectively.

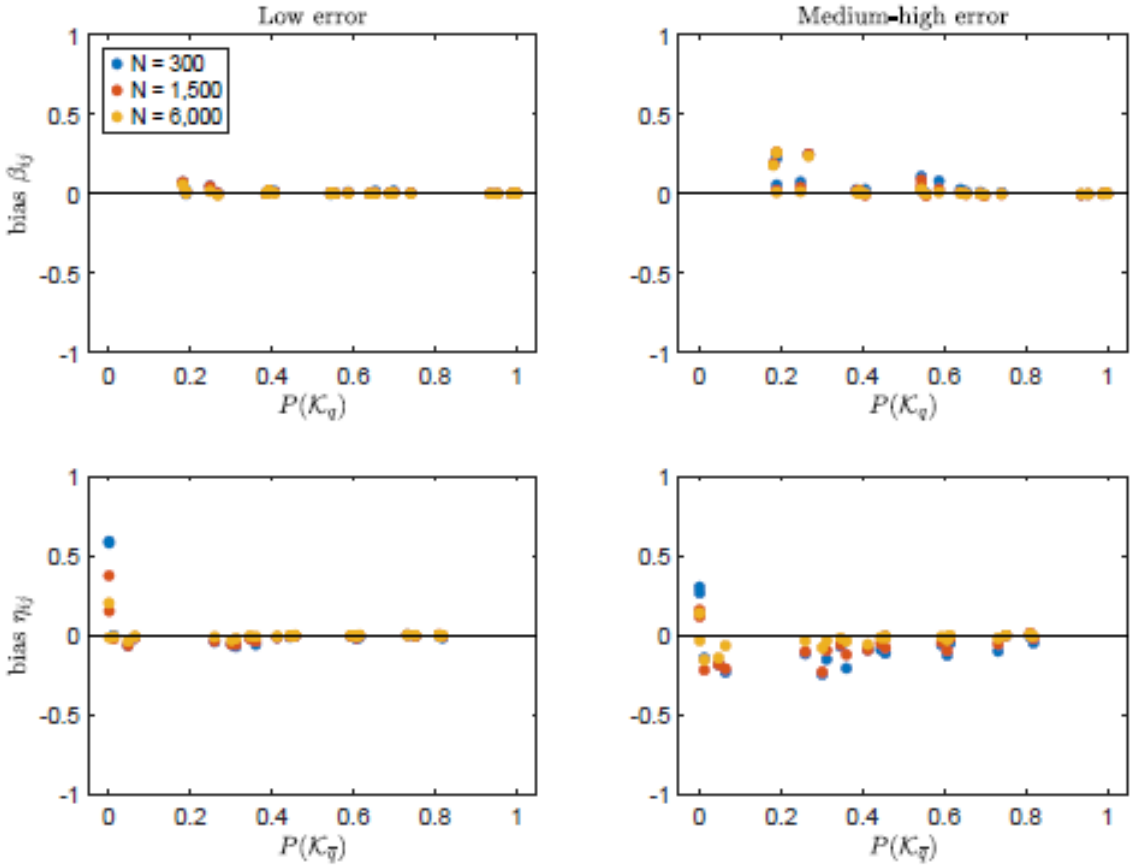


Figure 6: Bias of β_{ij} (top panels) and η_{ij} (bottom panels) as a function of their marginal problem probability (x-axis), across three sample sizes: 300 (blue), 1,500 (red), and 6,000 (yellow). The left panels correspond to low error conditions, while the right panels correspond to medium-high error conditions.

The transitions whose parameters have the largest bias are exactly those with the lowest marginal probabilities of being observed. For instance, $\eta_{31,g}$ and $\eta_{21,g}$ have marginal problem probabilities as low as .0012 and .0007, respectively (bottom left panel of Figure 6). In other words, the probabilities to simulate a knowledge state $K \in \mathcal{K}$ that does not include problems (31, g) or (21, g) are .0012 and .0007, respectively. As a result, in a data set with a sample size of 300, the expected frequencies for these transitions are $0.0012 \times 300 = 0.36$ and $0.0007 \times 300 = 0.21$, corresponding to less than one individual per transition.

As the sample size increases, the number of individuals contributing to the estimation of the parameters increases and the bias of the parameter estimates decreases (red and yellow data points in Figure 5). Two aspects are worth noting. First, even with samples sizes of 1,500 and 6,000, the number of individuals available for estimating parameters $\eta_{31,g}$ and $\eta_{21,g}$ remains rather small (i.e., 1.8 and 7.2 for $\eta_{31,g}$; 1.05 and 4.20 for $\eta_{21,g}$). Second, these two parameters correspond to transitions that are one step away from the goal state. In the context of the ToL task, these transitions almost never failed in the general population. Specifically, the probability that an individual fails a one-move problem is close to zero, making such transitions highly infrequent.

Concerning the medium-high error conditions, with the sample size held fixed, the biases of the parameter estimates increase. Interestingly, however, the bias is slightly higher across all transitions, even though the maximum bias is smaller compared to the condition with a smaller sample size.

Concerning conditions from 7 to 12, Table 4 displays the mean absolute bias for $\hat{\beta}_{ij}$ and $\hat{\eta}_{ij}$ calculated across all transitions (second and third columns), and their mean standard deviation (fourth and fifth columns). As expected the biases are smaller in the low error conditions (i.e., 7, 9, and 11), rather than the high error conditions (i.e., 8, 10, and 12). It is also worth mentioning that in all conditions bias($\hat{\eta}_{ij}$) is bigger than the bias($\hat{\beta}_{ij}$). The last result can be explained by looking at Figure 6 where it is clear that there is more information for estimating the β_{ij} parameters than for estimating the η_{ij} parameters. Indeed, even for the β_{ij} parameters with the least information, there is a marginal problem probability of about .20, whereas for the η_{ij} parameters with the least information, there is a marginal problem probability of about zero.

Table 4: Bias obtained in the simulation study for MSPM2 under each condition of sample size and error level. Columns 2 and 3 display the mean absolute bias for $\hat{\beta}_{ij}$ and $\hat{\eta}_{ij}$, respectively, while Columns 4 and 5 show the mean values of the corresponding standard deviations.

Condition	mean $\lvert bias(\hat{\beta}_{ij})\rvert$	mean $\lvert bias(\hat{\eta}_{hj})\rvert$	mean $\sigma(\hat{\beta}_{ij})$	mean $\sigma(\hat{\eta}_{hj})$
7	.012	.072	.015	.051
8	.050	.117	.030	.064
9	.009	.041	.007	.042
10	.044	.088	.019	.059
11	.006	.017	.005	.024
12	.035	.041	.010	.043

Finally, in all conditions, the proportion p of instances in which the AIC selects the model consistent with the assumption that generated the data is 1, with the sole exception of Condition 8, where p = 0.996. It is worth noting that, since MSPM1 and MSPM2 have the same number of parameters, comparing their AIC values is equivalent to directly comparing their likelihoods.

5. Discussion

Overall, MSPM2 demonstrates good parameter recovery performance, particularly for the majority of transitions. Even under relatively small sample sizes (i.e., N = 300), estimates for both β_{ij} and η_{ij} parameters are generally unbiased, with deviations emerging only in those transitions where failure is intrinsically rare (i.e., with very low variability in the data). Importantly, the two parameters susceptible to bias – namely η_{31},g and η_{21},g – refer to transitions toward the goal state in single-move problems, which are of little behavioral importance as these transitions are very rarely failed in the general population. As such, these aspects do not compromise the interpretability or practical utility of the model.

A potential limitation of this simulation study should also be mentioned. Indeed, data sets were obtained by using only 12 generating parameter sets. While this ensures internal consistency, it could not adequately capture the full variability of parameters encountered in real-world contexts. Future work could address this by employing a broader and more systematically varied set of generating conditions to enhance generalizability.

6. Empirical application

The literature on the development of planning ability highlights that there are some transition ages in which the performance at tower tasks and the type of planning employed may change. Specifically, children aged 4 to 8 seem to have lower performance than older children (McCormack & Atance, 2011b), with some authors suggesting that 4-year-old children have qualitatively different response processes (Kaller, Rahm, Spreer, Mader, & Unterrainer, 2008; McCormack & Atance, 2011b). Whereas, Asato et al. (2006a) report that from 8 years and older, the pre-planning time starts to increase, reaching adult-level performance around 14 to 15 years old (Luna et al., 2004).

In the previous section, a suggestion was made that the assumption behind MSPM1 represents a response behavior in which the planning process occurs predominantly at the onset of the solution process. Conversely, the assumption behind MSPM2 would represent a response behavior where multiple planning instances occur at various steps throughout the solution process.

Merging the insights gained from the literature on the development of planning ability with those of MSP models, the following three hypotheses seem plausible:

- (H1) Children younger than 8 years old do not preplan their responses but they might resort to an interim planning. Therefore, the MSPM2 should better fit the data than MSPM1;
- (H2) Children aged 9 to 13 begin to preplan their moves, although they may still demonstrate inconsistencies in their strategy. Therefore, both MSPM1 and MSPM2 are expected to fit the data;
- (H3) Individuals older than 14 years old should consistently preplan their responses, therefore, MSPM1 should better fit the data than MSPM2.

The analysis of the response times should also bring evidence in favor of these hypotheses. In fact, the preplanning time is expected to be longer when the problem solution is planned in advance (i.e., before moving) and it is expected to be shorter when the problem is solved with an interim-planning

strategy (i.e., between one move and the other). A response time analysis can serve as an external criterion for assessing the validity of the models.

7. Materials and Methods

7.1 The assessment tool: Adap-ToL

This study is part of a research project founded by the Italian Ministry of Education and Research aimed at the development of an intelligent system, named PsycAssist (de Chiusole et al., 2024), for the adaptive assessment of executive functions and fluid intelligence among general and clinical populations. The study here presented shows the results obtained from the application of the MSPM1 and MSPM2 to the data collected with one of the tools available on PsycAssist, which is Adap-ToL.

Adap-ToL is a test for the assessment of planning skills in individuals aged four and above. It resembles the ToL test, where participants move beads from an initial setup to a goal configuration, within a set number of moves and following specific rules. Unlike the traditional ToL test, where the initial state is constant among the problems but the goal state varies, Adap-ToL maintains a fixed goal state while varying the initial configuration. This was designed to maintain coherence with the PKST definition of a problem in a goal space.

Different versions of the test were designed. Among all the 35 problems having the same tower configuration (i.e, configurations in the first row of Figure 1) as the goal state, those with a number of moves 1 to 5 were used for creating a version of the test designed for individuals aged 4 to 8; those of a number of moves 1 to 6 were used for creating a version of the test designed for individuals aged 9 to 13; and all the 35 problems were used in the version for individuals older than 14.

The test is administered by using a tablet (IoS operating system with a 10.9in screen) oriented vertically. Each problem of the test consists of two images, one above the other (Figure 7). The upper image represents the configuration of the goal state, whereas the lower image represents the configuration of the initial state of the problem. The task is to replicate the goal state from the initial

configuration, within a specified minimum number of moves (specified at the top of the upper image). Acting on the lower image, participants tap the screen twice to make a move: first to select the bead, and then to choose the peg in which to move the bead. The problem is successfully solved when the goal state is reached within the minimum number of moves.

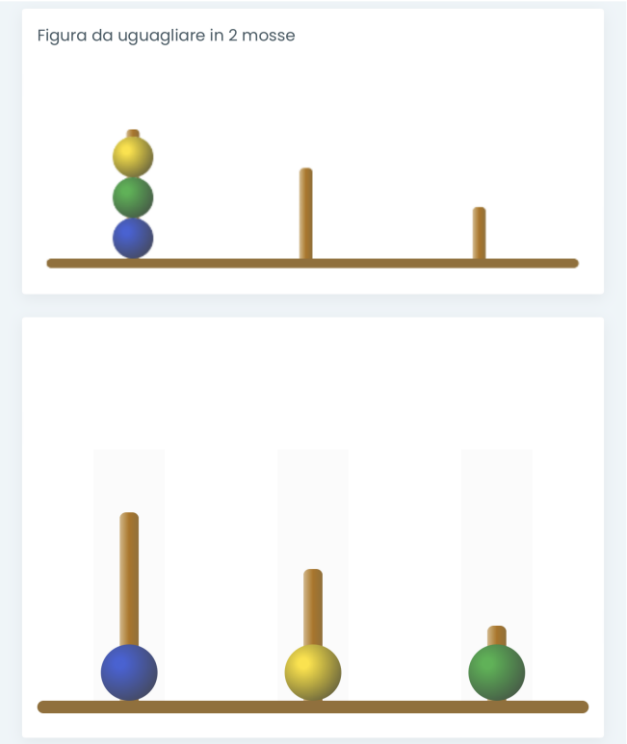


Figure 7: One of the problems in Adap-Tol. Each problem of the test consists of two images, one above the other. The upper image represents the configuration of the goal state, whereas the lower image represents the initial configuration of the problem. The task is to replicate the goal state from the initial configuration, within a specified minimum number of moves (specified at the top of the upper image). In the depicted configuration, “Figura da uguagliare in 2 mosse” translates to “Figure to be matched in 2 moves”). Acting on the lower image, participants tap the screen twice to make a move: first to select the bead, and then to choose the peg in which to move the bead.

The test is introduced by a brief animated video (95 seconds), which presented the test, explained the task, and gives the following two rules: (a) the highest peg can hold three beads, the central peg can hold a maximum of two beads, and the shortest peg can hold only one bead; (b) a bead cannot be

moved if it has another bead on its top. The system gives to the test-taker a feedback in case of the violation of one rule, an excessive number of moves made to complete the task, and when the task is correctly executed. Each feedback is presented within a colored box (red, orange, or green, respectively) by using a simple text and an emoticon. For children who can not read yet or who may face challenges in reading, the evaluator reads the message aloud.

After the video presentation and before the test administration, the test-taker were presented with four practice problems. If the test-taker struggled to find any answer to a problem, the evaluator could decide to skip the problem and move to a next one. It is worth noticing that, the decision to skip a problem was made solely by the evaluator and only when the test-taker was unable to reach the goal configuration and showed signs of frustration. The problems were presented one at a time in random order.

As in traditional ToL tests, the "pre-planning time" (i.e., the time elapsed between the presentation of the problem and the first click), the "execution time" (i.e., the time elapsed between the first click and the completion of the problem), the "total time" (i.e., the sum of the pre-planning and execution times), and the accuracy of each problem were recorded. Concerning accuracy, the problem is considered correctly solved when the goal state is reached within the minimum number of moves; otherwise, it is wrong.

Differently from other ToL tests, where at maximum three attempts are given to solve a problem, Adap-ToL prescribes only one attempt. Moreover, Adap-Tol registers the whole solution path made by a problem-solver. This aspect is fundamental for applying MSPM1 and MSPM2.

The computerized version of the test allows for registering the time at each click made by a test-taker. Consequently, other two types of sub-times can be registered, that is: (i) the "move-execution time", which is the time spent moving a bead, is computed as the time elapsed from the click on a bead to the click on a peg, and (ii) the "move-planning time", which is time spent selecting a new bead to

move, is computed as the time from the click on a peg to the click on a new bead. Figure 8 shows how the entire solution process of a 2-move problem is decomposed among the "traditional" pre-planning and execution times and the new move-execution and move-planning times.

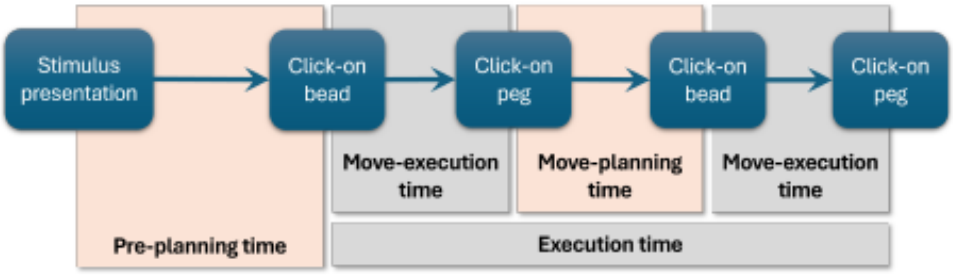


Figure 8: Decomposition of the solution process of a 2-move problem in the "traditional" pre-planning and execution times and the new move-execution and move-planning times.

7.2 Participants

Adap-Tol was administered to a sample, extracted from the general population, of N = 1,466 participants (Female = 52.46%, Age range: 4.20 – 71.05 years, Mean age: 19.14±16.35 years). Data were collected in different regions of the North, Centre, and South of Italy (i.e., Abruzzo, Basilicata, Calabria, Emilia-Romagna, Lazio, Liguria, Lombardia, Marche, Puglia, Toscana, Umbria, Veneto) and schools of all grades as well as adults in the general population were involved. Before participating in the study, participants and/or their parents received detailed information about the study purposes and procedure, and they (or their parents/legal guardians) provided a signed informed consent, in accordance with the declaration of Helsinki.

Data from 69 people with a diagnosis (e.g., ADHD, autism spectrum disorder, intellectual disability) were removed. The removal of individuals with a diagnosis is due to the fact that this study focuses on assessing the sensitivity of the two models to age-related differences in the general population. In fact, including individuals with specific difficulties could bias the results, potentially making older participants perform more similarly to younger ones. No problem were skipped, thus the data set does

not have any missing answers. Moreover, a maximum age limit was set at age 57. Thus, data from 105 participants aged 57 and above were excluded. This age limit was chosen to (i) prevent performance decline due to age-related factors and (ii) ensure a sample size comparable to the other age groups. The final sample was composed of 1,053 respondents (Female = 51.85%, Age range: 4.41–56.97 years, Mean age: 16.62 ± 11.50 years). Age group 4 to 8 was composed of 269 children (Female = 45%, Mean age: 7.08 ± 1.36), age group 9 to 13 was composed of 343 children (Female = 51%, Mean age: 11.70 ± 1.51), age group 14 to 57 was composed of 441 individuals (Female = 56.70%, Mean age: 26.20 ± 12.10).

A two-proportion z-test was used to investigate any significant difference in the proportion of correct responses between female and male children in each age group (Type I error probability $\alpha = .05$). The proportion of correct responses were computed considering the twenty problems that were presented to all respondents regardless of their age. The Benjamini-Hockberg (Benjamini & Hochberg, 1995) correction for multiple comparison was applied. The results are reported in Table 5. After correcting the p-values for multiple comparison, the difference in the proportion of correct responses in female and male participants remained significant in the 4 – 8 group and in the 30 – 57 group. In the former case, female participants performed significantly better than male participants, while the opposite effect was observed in the latter case.

Table 5: Proportion of correct responses of female (F) and male (M) participants in each age group. The *p*-values reported in the last column are already corrected according to the Benjamini-Hockberg correction.

Age group	F (*SD*)	M (*SD*)	*p*-value
4 – 8	.89 (0.09)	.85 (0.12)	< .001
9 – 13	.92 (0.07)	.92 (0.08)	.68
14 – 29	.91 (0.08)	.91 (0.08)	.76
30 – 57	.88 (0.10)	.92 (0.07)	< .001

7.3 Data Analysis

For maximizing the number of problems shared by the three age groups, only problems with 1 to 5 moves were considered.

Both MSPM1 and MSPM2 were fitted to the data of each age group by using knowledge space K and goal space P of the ToL introduced in Section 3.

Hypotheses (H1), (H2), and (H3) were tested by comparing both the absolute and the relative goodness-of-fit of the models. The absolute goodness-of-fit of the models was tested by the Pearson Chi-square statistic. In particular, a parametric bootstrap procedure (Efron, 1979) over 500 replications was used to compute the p-value of the Chi-square. When both models fitted the data, the AIC index and Bayes Factor were used to select the best model. The Bayes factor was approximated using the Bayesian information criterion (BIC; Schwarz, 1978).

Concerning the response time analysis, the entire time spent to complete a problem was split into three sub-times. The first one is the "pre-planning time", which is the time spent from the presentation of the task to the first click on a bead. The second one is the "move-execution time", which is the time spent from the click on the bead to the click on the peg. The third one is the "move-planning time", which is the time spent from the click on a peg to the click on a new bead. It follows that the sum of these three times equals the total time spent to complete the problem. Thus, for each participant, a standardized version of the three sub-times can be obtained by dividing each of them by the total time. Only the response times of problems correctly solved were considered in this analysis. Moreover, average values across test-takers were obtained separately for each age group.

Standardizing sub-times in this way allows for obtaining indexes whose range is in the interval [0, 1], and that express the average proportion of time spent on the specific "action". Since the total time spent to solve a problem strictly depends on the number of moves, these indexes might be misleading if used for comparing the response behavior of the same group (as well as of different groups) to

different problems. Nonetheless, these indexes should provide useful insights when used for comparing the response behavior of different age groups while keeping the problem fixed.

8. Results

The MSPM1 obtained a strong rejection (bootstrapped p-value = .004), whereas the MSPM2 adequately fit the data with a bootstrapped p-value = .312 on the group aged 4 to 8.

Concerning age group 9 to 13, a first attempt to fit the models to the data obtained a strong rejection for both models, with bootstrapped p-values < .005. To investigate the lack of fit of the models, the standardized residual associated with the Chi-squared statistics was computed. In both the MSPM1 and MSPM2 the same response pattern had the largest standardized residual (i.e, 10.0 for the MSPM1 and 10.6 for the MSPM2). Extreme outliers can distort the overall model fit, especially when using goodness-of-fit measures like Chi-square. Such a response pattern can be considered as an extreme outlier. Indeed, an additional qualitative check of the response pattern revealed frequent violations of the model's expected difficulty hierarchy: easier problems (requiring one to three moves) were failed, while a more complex seven-move problem was solved. In addition, most failures occurred at the final step before reaching the goal. Therefore, it was excluded from the analysis. This decision follows standard practices to ensure a reliable evaluation of the model parameters (Rousseeuw & Leroy, 1987).

After removing the response pattern, both MSPM1 and MSPM2 exhibit a good fit to the data, with bootstrapped p-values of .210 and .222, respectively. Concerning the model comparison, the AIC index of MSPM1 is 7,536, while that of MSPM2 is 7,582. Thus, MSPM1 is the selected model for individuals aged 9 to 13. The Bayes factor also shows evidence in favor of MSPM1 with a value of $BF_{12} = 7.6 \times 10^9$.

Finally, in age group 14 to 57 both MSPM1 and MSPM2 fit the data well with bootstrapped p-values of .138 and .144, respectively. The AIC comparison selects MSPM1 with an AIC statistic of 9,526,

compared to 9,620 for MSPM2. The Bayes factor shows evidence in favor of MSPM1 with a value of $BF_{12} = 3.6 \times 10^{20}$.

Figure 9 shows the results of the response time descriptive analysis obtained with the three age groups (x-axis in each panel). The y-axis of each panel gives the proportion of the total time spent to complete a problem. Rows display the results obtained with problems having a different number of moves, going from 1 (top panel) to 5 (bottom panel). Pre-planning time is in blue, move-execution time in yellow, and move-planning time in red. The three standardized sub-times are represented by stacked bar graphs (panels at the left) and line plots (panels at the right) with error bars representing standard deviations. Stacked bar graphs can be useful for analyzing how participants of the three age groups distributed the response time among the three actions. Whereas, line plots can serve for studying the trend of each action across age groups.

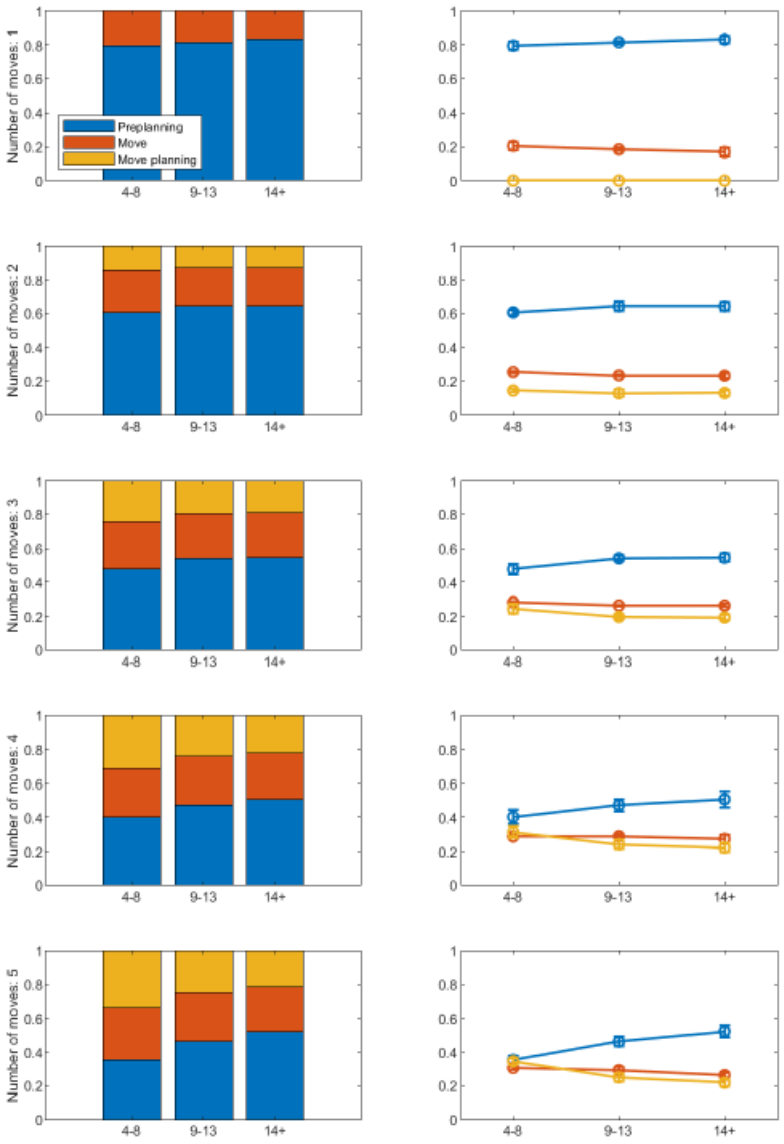


Figure 9: Results on the response time descriptive analysis. See text for more details.

In problems with a smaller number of moves (i.e., 1 and 2) a similar response time behavior was observed among the three age groups. In particular, most of the total time is spent on pre-planning. It is worth noticing that in 1-move problems the proportion of the total time spent in pre-planning time and that spent in Move planning coincide (this is why this last does not appear in the graph). Concerning the trend across age groups, a slightly monotonic increase was observed for the proportion of the total time spent in pre-planning time, even if with a small amount (the gain is .04 in both 1-move and 2-move problems).

Instead, the response time behavior seems different across age groups when the number of moves increases. In 5-move problems, despite the proportion of time spent is executing the move is quite similar in the three groups (the biggest drop is .02), the proportion of time spent on pre-planning increases (with a gain of .17) and the proportion of time spent in move-planning decreases (with a drop of .13) as age increases. Moreover, it seems that children aged 4 to 8 spent a proportion of time in pre-planning and in move-planning that is very similar (i.e, .35 and .34, respectively). Conversely, individuals older than 14 spent a proportion of time in pre-planning and in move-planning which is quite different, with the former greater than the latter (i.e., .52 and .21 respectively). In summary, the outcomes derived from both the response time and solution process analysis support Hypotheses (H1), (H2), and (H3).

9. Discussion

MSPM1 and MSPM2 are based on different assumptions about how planning occurs during problem-solving. MSPM1 assumes that planning takes place predominantly at the onset of the solution process, leading to a more rigid execution of an initial strategy. In contrast, MSPM2 allows for multiple instances of planning throughout the task, supporting a more flexible and adaptive approach.

The superior fit of MSPM2 in the 4 to 8 age group suggests that younger children do not rely solely on an initial plan but rather adjust their strategy dynamically as they progress. In contrast to this age group, the results for older participants indicate a shift in planning behavior. For individuals aged 9 to 13, both MSPM1 and MSPM2 exhibit a good fit to the data, but model comparison criteria favor MSPM1. The lower AIC value and the strong Bayes factor in favor of MSPM1 suggest that, in this age range, planning tends to occur primarily at the onset of the solution process, with less reliance on dynamic adjustments during task execution. A similar pattern emerges in the 14 to 57 age group, where MSPM1 remains the preferred model based on AIC and Bayes factor comparisons.

These findings align with the idea that, as individuals develop greater cognitive control and experience with problem-solving, they rely more on pre-planned strategies and execute solutions with reduced need for mid-task adjustments. This developmental trend supports the interpretation that younger children engage in more iterative and corrective planning, whereas older individuals increasingly favor an upfront, goal-directed approach to problem-solving.

10. General Discussion

Beyond providing the algorithms for the maximum likelihood estimation of the model based on the interim-planning, this manuscript brings evidence in favor of the usefulness of considering the sequences of moves undertaken by the respondents in solving tower tasks.

Considering only the accuracy performance at tower tasks might be reductive and might not allow for getting a comprehensive understanding of the cognitive processes involved in the completion of the task. By capitalizing the information provided by the sequences of moves, the approach presented in this contribution allowed for the distinction between respondents of different age who employ different strategies for solving tower tasks.

The results of the empirical application showed that the model based on the interim-planning assumption best fitted with respondents aged 4 to 8, while that based on the pre-planning assumption best fitted with respondents aged 14 on. These results are consistent with cognitive theories asserting that young children have not yet entirely developed the executive functions related to planning (Asato et al., 2006a; Kaller et al., 2008; Luna et al., 2004; McCormack & Atance, 2011b).

Interestingly, both models are plausible given the data of respondents aged 9 to 13. This result might be explained considering both the intra- and inter- individual differences. From the intra-individual perspective, the same individual might change their solution strategy depending on the difficulty of the problem. For instance, they might use a pre-planning strategy for problems with a lower number of moves, while they might use an interim-planning strategy when facing problems with a higher

number of moves. From the inter-individual perspective, the respondents in this age group might differ from one another with respect to their solution strategy, presenting a high degree of heterogeneity. Future studies employing a think-aloud protocol might provide further evidence about the planning strategies used by the individuals in this age group. In this light, a new model able to combine the pre-planning and interim-planning assumptions would be useful to better understand the potential occurrence of the shifting between planning strategies. Another possibility to extend the models considered so far would be to consider as knowledge states subsets of solution paths rather than subsets of problems.

This manuscript demonstrated the usefulness of the two Markov solution process models for both the accuracy and the observed sequences of moves in the discrete time. The preliminary results on time performances suggest that a promising future direction could be the development of Markov models in the continuous rather than discrete time, in analogy to Anselmi, Stefanutti, de Chiusole, and Robusto (2021).

In summary, the models based on the pre-planning and interim-planning assumptions are able to discriminate between respondents of different ages and provide insights into the cognitive processes involved in the solving of tower tasks.

Study 5

Error Patterns on a Computerized Version of Raven's Progressive Matrices in Specific Learning Disorders (SLD)

Spinoso, M., Benassi, M., Riccardi, A., Mazzoni, N., Orsoni, M., Brancaccio, A., Epifania, M. O., Muccinelli, M., Novelli, C., de Chiusole, D., Anselmi, P., Stefanutti, L., Bacherini, A., Pierluigi, I., Balboni, G., & Giovagnoli, S. (in preparation). *Error patterns on a computerized version of Raven's Progressive Matrices in specific learning disorders.*

Abstract

Specific Learning Disorders (SLD) are neurodevelopmental conditions characterized by persistent difficulties in reading, writing, or mathematics despite average intellectual functioning (American Psychiatric Association, 2013; World Health Organization, 1992). However, intelligence is multidimensional, and emerging research on SLD reveals heterogeneity in SLD cognitive profiles and subtle weaknesses in working memory (WM) and processing speed, which are closely intertwined with Fluid Intelligence (FI). FI, defined as the ability to think logically, identify patterns, and solve novel problems independently of prior knowledge, plays a fundamental role in learning and academic achievement (Cattell, 1963, 1971, 1987; Green et al., 2017). Existing research has not found evidence that SLD exhibits lower levels of FI. However, most studies have focused exclusively on quantitative measures (total accuracy or global scores), which may obscure underlying cognitive processes. The qualitative dimensions of FI - error types, and problem-solving strategies- remain largely unexplored in this population. However, examining error patterns can reveal subtle inefficiencies in reasoning and executive processes that quantitative scores alone cannot capture, thereby providing deeper insight into cognitive mechanisms contributing to learning difficulties. To address this gap, the present study investigated FI accuracy and error patterns using a new computerized version of Raven's Progressive Matrices: MatriKS (de Chiusole et al., 2024). A sample of 106 participants aged 7–17.11 years (46 with SLD, 60 typically developing [TD]) completed MatriKS alongside standardized cognitive and academic assessments. Consistent with prior research, no significant group differences emerged in overall accuracy, although SLD participants consistently obtained lower scores compared to TD participants. Qualitative error analysis revealed distinct patterns: younger children with SLD committed more errors than TD peers, particularly in Incomplete Correlate and Repetition categories, suggesting vulnerabilities in visuospatial WM and executive control. Age effects on error patterns emerged only in younger children, consistent with

developmental models of FI, while no significant differences were found in adolescents. These findings show that while global FI in SLD is preserved, specific error patterns characterize SLD. Thus, qualitative error analysis uncovers specific FI inefficiencies contributing to academic challenges. In this context, the computerized assessment format of MatriKS offers advantages for capturing nuanced performance patterns through automated recording. These findings offer valuable clinical implications, suggesting that incorporating qualitative error analysis into cognitive assessment enhances understanding of individual SLD cognitive profiles and provides insights for designing targeted interventions addressing underlying processing vulnerabilities rather than solely academic skill deficits.

Keywords: fluid intelligence, specific learning disorders, Raven's Matrices, computerized assessment, error analysis

1. Introduction

Following DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, 5th edition; American Psychiatric Association [APA], 2013) and ICD-10 (International Classification of Diseases, 10th revision; World Health Organization [WHO], 1992) diagnostic criteria, individuals with Specific Learning Disorders (SLD) have average global intellectual functioning. Nonetheless, given that intelligence is a multidimensional construct comprising several interrelated components, numerous researchers (e.g., Cornoldi et al., 2014, 2019a, 2019b; De Clercq-Quaegebeur et al., 2010; Giofrè & Cornoldi, 2015; Iaia et al., 2025; Poletti, 2016; Scorza et al., 2023; Toffalini et al., 2017) have investigated whether specific cognitive profiles or patterns can be identified in individuals with SLD. Among the various dimensions of intelligence that are crucially involved in learning, a core component is Fluid Intelligence (FI), which refers to the ability to reason logically and solve problems in novel situations, independently of prior knowledge (R. B. Cattell & Cattell, 1987). According to Cattell's Investment Theory (Cattell, 1963 1971, 1987), Fluid Intelligence contributes to the development of crystallized intelligence and to the acquisition of academic skills. The results of several experimental studies have supported this claim, showing that fluid reasoning abilities predict achievement in various domains, including reading and mathematics (e.g., Green et al., 2017; Johann et al., 2020; Passolunghi et al., 2022; Peng et al., 2019).

The role of FI in supporting learning has also been investigated in studies focusing on neurodevelopmental disorders, as these populations often exhibit deficits in FI and other cognitive processes essential to learning, such as executive functions (Semrud-Clikeman, 2012; Tamm & Juranek, 2012; Thorell, 2007). For instance, some studies regarding Attention-Deficit/Hyperactivity Disorder (ADHD) (e.g., Droder et al., 2024; Mano et al., 2018) have found both direct and indirect effects of FI on phonemic decoding and word recognition, suggesting that reduced fluid reasoning abilities may hinder reading acquisition in this population.

Despite this evidence on the importance of FI for learning, literature on the development of fluid abilities in children with learning disorders is lacking. A first insight into this topic comes from studies on the performance of individuals with SLD on intelligence tests, especially on the Wechsler Scales (e.g., WISC-IV [Wechsler, 2003]; WAIS-IV [Wechsler, 2008]). Indeed, the WISC-IV and WAIS-IV batteries are some of the most widely used tools for assessing intellectual functioning (Evers et al., 2012), given that they evaluate different dimensions of intelligence through their four indices of verbal comprehension (VCI), perceptual reasoning (PRI; mainly assessing non-verbal fluid abilities), working memory (WMI) and processing speed (PSI). Most studies on SLD (e.g., Cornoldi et al., 2014, 2019a; Iaia et al., 2025; Poletti, 2016; Scorza et al., 2023; Toffalini et al., 2017) have focused on the discrepancy between the General Ability Index (GAI; derived from VCI + PRI) and the Cognitive Proficiency Index (CPI; derived from WMI + PSI), showing that, in both children and adults with SLD, the former index typically falls within the average range, whereas the latter is usually lower than in controls. Thus, this discrepancy has been suggested as a hallmark of learning difficulties, characterized by average intellectual abilities, especially in the visuo-perceptual domain, and deficits in working memory and processing speed (Cornoldi et al., 2014, 2019a; Iaia et al., 2025; Poletti, 2016; Scorza et al., 2023; Toffalini et al., 2017).

However, SLDs are heterogeneous disorders, and individuals with learning difficulties frequently show considerable variability in their cognitive profiles and in the combination of deficits they present (Cornoldi et al., 2019b; McGrath et al., 2019; Menghini et al., 2010; Pennington, 2006; Pennington et al., 2012; Poletti, 2016; Toffalini et al., 2017; Willcutt et al., 2013). In line with this, Cornoldi & colleagues (2019b) identified two distinct subgroups among children with SLD based on WISC-IV profiles: a larger group with a lower Verbal Comprehension Index than Perceptual Reasoning Index, and another showing the opposite pattern. Interestingly, children with lower visuo-perceptual skills were also weaker in processing speed (Cornoldi et al., 2019b). This is not surprising, given that

processing speed, as well as working memory, has been shown to be closely related to the development of FI (Fry & Hale, 1996, 2000). Interesting results have also emerged when considering the specific domain of academic impairment (e.g., Poletti, 2016; Scorza et al., 2023; Toffalini et al., 2017). For instance, Toffalini & colleagues (2017), by comparing the WISC-IV profiles of children with different subtypes of SLD (i.e., reading, spelling, arithmetical, and mixed disorder, based on the ICD-10 classification; World Health Organization, 1993, cited in Toffalini et al., 2017), found not only shared characteristics across the different disorders, such as the GAI-CPI discrepancy, but also partially distinct patterns upon further comparison of the four main indexes. In particular, by comparing the Perceptual Reasoning Index (PRI) and the Verbal Comprehension Index (VCI), they found that the PRI was higher than the VCI in children with reading disorder, lower than the VCI in those with arithmetical disorder, and similar to the VCI in children with spelling disorder; regarding mixed disorder, even though the PRI-VCI pattern did not significantly differ from that in reading disorder, it nonetheless exhibited the weakest overall intellectual profile compared to the other three groups (Toffalini et al., 2017). Similar differences in the WISC profiles, particularly between reading and mathematical disorders, have emerged in the study by Poletti (2016), who found that children with different subtypes of SLD (based on the DSM-5 [APA, 2013] and the Italian guidelines [Consensus Conference, 2007; DSA, 2011, cited in Poletti, 2016]) showed different cognitive fragilities: those with mathematical impairment (isolated or combined with other learning impairments) had deficits not only in working memory and processing speed, but also in perceptual reasoning, whereas children with isolated reading impairment had deficits only in processing speed.

In addition to the Wechsler Scales, other widely used measures of FI are Raven's Matrices, including Raven's Progressive Matrices (RPM; Raven & Raven, 1938) and the Coloured Progressive Matrices (CPM; Raven, 1947; 1962), which require completing visual patterns by selecting the correct element from a series of distractors. Given their non-verbal format and their shorter administration time

compared to the WISC-IV, Raven's Matrices have been considered a valid alternative for the evaluation of intellectual abilities in clinical conditions, such as Intellectual Disability (Mungkhetklang et al., 2016), Autism (Nader et al., 2016) and Cerebral Palsy (Pueyo et al., 2008). This has also been suggested in relation to SLD, given that several WISC-IV indices may sometimes be biased in children with learning difficulties (Giofrè & Cornoldi, 2015).

Consistent with findings from WISC-based studies, those using Raven's Matrices have found no evidence that children with learning difficulties are more impaired in FI than controls (e.g., Bayliss et al., 2005; Belelli et al., 2010). For instance, Belelli et al. (2010) found that children with moderate SLD performed comparably or even better than typically developing peers on the Coloured Progressive Matrices. However, when examining the types of errors made by participants, children with SLD were found to commit different error patterns compared to controls (Belelli et al., 2010). This finding suggests that focusing solely on the overall accuracy score on Raven's Matrices (i.e., number of correct answers) may overlook meaningful variations in problem solving strategies and that a qualitative error analysis could shed light on the underlying cognitive processes involved in the resolution, which may vary depending on age or in presence of cognitive impairments (see Kunda et al., 2016, for an extensive discussion).

The importance of a qualitative analysis of the performance on Raven's Matrices has been highlighted by Raven himself, who, in his test manuals (Raven et al., 2003), classified possible errors into four categories: Repetition (i.e., the selected distractor mirrors a matrix piece located next to the blank space), Difference (i.e., the chosen distractor is qualitatively distinct from the other alternatives, for example by being blank or containing elements not found elsewhere in the matrix), Wrong Principle (i.e., the selected distractor replicates or combines elements from different parts of the matrix according to an incorrect rule, due to a misinterpretation of the underlying relationships), and Incomplete Correlate (i.e., the chosen distractor closely approximates the correct answer but is not

entirely accurate) (see Kunda et al., 2016). In line with that, several studies have investigated the possible presence of specific error patterns on Raven's Matrices, in both clinical and non-clinical samples. For instance, Babcock (2002) found that adults with high versus low ability (based on correct answers) made different types of errors, with the most frequent error in high ability adults being incomplete correlate, while no specific error pattern was found in low ability participants.

Research on clinical populations has yielded similar results, identifying specific error patterns in children with neurodevelopmental disorders, such as Down Syndrome (Belacchi et al., 2012; Gunn & Jarrold, 2004), Cerebral Palsy (Asano et al., 2023), and Intellectual Disability (Goharpey et al., 2013). These findings suggest that analyzing error types can provide valuable information about problem-solving strategies and logical reasoning in these populations. However, studies investigating error patterns on Raven's Matrices among children with SLD are still scarce. Although preliminary evidence suggests that individuals with SLD may commit different types of errors compared to their typically developing peers (Belelli et al., 2010), further studies are needed to drawn consistent conclusions.

Research has also explored age-related changes in performance on Raven's Matrices. Among the general population, most studies on adults have shown that older individuals tend to perform more poorly on Raven's Matrices compared to younger ones (e.g., Panek & Stoner, 1980); conversely, among children's and adolescents', performance tends to improve with age (e.g., Gunn & Jarrold, 2004). This is in line with theories of FI development, which state that it develops rapidly during childhood, continues to increase during adolescence through early adulthood, before starting to decline gradually in adulthood and older age (Fry & Hale, 2000; McArdle et al., 2002).

Regarding error patterns, a few studies have investigated the potential effect of age. While Babcock (2002) found no age-related differences in error types among adults, Gunn and Jarrold (2004) showed that, in their sample of typically developing children, the proportion and distribution of error types

changed in relation to age, with older children making more repetition and incomplete correlate errors and less difference and individuation (i.e., wrong principle; Kunda et al., 2016) errors than younger children. In contrast, among individuals with Down Syndrome, they found a positive correlation between age and performance, but no significant differences in the error pattern across age (Gunn and Jarrold, 2004).

Studies on age-related differences in Raven's Matrices performance among children with SLD are currently lacking. However, there is evidence that FI development is closely linked to the development of working memory and processing speed (e.g., Fry & Hale, 1996, 2000), two cognitive functions that appear to be often impaired in children with learning disorders (e.g., Cornoldi et al., 2014, 2019a; De Clercq-Quaegebeur et al., 2010; Poletti, 2016; Toffalini et al., 2017). Therefore, examining potential age-related changes in the performance on Raven's Matrices in SLD, both quantitatively and qualitatively, seems warranted.

In conclusion, the reviewed evidence highlights the importance of assessing fluid abilities in children with SLD, as they may display different levels of FI development, which can play a dual role in learning: when efficient, FI could serve as a strength that supports the development of scholastic skills (Cattell, 1971, 1987; Green et al., 2017; Johann et al., 2020; Passolunghi et al., 2022; Peng et al., 2019); whereas when inefficient or poor, it may represent a vulnerability that hinders learning progress (Droder et al., 2024; Mano et al., 2018). Moreover, when assessing FI with Raven's Matrices, going beyond the overall accuracy score and adding a qualitative error analysis can offer deeper insight into the cognitive functioning of children with SLD (Asano et al., 2023; Belacchi et al., 2012; Belelli et al., 2010; Goharpey et al., 2013; Gunn & Jarrold, 2004; Kunda et al., 2016). In this context, recently developed computerized versions of the Raven's Matrices (e.g., MatriKS; de Chiusole et al., 2024) may provide a more efficient and information-rich alternative solution compared to the paper-and-pencil format, as they automatically record a large amount of detailed

information and parameters on the test-taker's performance, making data collection easier and more time efficient (Mead & Drasgow, 1993; Witt et al., 2013).

Based on these considerations, the present study aimed to further investigate fluid intelligence (FI) in children with specific learning disorders (SLD) using the new digital assessment tool MatriKS (de Chiusole et al., 2024). Specifically, four main objectives were addressed:

1) To compare the total accuracy score obtained in MatriKS between healthy controls and specific learning disorders group (dyslexia and dysorthographia, dyscalculia, or mixed) and observe the distribution of the total number of errors.
2) To investigate the presence of differences in the proportion of errors (R, WP, D, IC) between the specific learning disorders group and the control group
3) To examine the effect of age on these measures (age included as a covariate)
4) To investigate the relationship between fluid intelligence performance on the MatriKS task (measured by total accuracy) and performance on assessments of cognitive abilities, academic skills (reading, writing, arithmetic), and cross-domain learning skills (attention, memory, and non-verbal logical reasoning).

2. Participants

The sample consisted of 106 children and adolescents (59 females, 56%; 47 males, 44%) aged 7.00 to 17.11 years (M = 11.74, SD = 1.96), divided into two age groups based on the MatriKS test version administered. The younger age group (7–11 years) comprised 56 participants (M = 10.30, SD = 1.41; 34 females, 61%; 22 males, 39%). Of these, 30 were typically developing (TD) children (54%) and 26 (46%) had a diagnosis of Specific Learning Disorder (SLD) according to DSM-5 (American Psychiatric Association [APA], 2013) and ICD-10 (World Health Organization [WHO], 1992) criteria. The older age group (12–17.11 years) comprised 50 participants (M = 13.34, SD = 1.02; 25 females, 50%; 25 males, 50%). Of these, 30 were TD adolescents (60%) and 20 (40%) met diagnostic

criteria for SLD based on DSM-5 and ICD-10 classifications. Overall, the sample included 60 TD participants (57%) and 46 participants with SLD (43%).

Table 1 reports the descriptive statistics of the total sample of participants (N = 106), divided into two groups based on age and the version of MatriKS test administered: MatriKS 4–11 years (N = 56) and MatriKS 12+ years (N = 50). Within each group, participants with TD and those with a SLD are distinguished.

Table 1. Sample Characteristics for MatriKS 4–11 and MatriKS 12+ Groups

Groups	MatriKS 4-11		MatriKS 12+	
	n	Age in years, M (SD)	*n*	Age in years, M (SD)
TD	30	10.72 (1.5)	30	12.99 (.45)
SLD	26	9.83 (1.17)	20	13.87 (1.38)
Total	56	10.30 (1.41)	50	13.34 (1.02)

Participants were recruited from two primary sources: (1) the Child and Adolescent Neuropsychiatry Units (UONPIA) of Forlì-Cesena and Rimini (AUSL Romagna), and (2) the Clinical Service SPEV (Servizio di Potenziamento cognitivo dell'Età Evolutiva) of the Department of Psychology, University of Bologna. Clinical participants presented for initial assessment of suspected SLD, reassessment of previously diagnosed SLD, or were receiving ongoing treatment at these facilities. Data collection was conducted at AUSL Romagna sites by licensed psychologists in collaboration with the principal investigators of the respective UONPIA sites and a research team from the Department of Psychology, University of Bologna. Data analysis was performed by the Laboratory of Psychometrics and Neuropsychology of the Department of Psychology, University of Bologna.

The inclusion criteria were as follows:

1. Age between 7.00 and 17.11 years

2. For the SLD group: formal diagnosis of SLD according to DSM-5 (APA, 2013) and ICD-10 (WHO, 1992) criteria, established through comprehensive neuropsychological assessment

3. For the TD group: absence of learning disorders, neurological conditions, or psychiatric diagnoses based on parent report and clinical history

4. Adequate vision and hearing (normal or corrected-to-normal) to complete tablet-based tasks

5. Ability to understand task instructions in Italian

The exclusion criteria were as follows:

1. Motor, visual, or auditory impairments that would compromise the ability to perform tablet-based tasks or communicate responses to the examiner

2. Intellectual disability, defined as Total Intelligence Quotient (TIQ) < 70 (more than 2 standard deviations below the mean, corresponding to the range of 55–69), as assessed by standardized cognitive batteries (WISC-IV, Orsini et al., 2012; WAIS-IV, Orsini & Pezzuti, 2013).

3. Primary diagnosis of other neurodevelopmental disorders (e.g., Autism Spectrum Disorder)

4. Primary diagnosis of emotional or behavioral disorders

5. Comorbid psychiatric conditions

The study was conducted in accordance with the Declaration of Helsinki (Williams, 2008) and approved by the Ethics Committee of Romagna (Comitato Etico di Area Vasta Romagna e IRST [CET–C.E.ROM]; protocol number 410/2025 I.5/46; approval date: July 9, 2025). Written informed consent was obtained from parents or legal guardians of all participants prior to enrollment in the study. Only participants for whom signed parental informed consent was available were included. Age-appropriate verbal consents were also obtained from all children and adolescent participants.

It should be noted that data collection and recruitment of participants with SLD are ongoing, with the objective of expanding the SLD group sample; therefore, given the preliminary nature of this study, interpretations should be considered exploratory.

3. Materials and Methods

3.1 Instruments

MatriKS (de Chiusole et al., 2024)

MatriKS is a computerized test, built on the characteristics and principles of Raven's Matrices, and developed to assess Fluid Intelligence in individuals of different ages (i.e., children, adolescents, and adults), with or without clinical conditions. As traditional Raven-like tasks, it is composed of a series of visual patterns, each containing a missing element that must be completed by selecting one of several alternatives, with the incorrect options also termed "distractors".

Two versions of MatriKS are available: MatriKS 4-11 for children aged between four and 11 years old, and MatriKS 12+ for adolescents and adults older than 12. In both versions, each stimulus consists of two parts. One is the matrix, which can be of three types: monothematic matrix (a single figure containing a missing piece), 2x2 matrix (a set of four cells with one missing), or 3x3 matrix (a set of nine cells with one missing). The matrix appears on the left of the screen and has the missing cell in its low-right corner. The other part of the stimuli consists of the response options, which are five or eight, depending on the type of matrix, and include the correct item and the distractors. The distractors were purposefully created to reproduce the error categories of traditional Raven's Matrices, previously described (see Introduction), with the aim of collecting information on error patterns. In addition to these error types (i.e., Repetition, Difference, Wrong Principle, and Incomplete Correlate), in the present study another error category, referred to as "Repetition and Incomplete correlate", was considered, which includes distractors that exhibit characteristics of both Repetition and Incomplete Correlate errors. In MatriKS, the response options are shown next to the corresponding matrix, on the right of the screen. The system records the participant's selected answer (i.e., correct, incorrect, non-provided), the selected distractor in case of an incorrect response, the time spent on the whole test and on each problem, and the average time per item.

MatriKS 4-11 includes 40 stimuli, of which 22 are colored and 18 in black and white. As for the dimensions, five matrices are monothematic, 20 matrices are 2x2, and 15 matrices are 3x3. The monothematic and 2x2 matrices include five response options, while the 3x3 matrices include eight response options. To be solved, the stimuli of MatriKS 4-11 require skills related to visual perception, elaboration of general configuration, reasoning and elementary inference. MatriKS 12+ includes 52 black-and-white stimuli, each with a dimension of 3x3 and eight response options, except for three stimuli with a dimension of 2x2 and five response options. The two versions have 11 stimuli in common. To be solved, the stimuli of MatriKS 12+ require the skills involved in MatriKS 4-11, plus advanced skills of reasoning and logic inference.

MatriKS stimuli (i.e., matrices and associated responses) were created with a dedicated R package (MatriKS; Brancaccio et al., 2023, cited in de Chiusole et al., 2024; available online https://CRAN.R-project.org/package=matRiks). This package allows for the automated generation of Raven-like matrices based on a series of parameters such as the dimension of the matrix, the objects included, the rule governing the manipulation, and the direction of the manipulation. Further information on the developmental stage of the package is available at https://ottaviae.github.io/stimuliGen/OfficialDocumentation.html.

MatriKS was administered to each child individually by psychologists or trained researchers. The test was presented on the PsycAssist platform (de Chiusole et al., 2024) available at https://psycassist.fisppa.unipd.it by using a tablet (IoS operating system with a 10.9in screen). The administration of MatriKS was carried out with the tablet held horizontally. The MatriKS test was introduced by a brief animated video (95 seconds), which presented the test, explained the structure of the stimuli and the correct way to select the response from the response list. After the video presentation, children were presented with four practice items, and then the test started. There were no time constraints. A response was registered once the child touched one of the response options. If

the child struggled to find an answer to an item, the researcher could decide to skip it. No feedback was given about the correctness of the responses provided by the children. The stimuli were presented one at a time in random order.

Neuropsychological Measures

The following standardized diagnostic instruments were administered to assess cognitive abilities, learning skills, and cross-domain learning competencies, according to the participant's age and level of schooling. Some of these data were included in the analyses reported below, while others could not be explored due to the limited sample size. However, data collection is still ongoing, and these aspects will be further investigated in future studies. All the instruments were administered individually in a quiet room, by trained psychologists (n = 2) or psychology interns (n = 9) in a counterbalanced order, both for the typology of test (computerized vs traditional, i.e., MatriKS vs CPM/SPM) and the traditional measures.

The Colored Progressive Matrices (CPM) (Raven, 1947;1962; Italian adaptation of CPM by Belacchi et al., 2008)

The CPM represent one of the most popular measures of non-verbal reasoning abilities for children and elders. It is composed of 36 items of increasing difficulty, organized into three series (i.e., A, AB, and B) having 12 items each. The task consists of identifying the piece that best completes the given matrix choosing among six alternatives. Serie A evaluates the ability to identify similarities (based on shape, dimension, direction, quantity, orientation, figure/background, density criteria), Serie AB assesses the ability to detect symmetry, and Serie B measures the conceptual thinking skills (i.e., detection of abstract relations according to operant-deductive logic and their retention in working memory). A score of one is assigned if the person identifies the correct option. Otherwise, the score is zero. Thus, the highest total score achievable is 36.

Raven's Standard Progressive Matrices (SPM) (Raven, 1938; Raven et al., 1989; Italian adaptation by Picone et al., 2018)

The SPM measure non-verbal reasoning and are standardized for individuals aged 6 to 60 years. They consist of a series of visual matrices, each containing a pattern with a missing piece that must be completed by selecting the correct element from a set of distractors. The battery includes 60 items divided into five series (A–E), each composed of 12 items with increasing difficulty. Each series addresses distinct cognitive themes: continuous patterns (Series A), analogical relationships (Series B), progressive pattern transformations (Series C), figure permutations (Series D), and decomposition of figures into constituent parts (Series E) (Burke, 1958). In the present study, the test was administered to children aged 12 to 17 years and 11 months.

The instruments were administered individually in a quiet room, by trained psychologists (n = 2) or psychology interns (n = 9). Instruments were administered in a counterbalanced order, both for the typology of test (computerized vs traditional, i.e., MatriKS vs CPM/SPM) and the traditional measures.

WISC-IV (Wechsler Intelligence Scale for Children-IV; Wechsler, 2003; Italian adaptation by Orsini et al., 2012) and WAIS-IV (Wechsler Adult Intelligence Scale-IV; Wechsler, 2008; Italian adaptation by Orsini & Pezzuti, 2013)

The WISC-IV and WAIS-IV are multi-component batteries for the assessment of cognitive abilities, developed for children aged between 6 years and 16 years and 11 months, and for adolescents and adults aged between 16 years and 89 years and 11 months, respectively. They comprise a series of tasks organized to provide four indices: Verbal Comprehension (VCI; assessing verbal reasoning and knowledge), Perceptual Reasoning (PRI; assessing visuo-perceptual and non-verbal fluid reasoning), Working Memory (WMI; assessing the ability to hold and manipulate information in memory), and Processing Speed (PSI; assessing the speed and accuracy of information processing in visuomotor

tasks). In addition to these indices, the Full-Scale Intelligence Quotient (FSIQ) is obtained, representing the child's overall cognitive abilities.

In the present study, the WISC-IV was administered only to a subset of participants aged between 7 years and 16 years and 11 months, whereas the WAIS-IV to a subset of participants aged between 7 years and 17 years and 11 months.

Digit Span subtest (WISC-IV; Wechsler Intelligence Scale for Children-IV; Wechsler, 2012; Italian adaptation by Orsini et al., 2012; WAIS-IV; Wechsler Adult Intelligence Scale-IV; Wechsler, 2008; Italian adaptation by Orsini & Pezzuti, 2013).

This subtest, included in both the WISC-IV and the WAIS-IV batteries, measures auditory short-term memory and working memory capacity. The participant is required to listen to and repeat a sequence of numbers that increases in length with each trial. In the Forward Digit Span, numbers must be repeated in the same order as presented, while in the Backward Digit Span, they must be repeated in reverse order.

In the present study, the WISC-IV subtest was administered to all children and adolescents aged between 7 years and 11 months and 16 years and 11 months, whereas the WAIS-IV subtest was administered to all adolescents aged between 17 years and 17 years and 11 months.

NEPSY-II (Korkman, Kirk & Kemp, 2007; Italian adaptation by Urgesi et al., 2011)

The NEPSY-II is a standardized neuropsychological battery designed to assess cognitive, language, and neuropsychological abilities in children and adolescents aged 3 to 16 years. The NEPSY-II includes six main domains, each composed of specific subtests: Attention and Executive Functions, Language, Memory and Learning, Sensorimotor Functions, Social Perception, and Visuospatial Processing. The battery can be administered in full or selectively, depending on the assessment goals. In the present study, only the Visual Attention subtest was administered (NEPSY-II; Korkman, Kirk

& Kemp, 2007; Italian adaptation by Urgesi et al., 2011). In the Visual Attention subtest, the child is required to identify as many target stimuli as possible among distractors within a specified time frame.

Leiter International Performance Scale revised (Leiter - R; Roid & Miller, 1997)

The Leiter-R battery is a nonverbal scale for assessing intelligence and cognitive abilities in individuals aged 2 to 21. For this specific study, we administer only the sustained attention subtest of the scale, in which the individual is required to identify as many target stimuli as possible among distractors within a set time limit.

DDE-2 (Battery for the Assessment of Developmental Dyslexia and Dysorthography; Sartori, Job & Tressoldi, 2007)

The DDE-2 is an Italian standardized battery for the assessment of reading and writing abilities in children from the second grade of primary school to the third grade of lower secondary school. In the present study, reading abilities were evaluated in terms of accuracy (number of errors) and speed (syllables per seconds) through word and nonword list reading tasks. Writing abilities, on the other hand, were assessed based on accuracy (number of errors) in tasks requiring the production of lists of words, nonwords, and sentences containing homophonic but non-homographic words.

BVSCO-3 (Battery for the Assessment of Writing and Orthographic Competence-3; Cornoldi et al., 2022)

The BVSCO-3 is an Italian standardized battery for the assessment of writing, dysgraphia and dysorthography in children aged 6 to 14 years. It includes tasks for the assessment of handwriting, orthographic competence, and written text production. In this study, the text dictation task was used to assess orthographic competence.

MT-3 Clinical (6-14) (Cornoldi & Carretti, 2016)

The MT-3 Clinical (6-14) is an Italian standardized battery for the assessment of reading and comprehension abilities in children aged 6 to 14 years. In the current study, reading abilities were

assessed in terms of speed (syllables per second) and accuracy (number of errors) using a written text reading task, while comprehension was evaluated based on accuracy through a written text comprehension task, which requires reading a passage and answering a series of multiple-choice questions.

MT Advanced-3 Clinical (Cornoldi et al., 2017)

The MT Advanced-3 Clinical is an Italian standardized battery designed to assess reading, comprehension, writing, and mathematics skills in upper secondary school students. In the present study, adolescents attending the first and second year of upper secondary school were administered the words, nonwords, and text reading tasks, which were evaluated in terms of both speed and accuracy, as well as the text comprehension task, which was evaluated in terms of accuracy. With regard to the mathematics domain, the tasks were administered to students attending from the first to the fifth year of upper secondary school. Specifically, arithmetic and calculation skills were assessed through tasks measuring numerical competence (accuracy), mental calculation (speed and accuracy), and retrieval of arithmetic facts (speed and accuracy).

MT 16-19 (Battery for the verification of learning and diagnosis of Dyslexia and Dysorthography; Cornoldi & Candela, 2015)

The MT 16-19 is an Italian standardized battery for the diagnosis of dyslexia and dysorthography in adolescents aged 16 to 19 years. It includes tasks designed to assess reading abilities of texts, words, and nonwords, comprehension of written texts, and writing skills through word dictation, dictation of sentences containing homophones, and writing numbers in words. In the present study, the words, nonwords, and text reading tasks were administered and evaluated in terms of both speed and accuracy; the written text comprehension tasks with multiple-choice answers were evaluated in terms of accuracy; and the words and sentence dictation tasks were evaluated in terms of both speed and accuracy.

BDE-2 (Battery for Developmental Dyscalculia; Biancardi et al., 2016)

The BDE-2 is an Italian standardized battery designed to assess numerical processing and calculation skills for the diagnosis of dyscalculia in children aged 8 to 13 years, including both primary school students (last three grades) and lower secondary school students. The battery comprises a series of tasks assessing three main areas: Number Area: tasks of forward and backward counting and numerical transcoding, including reading, writing, and repeating numbers; Calculation Area: tasks of mental calculation, retrieval of arithmetic facts, and timed written calculation; Number Sense Area: tasks of approximate calculation and determination of the order of magnitude of numbers. In addition to the final scores for each of the three areas, the BDE-2 provides a cumulative overall score. An optional arithmetic problem-solving task is also included, which yields a separate score. In this study, all mandatory tasks of the full battery were administered to children from the third grade of primary school to the third grade of lower secondary school, while optional tasks were only partially included. For each test, raw scores and age-corrected scores are considered according to the normative data provided in the manuals, following the standard assessment protocol.

3.2 Data Analysis

All statistical analyses were performed using IBM SPSS Statistics version 27 (IBM Corp, 2017) and JASP version 0.18.3.0 (JASP Team, 2024).

To address the study objectives, the following analyses were conducted. To compare overall accuracy between typically developing (TD) children and children with specific learning disorders (SLD), two separate General Linear Models (GLMs) were conducted for the MatriKS 4–11 (7–11 years) and MatriKS 12+ (12–17.11 years) versions. In these models, total accuracy (number of correct items) was entered as the dependent variable and group (TD vs. SLD) as the independent variable.

To investigate differences in error types between groups, non-parametric repeated-measures analyses were conducted using Generalized Estimating Equations (GEE). Separate analyses were performed

for MatriKS 4–11 and MatriKS 12+ versions. The models included error type (Repetition [R], Wrong Principle [WP], Difference [D], Incomplete Correlate [IC]), Repetition and Incomplete correlate category [R/IC], as the within-subject factor, group (TD vs. SLD) as the between-subject factor, and age as a covariate. Age effects on both overall accuracy and error patterns were examined by including age as a covariate.

Exploratory non-parametric correlation analyses (Spearman's rho) were conducted to examine relationships between fluid intelligence performance on MatriKS (total accuracy) and other cognitive and academic measures, including attention, memory, non-verbal logical reasoning, and academic skills (reading, writing, mathematics). These correlations were computed separately for each age group (TD and SLD).

4. Results

General Linear Model (GLM) analyses were conducted separately for each age group to assess potential differences in overall accuracy on MatriKS between typically developing (TD) children and children with specific learning disorders (SLD).

Results overall accuracy MatriKS 4–11 Version (Ages 7–11 Years):

For the younger age group, the GLM analysis revealed no statistically significant difference on MatriKS 4-11 between TD children (M = 30.70, SD = 6.54) and children with SLD (M = 27.31, SD = 5.40; Likelihood ratio $\chi^2(1) = 3.42$, $p = .064$; see Figure 1 and Table 2), though the result approached the conventional significance threshold. Children with SLD scored numerically lower than TD peers, with a mean difference of 3.4 correct items (Cohen's d = 0.56, medium effect size).

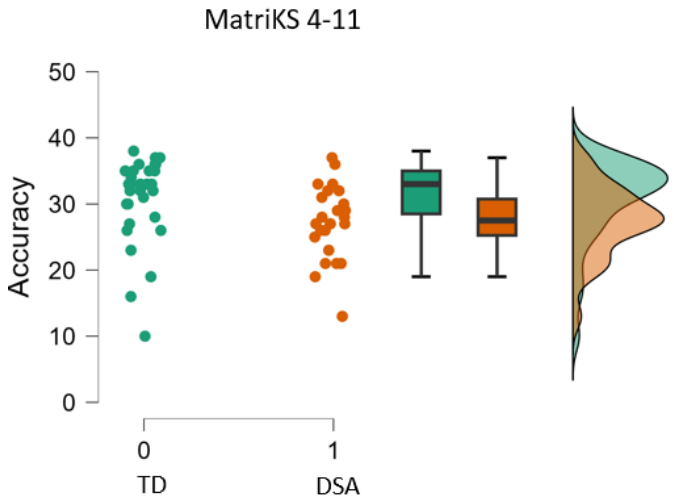


Figure 1. Distribution of overall accuracy scores by group on the MatriKS 4–11 version. Box plots show medians, interquartile ranges, and score distributions for TD and SLD groups.

Table 2. Descriptive Statistics for Overall Accuracy by Group (MatriKS 4–11)

Group	*n*	M (SD)
TD	30	30.70 (6.54)
SLD	26	27.31 (5.40)

Note. TD = Typically Developing; SLD = Specific Learning Disorder; M = Mean; SD = Standard Deviation.

Results overall accuracy MatriKS 12+ Version (Ages 12–17.11 Years):

Similarly, for the older age group, the GLM analysis did not reveal a statistically significant difference on MatriKS 12+ between TD adolescents (M = 30.17, SD = 8.87) and those with SLD (M = 26.55, SD = 7.64; Likelihood ratio $\chi^2(1) = 1.67$, p = .196; see Figure 2 and Table 3). However, the observed mean difference of 3.6 correct items (Cohen's d = 0.44, small-to-medium effect size) was comparable in magnitude to that observed in the younger age group.

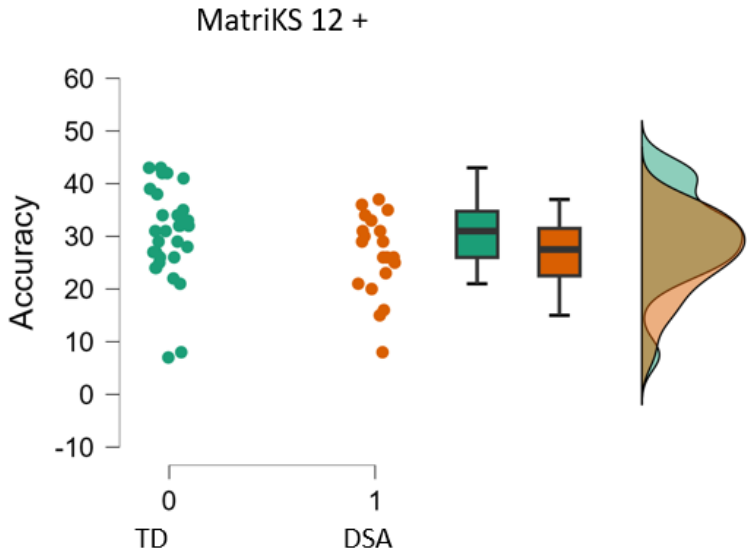


Figure 2. Distribution of overall accuracy scores by group on the MatriKS 12+ version. Box plots show medians, interquartile ranges, and score distributions for TD and SLD groups.

Table 3. Descriptive Statistics for Overall Accuracy by Group (MatriKS 12+)

Group	*n*	M (SD)
TD	30	30.17 (8.87)
SLD	20	26.55 (7.64)

Note. TD = Typically Developing; SLD = Specific Learning Disorder; M = Mean; SD = Standard Deviation.

Across both age groups, no statistically significant differences emerged in overall accuracy on MatriKS between TD and SLD groups, although the difference in the younger group approached significance (younger: p = .064; older: p = .196). However, children and adolescents with SLD demonstrated consistently lower scores than TD peers (younger: mean difference = 3.4 items, d = 0.56; older: mean difference = 3.6 items, d = 0.44). These medium effect sizes suggest that, despite the lack of conventional statistical significance in this preliminary sample, there may be subtle differences in fluid reasoning performance between groups. The consistent pattern across age groups

and the magnitude of effect sizes indicate that these findings warrant replication with larger samples to achieve adequate statistical power.

Error Type Results

Non-parametric repeated-measures analyses using Generalized Estimating Equations (GEE) were conducted to investigate potential differences in the distribution of error types (i.e., Repetition, Wrong Principle, Difference, Incomplete Correlate, Repetition and Incomplete correlate) between typically developing (TD) children and children with specific learning disorders (SLD) for both MatriKS versions.

Error Type Results: MatriKs 4–11 Version

For the younger age group, the GEE analysis revealed significant main effects of age (Wald χ^2 (1) = 9.26, p = .002) and error type (Wald χ^2 (4) = 59.51, p < .001), indicating that error frequency varied systematically with age and that certain error categories were more prevalent than others. The main effect of group approached statistical significance (Wald χ^2 (1) = 3.34, p = .068), as did the group × error type interaction (Wald χ^2 (4) = 8.88, p = .064). These marginally significant effects suggest that children with SLD and TD children may exhibit different error patterns, though the current sample size limits statistical power to detect these differences reliably. Descriptive analyses showed that, across both groups, Incomplete Correlate (IC) errors were the most frequent error type. However, children with SLD tended to produce a higher number of errors overall compared to TD children, particularly in the IC and Repetition (R) categories. Participants with SLD also displayed a more pronounced imbalance between IC and R errors compared to other error types, whereas TD participants exhibited a more uniform distribution across error categories. The distribution of errors is illustrated in Figure 3, and descriptive statistics are reported in Table 4.

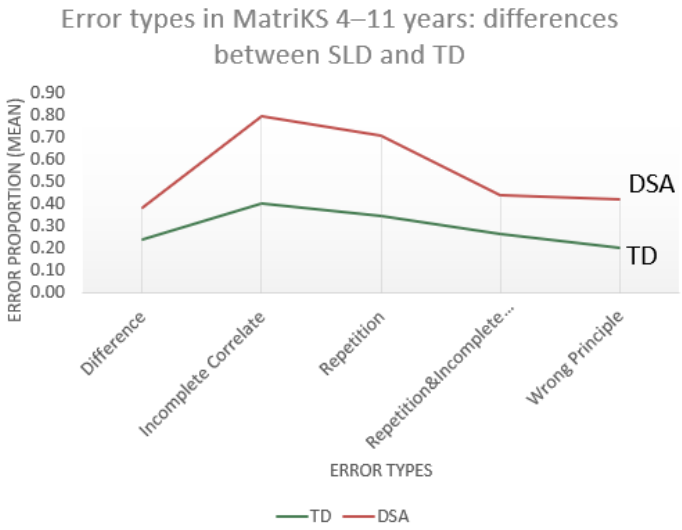


Figure 3. Analysis of error types committed by TD and SLD participants in the MatriKS 4-11 version using the Generalized Estimating Equations (GEE) model. The x-axis represents the different error types, and the y-axis shows the mean number of errors committed by each group.

Table 4. Mean (M) and standard error (SE) of the number of errors in MatriKS 4-11

MatriKS 4-11	*TD GROUP*		*SLD GROUP*	
Errors	*M*	SE	*M*	SE
Difference	.24	.03	.14	.03
Incomplete Correlate	.40	.04	.40	.04
Repetition	.35	.04	.36	.04
Repetition/Incomplete Correlate	.26	.04	.18	.02

Wrong Principle	.20	0.3	.22	.03

Note. TD = Typically Developing; SLD = Specific Learning Disorder; M = Mean; SE = Standard Error.

Error Type Results: MatriKs 12+ Version

For the older age group, the GEE analysis revealed a significant main effect of error type (Wald χ^2 (4) = 468.65, $p < .001$), indicating substantial differences in the frequency of different error categories. However, no significant main effects emerged for group (Wald χ^2 (1) = 2.23, $p = .136$) or age (Wald χ^2 (1) = 1.38, $p = .240$), and the group × error type interaction was not significant (Wald χ^2 (4) = 2.19, $p = .702$). These results indicate that, unlike the younger age group, TD and SLD adolescents exhibited comparable overall error rates and similar error profiles. Descriptive analyses showed that, across both groups, Incomplete Correlate (IC) and Wrong Principle (WP) errors were the most frequent error types. However, adolescents with SLD exhibited a numerical trend toward higher error rates compared to TD, particularly in the WP and IC categories. Participants with SLD displayed a more pronounced tendency toward WP and IC errors compared to other error types, with the largest absolute differences observed for these two categories. In contrast, Difference (D) errors were relatively infrequent in both groups. The distribution of errors is illustrated in Figure 4, and descriptive statistics are reported in Table 5.

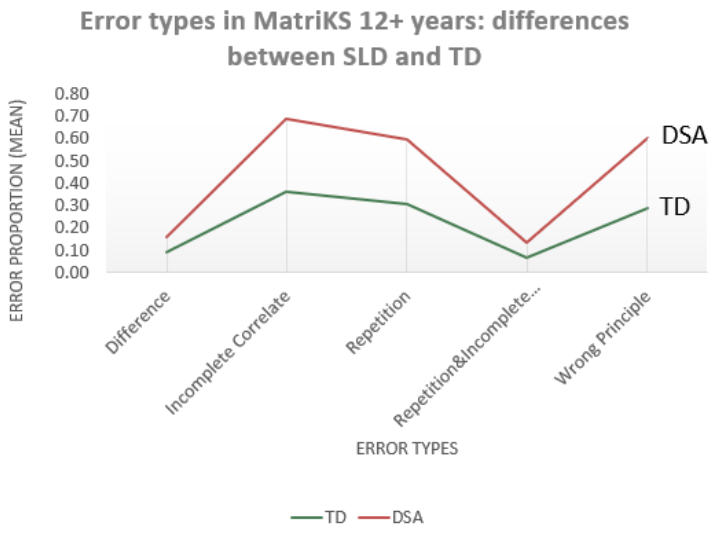


Figure 4. Analysis of error types committed by TD and SLD participants in the MatriKS 12+ version using the Generalized Estimating Equations (GEE) model. The x-axis represents the different error types, and the y-axis shows the mean number of errors committed by each group.

Table 5. Mean (M) and standard error (SE) of number of errors in MatriKS 12+

MatriKS 12+	*TD GROUP*		*SLD GROUP*	
Error	*M*	SE	*M*	SE
Difference	.09	.02	.07	.01
Incomplete Correlate	.36	.03	.33	.03
Repetition	.31	.03	.29	.02
Repetition/Incomplete Correlate	.07	.01	.06	.01

Wrong Principle	.29	.02	.31	.02

Note. TD = Typically Developing; SLD = Specific Learning Disorder; M = Mean; SE = Standard Error.

Descriptive analyses of the different error types for the 4–11 and 12+ versions are presented below, illustrated using boxplots and raincloud plots. For each error, the left panel represents the 4–11 version, and the right panel represents the 12+ version of MatriKS. The typically developing (TD) group is shown in green, and the specific learning disorder (SLD) group in orange.

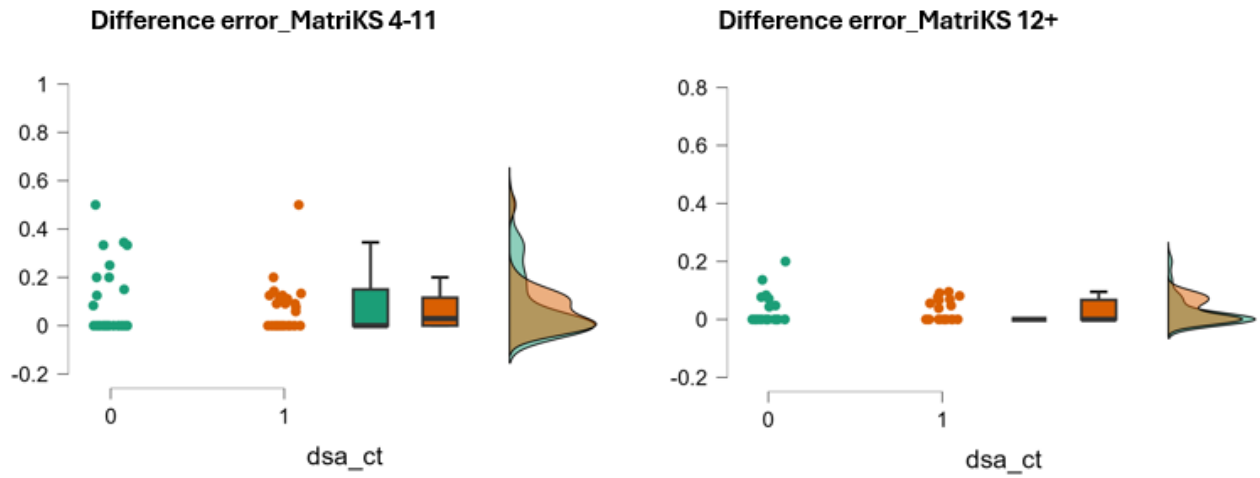


Figure 5. Boxplots showing the proportion of "Difference" errors by group and test version. The typically developing (TD) group is shown in green, and the specific learning disorder (SLD) group in orange. In both versions, the SLD group exhibited more error compared to their age-matched TD peers, although this difference was not statistically significant.

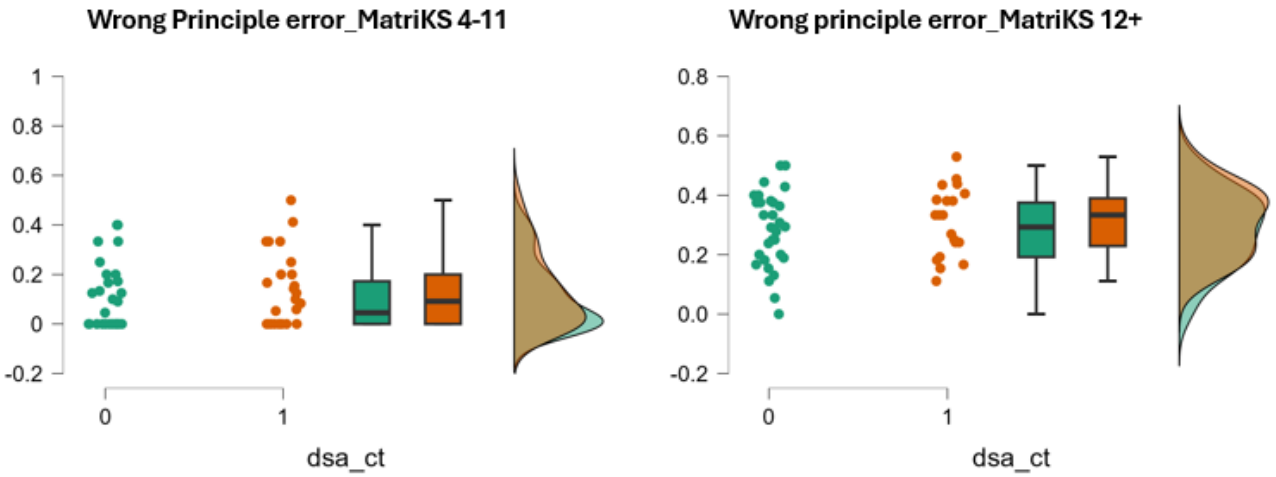


Figure 6. Boxplots showing the proportion of "Wrong Principle" errors by group and test version. For this error type as well, the SLD group committed more errors than the TD group, although these differences were not statistically significant.

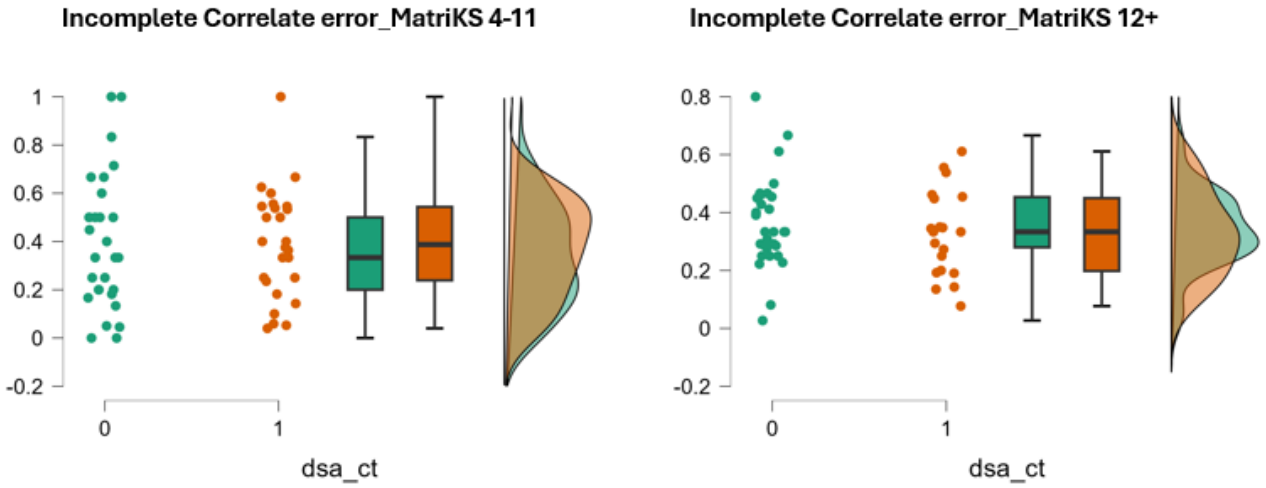


Figure 7. Boxplots showing the proportion of "Incomplete Correlate" errors by group and test version. The pattern of errors across the two groups follows a similar trend in both versions, with the SLD group showing more errors than TD group.

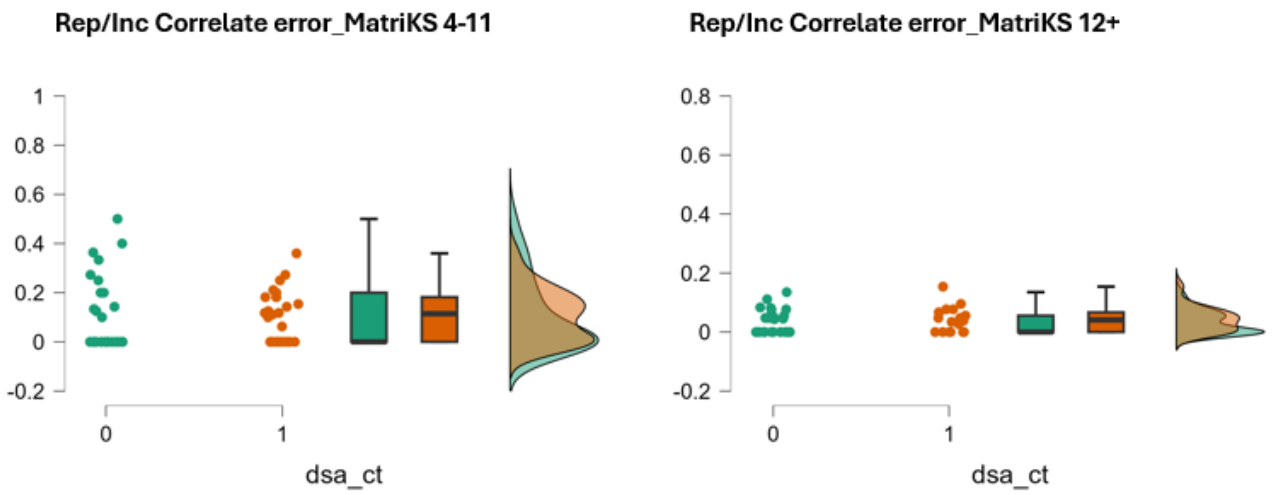


Figure 8. Boxplots showing the proportion of errors in the category "Rep/Inc Correlate", including distractors that exhibit characteristics of both Repetition and Incomplete Correlate errors. The TD group demonstrated higher accuracy compared to the SLD group, although this difference was not statistically significant. Both groups exhibited more variable performance in the 4–11 version than in the 12+ version.

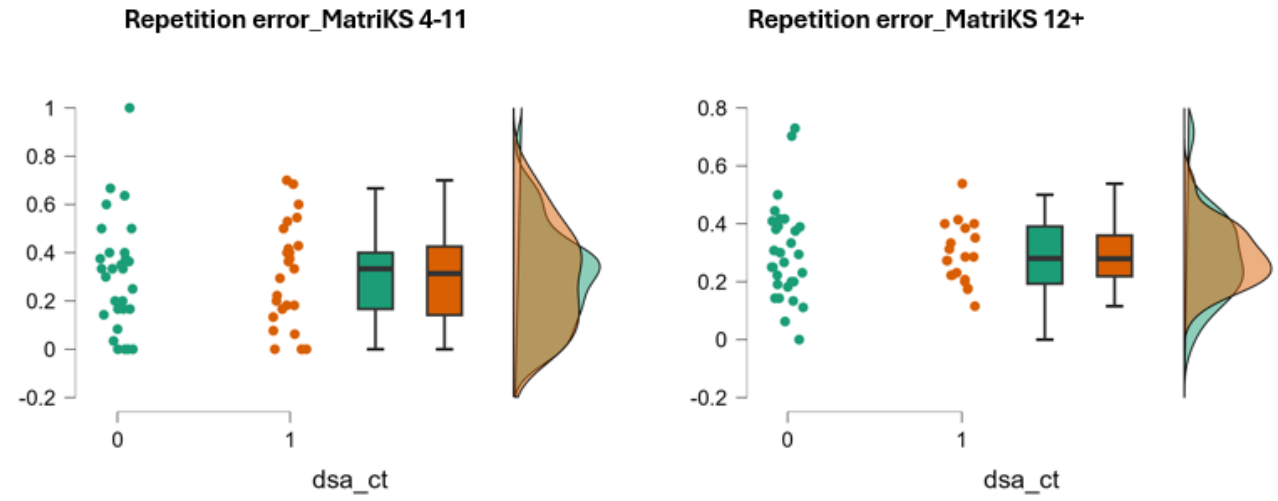


Figure 9. Boxplots showing the proportion of "Repetition" errors by group and test version. Performance in both groups followed a similar pattern across versions, with some variability in scores.

The TD group committed slightly fewer errors than the SLD group, although this difference was not statistically significant.

Correlation results:

Exploratory Spearman correlation analyses conducted to examine the relationships between MatriKS performance (both 4-11 and 12+ versions), and other cognitive and academic measures showed no significant correlations between overall accuracy on the MatriKS and any of the assessed variables (all $p > .05$), except the comprehension subtest and MatriKS 12+. Although these correlations did not reach statistical significance, they are reported in Tables 6 and 7 and will be discussed further below (see Discussion section).

Table 6. Correlations between MatriKS 4-11 and Learning and cognitive tasks

Learning and cognitive tasks	MatriKS 4-11
Writing	
Words	-0.161
Non-words	-0.231
Sentences	0.486
Text	0.200
Reading	
Words	
Accuracy	-0.460
Speed	-0.277
Non words	
Accuracy	0.174
Speed	-0.286
Text	

Accuracy	0.083
Speed	-0.414
Comprehension	0.254
Memory (WISC-IV)	0.286
Attention (Nepsy)	0.241
Colored Progressive Matrices (CPM)	0.205

Table 7. Correlations between MatriKS 12+ and Learning and cognitive tasks

Learning and cognitive tasks	**MatriKS 12+**
Writing	
Words	-0.464
Non-words	0.257
Sentences	-0.500
Text	**0.775***
Reading	
Words	
Accuracy	0.126
Speed	-0.042
Non words	
Accuracy	-0.295
Speed	0.025
Comprehension	0.714
Memory (WISC-IV)	0.105
Visual attention (Nepsy)	-0.633

Standard Progressive Matrices (SPM)	0.369

. Correlation is significant at the 0.05 level (2-tailed).

5. Discussions

The present study addressed four main objectives regarding the assessment of fluid intelligence in children and adolescents with Specific Learning Disorders (SLD) using MatriKS, a novel computerized adaptation of Raven's Progressive Matrices (de Chiusole et al., 2024). First, we examined whether MatriKS could detect differences in overall accuracy between typically developing (TD) children and those with SLD across two developmental periods (ages 7–11.11 and 12–17.11 years). Second, we investigated whether distinct error patterns would emerge between groups (i.e., SLD and TD), specifically examining the distribution of specific error types. Third, we evaluated the effect of age on both overall accuracy and error patterns by including age as a covariate in our analyses. Fourth, we explored the relationships between MatriKS performance and other cognitive and academic measures, including attention, memory, non-verbal logical reasoning, and academic skills (reading, writing, comprehension and mathematics).

Regarding the first objective, General Linear Model analyses revealed no statistically significant differences in overall accuracy between TD and SLD groups across either age range, although the difference in the younger group approached the conventional significance threshold ($p = .064$). However, children and adolescents with SLD consistently scored lower than their typically developing peers across both age groups. Although these differences did not reach statistical significance in this preliminary sample, they suggest the possible presence of subtle variations in fluid reasoning abilities, which should be further investigated in studies with larger samples and adequate statistical power. These findings are consistent with the body of prior research documenting average global intellectual functioning in individuals with SLD. Such evidence emerges both from studies

conducted with the Wechsler scales, which demonstrate that the GAI (General Ability Index) typically falls within the average range (Cornoldi et al., 2014, 2019a; Poletti, 2016; Toffalini et al., 2017), and from those employing Raven's Progressive Matrices (Belelli et al., 2010). However, as extensively documented by the characteristic GAI-CPI (Cognitive Proficiency Index) dissociation in SLD, global scores may mask specific cognitive vulnerabilities: while the GAI remains in the average range, the CPI (Cognitive Proficiency Index), which reflects working memory and processing speed functions, is typically lower in SLD compared to TD (Cornoldi et al., 2014, 2019a; Poletti, 2016; Toffalini et al., 2017). Given that these processes are closely interconnected with the development of fluid intelligence (Fry & Hale, 1996, 2000), the subtle differences observed in MatriKS accuracy, despite falling within the average range, may reflect precisely these vulnerabilities in the cognitive processes underlying fluid reasoning, rather than a general intellectual impairment. This interpretation finds theoretical support in Cattell's Investment Theory (1971, 1987), according to which fluid intelligence constitutes the foundation for the development of crystallized intelligence and academic skills. Empirical evidence has consistently demonstrated that fluid reasoning abilities predict achievement across various academic domains, including reading and mathematics (Green et al., 2017; Johann et al., 2020; Passolunghi et al., 2022; Peng et al., 2019). Consequently, even mild inefficiencies in fluid reasoning processes, when combined with deficits in working memory and processing speed, may contribute significantly to the learning difficulties characteristic of SLD. The marginally significant result observed in the younger group compared to the non-significant result in the older group raises interesting questions regarding the developmental trajectories of fluid reasoning in SLD. On the one hand, this pattern is consistent with theories of fluid intelligence development, which document rapid growth during childhood followed by continued increase throughout adolescence (Fry & Hale, 2000; McArdle et al., 2002). The greater detectability of differences in childhood may reflect genuine processes of developmental convergence, through which adolescents with SLD develop

compensatory strategies that progressively attenuate the impact of their underlying cognitive vulnerabilities. On the other hand, methodological factors related to the psychometric properties of the MatriKS 12+ version, which may reduce discriminative power in the older group, cannot be ruled out. These alternative interpretations will be discussed in detail in subsequent sections, particularly in light of the analysis of error patterns and age effects on different error types.

Regarding the second objective, qualitative analysis of error types provided the nuanced insights anticipated in the previous section. In the younger age group (MatriKS 4–11), GEE analysis revealed marginally significant effects for both group ($p = .068$) and the group × error type interaction ($p = .064$), with Incomplete Correlate (IC) and Repetition (R) errors more frequent in children with SLD compared to TD. In the older age group (MatriKS 12+), no significant group differences emerged (interaction: $p = .702$), although descriptive trends suggested numerically higher error rates in adolescents with SLD, particularly for IC and Wrong Principle (WP) errors.

The convergence between marginally significant differences in overall accuracy and error patterns in the younger group confirms that qualitative analysis of performance, which extends beyond simple global accuracy scores, can reveal differences in underlying cognitive processes that are clinically significant yet remain otherwise hidden in traditional quantitative analysis, as emphasized by Kunda and colleagues (2016) and empirically demonstrated by Belelli and colleagues (2010). In particular, IC and R errors reveal specific vulnerabilities demonstrating how different cognitive processes underlie superficially similar levels of performance. It is important to note that different error types differ in severity and implicated cognitive processes: Difference (D) and Wrong Principle (WP) errors are considered the most severe, reflecting fundamental failure to understand the logical problem; Repetition (R) errors are intermediate, attributable to simple associations; while Incomplete Correlate (IC) errors are the least severe, representing partially correct responses (Belelli et al., 2010). In terms of processes involved, D, WP, and R errors primarily implicate logical and visuospatial analyses,

whereas IC errors more directly involve visuoperceptual processes related to the orientation and direction of elements (Kunda et al., 2016).

Specifically, IC errors occur when the selected distractor approximates the correct answer but differs in a perceptual feature (orientation, direction, number of elements; Raven et al., 2003), representing a partially adequate response in which the individual has identified the solution at least in part but cannot fully utilize it (Kunda et al., 2016). Belelli et al. (2010) found that children with moderate SLD, despite comparable performance in overall accuracy, exhibited higher frequency of IC errors, interpreted as reflecting difficulties in visuospatial working memory. Consistently, Belacchi and Cornoldi (2009) documented that IC errors increase with age in the general population as children attempt more sophisticated reasoning strategies that place greater demands on working memory capacity. The theoretical relevance of IC errors to working memory functioning was well established by Carpenter et al. (1990), who demonstrated in the Advanced Progressive Matrices that good performance is associated with the ability to actively maintain task resolution rules in memory and with effective control and monitoring in the use of rules and information present in the problem. This underscores the crucial importance of executive functions, particularly the control component of working memory. It is plausible that the greater the attentional effort required by the matrix, the greater the number of IC errors committed. The elevated frequency of IC errors in children with SLD can be explained through multiple interconnected cognitive mechanisms. First, working memory deficits, which have been consistently found in this population (Cornoldi et al., 2014, 2019a; Poletti, 2016; Toffalini et al., 2017; Willcutt et al., 2013), could prevent actively maintaining all necessary rules or lead to forgetting them at the crucial moment when distractor characteristics must be compared with matrix features to select the correct alternative. This operation is particularly demanding from a working memory perspective: if capacity is limited, the individual might select a distractor similar to the correct one in all but one feature, corresponding exactly to an IC error. Second,

factors related to impulsivity (Donfrancesco et al., 2005; Purvis & Tannock, 2000; Venkatesan & Lokesh, 2019) or distractor position could intervene, leading the individual to rapidly select among the first similar elements without systematically exploring all available options. Third, deficits in attentional systems (Bolster et al., 1986; Crisci et al., 2021), specifically the Central Executive (Baddeley & Hitch, 1974) and the Supervisory Attentional System (Shallice, 1988), could compromise strategy selection, maintenance of activation of relevant information, inhibition of irrelevant material, and the constant "updating" of working memory that must replace obsolete information with new and more pertinent material.

Concerning the Repetition errors (R), it's known that they occur when the selected distractor replicates an element adjacent to the blank space, representing an inadequate intrusion response caused by dependence on the matrix space (Belacchi et al., 2012; Kunda et al., 2016). This pattern suggests a failure to grasp relationships between elements across rows and/or columns of the matrix: it is as if the matrix contained too much information, leading attention to focus on elements immediately surrounding the missing cell rather than on the overall relational structure (Carpenter et al., 1990). Belacchi and Cornoldi (2009) showed that in typical development, R errors increase until approximately age 7 and then decline progressively. This pattern might reflect the fact that in some items the choice of the alternative adjacent to the target is adequate, leading younger children to inappropriately generalize this strategy. With maturation and development of working memory capacity, children become progressively capable of maintaining a holistic view of the matrix structure. The persistence of elevated R errors in children with SLD, even in the older age range, could suggest developmental delay or an atypical trajectory. This is consistent with evidence from clinical populations: Belacchi et al. (2012) found R errors more common in children with Down Syndrome, interpreting them as evidence of reduced processing capacity, similar to studies on neurological damage, psychiatric disorders, and intellectual disability (Goharpey et al., 2013). The most plausible

explanation concerns the crucial role of working memory abilities: weaker and deficient capacities would lead to poorer performance with a singular error pattern distant from that observed in typically developing peers of matched mental age (Carpenter et al., 1990).

The joint presence of IC and R errors in SLD suggests that difficulties encompass both the capacity to maintain and manipulate information in working memory with adequate executive control (IC) and the ability to maintain an integrated and holistic representation of complex visual information while avoiding cognitive overload (R). These vulnerabilities extend beyond domain-specific academic skills to encompass more general aspects of executive functioning and information processing.

Regarding the third objective, age effects emerged differently across the two test versions. In the MatriKS 4–11 version, age was a significant predictor of error frequency ($p = .002$), indicating that error patterns varied systematically with age in the younger group. Conversely, in the MatriKS 12+ version, no significant age effect was detected ($p = .240$). This differential pattern, with marginally significant effects for both accuracy ($p = .064$) and error patterns ($p = .064$) in the younger group, but absence of significant effects in the older group (accuracy: $p = .196$; error interaction: $p = .702$), raises important questions admitting alternative interpretations with both theoretical and methodological implications.

A first interpretation suggests that the gap in fluid reasoning abilities between TD and SLD groups genuinely narrows with age, reflecting processes of compensation and maturation. Adolescents with SLD might develop strategic approaches to complex reasoning tasks that partially compensate for their working memory limitations. This developmental convergence hypothesis would be consistent with evidence that executive functions, including working memory and fluid intelligence continue to develop throughout childhood and adolescence (Fry & Hale, 2000), and that individuals with learning difficulties may show catch-up growth in some cognitive domains. With age and experience, children might develop metacognitive strategies for approaching matrix tasks, such as systematic analysis of

row and column relationships, that reduce reliance on working memory capacity alone. However, this hypothesis must be evaluated cautiously given the cross-sectional design of the study; only longitudinal research could definitively establish whether individual trajectories show genuine convergence or whether the pattern reflects group differences.

An alternative interpretation focuses on the psychometric properties of the MatriKS 12+ version. Preliminary observations suggest that this version contains items of considerable difficulty that might challenge even high-performing adolescents, potentially creating a floor effect that obscures group differences. If the test is excessively difficult for the target age range, both TD and SLD adolescents might perform suboptimally, reducing sensitivity to detect meaningful differences. This interpretation is supported by the error pattern data: in the MatriKS 12+ version, Wrong Principle (WP) errors, which reflect fundamental misunderstanding of the logical relationships in the matrix (Raven et al., 2003; Kunda et al., 2016), were among the most frequent error types in both groups. The high frequency of WP errors suggests that many items exceeded the optimal difficulty range, leading both groups to guess or apply incorrect rules. In contrast, WP errors were less common in the younger age group, where items appear more appropriately calibrated. In other words, MatriKS appears sensitive to measuring fluid intelligence improvement during childhood but not adolescence, potentially because the 12+ version contains very difficult items that may have particularly penalized younger adolescents.

A third consideration concerns some limitations of the present study. The relatively modest sample may have limited statistical power to detect subtle age-related effects and group differences. This is particularly relevant given that the 12+ version spans a wide developmental range (12–17.11 years), a period during which substantial cognitive maturation occurs. The current analytical approach, which treated participants by test version rather than subdividing by narrower age bands, may have obscured more fine-grained developmental patterns. Future studies with larger samples would enable more

sophisticated analytical strategies, such as dividing participants into narrower age groups (e.g., 12–13 years; 14-15; 16–17 years) which could reveal whether developmental trajectories differ systematically between TD and SLD groups across adolescence. Such approaches would help distinguish genuine developmental convergence from psychometric limitations and constraints related to sample size.

Regardless of which interpretation is more appropriate, these findings suggest that the MatriKS 12+ version would benefit from psychometric refinement. Systematic item analysis examining difficulty, discrimination, and distractor functioning could identify suboptimal items. Revisions might include: (1) removing or modifying excessively difficult items, (2) adding easier items to create a more balanced difficulty range, (3) potentially developing separate versions for younger and older adolescents. Such revisions could enhance the test's discriminative validity, making it more useful for both research and clinical assessment. Future studies should include larger samples to enable more sophisticated psychometric analyses, including more advanced approaches that can provide detailed information about item functioning across the ability continuum.

Regarding the fourth objective, exploratory correlation analyses revealed no significant relationships between MatriKS performance in terms of accuracy and other cognitive or academic measures in the current sample (except for comprehension subtest and MatriKS 12+). Overall, this finding diverges from the body of literature documenting substantial relationships between fluid intelligence and academic achievement (Green et al., 2017; Johann et al., 2020; Passolunghi et al., 2022; Peng et al., 2019). However, the non significant correlations are most plausibly attributable to the study limitations rather than a real construct independence. Three factors appear particularly relevant. First, the relatively small sample size (N = 106), divided into subgroups (SLD and TD) and further subdivided by age range according to the MatriKS 4–11 and MatriKS 12+ versions, limited the statistical power to detect potential correlations. Second, data availability was incomplete, as not all

participants completed the full battery of neuropsychological assessments due to strictly clinical constraints. Finally, sample heterogeneity, comprising typically developing children and children with SLD presenting different primary areas of academic difficulty (reading, writing, mathematics, or mixed profiles) but analyzed as a single group, may have obscured domain-specific relationships. Future studies with larger samples, more comprehensive assessment protocols, and separate analyses for different SLD subtypes would be valuable to clarify the relationships between MatriKS performance and other cognitive and academic abilities, as well as to examine whether error patterns show distinct associations with specific cognitive functions.

This constellation of findings demonstrates that children with SLD do not exhibit global intellectual impairment but rather show weaknesses in nonverbal logical reasoning processes that underline complex reasoning and, consequently, academic learning. The dissociation between quantitative performance (overall accuracy) and qualitative patterns (error types) underscores that comprehensive assessment must extend beyond global scores to capture clinically meaningful cognitive vulnerabilities. These findings, although preliminary and to be interpreted with caution given the study's limitations, suggest important implications for both the theoretical understanding of learning disorders and for clinical practice in assessment and intervention.

From a clinical perspective, these general findings suggest that comprehensive assessment of children with suspected or diagnosed SLD should include qualitative analysis of test performance, not merely quantitative scores. For instance, error analysis in tasks like MatriKS or traditional Raven's Matrices can reveal specific cognitive vulnerabilities, such as working memory limitations or visuospatial processing difficulties, that may be highly relevant for intervention planning. For example, children who commit frequent IC errors might benefit from interventions targeting visuospatial working memory, such as training in visualization strategies or use of external memory aids. On the other hand, children who commit frequent R errors might benefit from training in strategic problem-solving

approaches that encourage systematic and holistic analysis rather than focusing on locally salient features. Moreover, understanding error patterns can help clinicians differentiate between children whose academic difficulties stem primarily from domain-specific deficits (e.g., phonological processing in dyslexia) versus those whose difficulties reflect more general cognitive limitations that affect learning across domains. Children with substantial working memory limitations, as evidenced by frequent IC and R errors, may require more intensive academic accommodations and a focus on strategies that reduce cognitive load during learning.

Finally, the computerized format of MatriKS offers distinct advantages for both research and clinical practice. For instance, automated data collection ensures precise recording of response patterns, reaction times, and error types, reducing human scoring errors and facilitating not only clinical assessment but also the conduct of large-scale studies. The digital format also opens possibilities for adaptive testing paradigms that could adjust item difficulty to individual ability levels, potentially improving diagnostic precision across a broader range of cognitive functioning. Future developments could include real-time tracking of problem-solving processes, such as response latencies for different error types, which could provide additional insights into the strategies children employ when compared with difficult reasoning problems. However, several limitations of the present study merit consideration and suggest directions for future research.

As previously mentioned, the sample size was relatively small, particularly when divided into two age groups and two diagnostic categories. This limited statistical power to detect group differences and likely contributed to the marginally significant results that approached but did not reach conventional significance thresholds. The consistent pattern of medium effect sizes across age groups suggests that meaningful differences exist, but replication with larger samples (ideally N = 150 per group to achieve 80% power for detecting medium effect sizes) is necessary to confirm these effects with adequate statistical confidence.

Second, the SLD group was heterogeneous, comprising children with different primary areas of academic difficulty (reading, writing, mathematics, or combinations). While this heterogeneity reflects the clinical reality that SLD encompasses different functioning profiles with diverse manifestations, it may have obscured subgroup-specific patterns. Future research should examine error patterns separately for different SLD subtypes, such as dyslexia, dysorthography, dyscalculia, and mixed learning disorders. For example, given evidence that mathematical SLD is associated with greater visuospatial working memory deficits while reading SLD is more strongly linked to verbal working memory deficits (Siegel & Ryan, 1989; Poletti, 2016; Toffalini et al., 2017), these different working memory profiles might translate into distinct error patterns on MatriKS. Children with dyscalculia might show particularly elevated IC errors due to visuospatial processing demands, while children with dyslexia might show relatively fewer IC errors but difficulties on items requiring verbal mediation of relationships.

Another aspect could be that the cross-sectional design precludes conclusions about developmental trajectories. Longitudinal studies following the same children over time from middle childhood through adolescence would clarify whether reduced group differences in adolescence reflect genuine developmental convergence, and whether specific error patterns in childhood predict later academic outcomes. Such studies could identify early markers of risk for persistent learning difficulties and could examine whether interventions targeting working memory or fluid reasoning abilities in childhood have lasting effects on academic achievement.

Moreover, as discussed extensively above, the psychometric properties of the MatriKS 12+ version require further refinement. The high frequency of WP errors in both TD and SLD groups, combined with the absence of age effects, suggests that item difficulty may not be optimally calibrated for the adolescent age range. Systematic item analysis using modern psychometric approaches such as Item Response Theory, followed by item revision or replacement, could improve the test's discriminative

validity. Moreover, the possibility of developing separate normative standards or even distinct test versions for younger (12–14 years) versus older (15–17 years) adolescents should be considered, given the substantial cognitive development that occurs during this period.

Furthermore, another limitation of the study concerns the incomplete data availability across participants. As previously described, not all children completed the entire battery of neuropsychological and academic assessments, reducing both statistical power and the completeness of conclusions about construct relationships. Future research should employ more comprehensive and standardized assessment protocols administered to all participants, allowing more definitive conclusions about the relationships between MatriKS performance, working memory, processing speed, and academic achievement. Such research could also examine whether MatriKS error patterns provide incremental validity beyond traditional neuropsychological measures in predicting academic outcomes.

Finally, although the current study focused on comparing TD and SLD groups, it would be valuable to examine MatriKS performance in other clinical populations, such as children with Attention-Deficit/Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), or Intellectual Disability (ID). Comparative studies across different neurodevelopmental disorders could clarify whether IC and R errors are specifically elevated in SLD or whether they reflect more general markers of cognitive limitations that cross diagnostic categories. Such research could also examine whether different clinical populations show distinct error profiles; for example, whether children with ADHD show elevated R errors due to attentional difficulties, while children with autism or intellectual disability show elevated WP errors due to difficulties in understanding and flexibly applying rules. Additionally, future research could examine further variables that might moderate the relationship between SLD status and MatriKS performance. For example, factors such as socioeconomic status, language background, intervention history, and comorbid conditions could all influence both fluid

reasoning performance in terms of accuracy and error patterns. Understanding these moderating effects would provide a more nuanced picture of cognitive functioning in children with SLD and could be more informative to plan more individualized assessment and intervention approaches.

6. Conclusion

In conclusion, the present study demonstrates that MatriKS, a computerized adaptation of Raven's Progressive Matrices, can detect subtle but meaningful differences between children with SLD and their TD peers when qualitative error analysis is incorporated alongside traditional accuracy measures. Although overall accuracy did not differ significantly between groups, the elevated frequency of Incomplete Correlate and Repetition errors in children with SLD provides evidence of working memory and visuospatial processing difficulties that may contribute to their academic challenges. These findings are consistent with previous research using traditional Raven's Matrices (Belelli et al., 2010) and extend this work by demonstrating similar patterns with a computerized assessment tool that offers advantages in terms of automated scoring, precise data recording, and potential for adaptive testing.

For children with SLD, who by definition show average intellectual functioning, such fine-grained analysis may be particularly valuable for identifying subtle cognitive weaknesses that contribute to academic difficulties. The differential results between age groups highlight important questions about both developmental trajectories and test characteristics that merit further investigation. Whether reduced group differences in adolescence reflect genuine developmental convergence or psychometric limitations remains an open question that longitudinal research could address.

As computerized assessment tools become increasingly prevalent in clinical practice, MatriKS represents a promising advancement that leverages digital technology to enable automated administration, precise data recording, and potential for adaptive testing algorithms. With continued psychometric refinement and validation in larger, more diverse samples, including specific SLD

subtypes and other neurodevelopmental populations, MatriKS has the potential to become a valuable tool for both research and clinical assessment of cognitive functioning in children and adolescents with learning difficulties.

Ultimately, this study advances our understanding of the cognitive profiles associated with SLD by revealing that qualitative error analysis can detect clinically significant processing differences beyond global performance metrics, this research provides both theoretical insights and practical guidance for developing more targeted, effective interventions that address the multifaceted nature of learning difficulties and support optimal educational outcomes.

General Discussion and Conclusions

As outlined in the introduction, this dissertation was conceived within a broader research project pursuing three interrelated goals: 1) to develop a new generation of adaptive neuropsychological tests assessing planning abilities and fluid intelligence; 2) to create an intelligent web-based system capable of administering these tests, analyzing data, and providing personalized reports; 3) to implement innovative methodological approaches that allow for efficient evaluation and individualized adaptation through dynamic item selection based on a person's performance.

These aims converged in the development of PsycAssist (de Chiusole et al., 2024), an artificial intelligence–based web platform designed to provide comprehensive adaptive neuropsychological assessment with automated scoring and clinical report generation. The platform integrates two principal instruments. Adap-ToL, an adaptive version of the Tower of London task for the evaluation of planning abilities, and MatriKS, a computerized adaptive test based on Raven-like matrices for the assessment of fluid intelligence.

The dissertation was structured into two complementary parts. The first part established the theoretical background in four sections: the evolution of neuropsychological assessment from traditional to digital adaptive approaches (Section 1); executive functions as higher-order cognitive abilities, with specific focus on planning abilities and the Tower of London task (Section 2); fluid intelligence through major theoretical frameworks and traditional instruments such as Raven's Matrices (Section 3); and an integrated analysis of these cognitive domains in clinical populations, with particular focus on Specific Learning Disorders (Section 4).

The second part presented five empirical studies that built progressively upon one another, illustrating how technological innovation, psychometric rigor, and cognitive modeling can be meaningfully combined within a unified system: the introduction of PsycAssist technological architecture and preliminary findings on usability and acceptability (Study 1); the development and validation of the

Usability and Attitude Scale for children and adolescents (Study 2); the multi-method psychometric validation of MatriKS (Study 3); the process-oriented analysis of planning through Markov models (Study 4); and the examination of error patterns in MatriKS in children and adolescents with SLD (Study 5).

Collectively, these studies demonstrate that advancing neuropsychological assessment requires not only new instruments and computational methods but also a coherent integration of usability, psychometric validity, and process-oriented understanding.

The first study laid the technological and conceptual foundations of this work by introducing PsycAssist as an innovative web-based system designed to support adaptive neuropsychological assessment and cognitive training. The platform's adaptive engine, based on Procedural Knowledge Space Theory (PKST; Stefanutti, 2019), employs deterministic and probabilistic models to organize individuals according to their problem-solving performance, offering a sophisticated representation of cognitive functioning that extends beyond simple scoring. The Web-based architecture allows psychologists and neuropsychologists to access and use the platform from any location and device, representing a substantial advancement in terms of flexibility and accessibility for both research and clinical practice.

The preliminary usability study confirmed a generally positive reception of the platform and its instruments. Evaluators judged both Adap-ToL and MatriKS to be appropriate and effective for their intended purposes, while participants described the assessment experience as engaging, comprehensible, and motivating. These results underline several clinical advantages of digital adaptive assessment. Automated scoring and report generation significantly reduce examiner workload and the potential for human error, thereby enhancing both efficiency and reliability. Moreover, the interactive digital format captures and sustains the participant's attention more effectively than traditional paper-and-pencil procedures. This advantage is particularly relevant for

individuals who may find conventional testing monotonous or frustrating, as the adaptive system can maintain an optimal level of challenge tailored to the individual's ability.

The combination of adaptive administration and automated reporting not only streamlines the assessment process but also generates data that can inform personalized interpretation. The feedback produced by PsycAssist provides valuable insights into individual cognitive strengths and weaknesses, facilitating the development of targeted training programs and enabling ongoing monitoring of intervention effectiveness through dynamic, data-driven adjustments.

Despite the generally positive findings, the study also identified several areas for improvement that can guide future refinements. The sample composition represented a notable limitation, as most participants were recruited from schools (N = 902) and only a smaller portion were adults (N = 337), suggesting the need for broader age representation in future research. Additionally, since the examiners who administered the assessments were not healthcare professionals, subsequent studies involving practicing clinicians will be crucial to verify the instruments' acceptability and practical integration into everyday clinical practices. Finally, the versions of Adap-ToL and MatriKS used in this study were preliminary, with psychometric validations still ongoing. Nevertheless, both participants' and evaluators' feedback indicated strong potential for these tools, confirming that the adaptive and user-centered principles guiding their design are well-suited to contemporary neuropsychological assessment needs.

Recognizing that technological innovation alone is not sufficient to ensure successful implementation, the second study addressed a crucial methodological gap by developing and validating the Usability and Attitude Scale (UAS), a psychometric instrument specifically designed to evaluate the usability and acceptability of digital technologies among children and adolescents. When introducing new digital assessment tools, it is essential not only to establish their psychometric soundness but also to assess how users perceive and interact with them. Usability and attitudes toward technology,

particularly perceptions of ease of use and engagement, represent fundamental determinants of acceptance and long-term adoption.

The UAS was conceived to capture these dimensions systematically and to provide a reliable instrument applicable to developmental populations, for whom existing usability measures—typically designed for adults—are often inappropriate due to linguistic or cognitive demands. The study employed both exploratory (EFA) and confirmatory factor analyses (CFA) to examine the structure and psychometric properties of the scale. Results revealed a coherent two-factor structure, distinguishing between usability and acceptability, with items loading strongly on their intended dimensions and eigenvalues supporting the hypothesized model. The CFA confirmed the adequacy of this structure, with fit indices indicating good model fit, thus reinforcing the scale's structural validity. Internal consistency, as indexed by Cronbach's alpha, was satisfactory for both subscales, providing further evidence of reliability. These findings align closely with the theoretical premises of the Technology Acceptance Model (TAM) and the System Usability Scale (SUS), both of which emphasize perceived ease of use and user attitude as key predictors of technological acceptance. The successful adaptation of these constructs to younger populations represents a significant contribution, demonstrating that children and adolescents are capable of providing valid metacognitive evaluations of their digital experiences when the instruments are designed appropriately for their developmental level. The validation of the UAS thus constitutes an important step forward in the field of digital assessment and educational technology. By offering a psychometrically robust instrument tailored to younger users, the UAS provides researchers, clinicians, and developers with a tool for systematically assessing how children and adolescents perceive and interact with digital systems. This allows not only for the identification of usability issues that may hinder engagement but also for the continuous improvement of digital interventions and assessment tools, ensuring that they remain both effective and motivating.

While the results were encouraging, the study acknowledged several limitations that suggest directions for future research. The data were collected exclusively in Italy, which may limit the cultural generalizability of the findings. Given that cultural factors can influence familiarity with technology, communication styles, and motivational responses, it would be valuable to replicate the validation process in diverse cultural contexts. Furthermore, although the UAS captures the essential dimensions of usability and acceptability, future refinements could expand the instrument to include additional facets such as long-term engagement, enjoyment, and satisfaction. Longitudinal research designs would also be beneficial for examining how users' perceptions evolve over time and with repeated exposure to digital tools, providing a more comprehensive understanding of sustained acceptance and engagement. Finally, convergent validation with other established measures—such as the SUS or TAM-based questionnaires adapted for developmental age—would strengthen the construct validity of the UAS and situate it within a broader network of usability research. Overall, the second study demonstrated that rigorous psychometric validation of digital tools must include attention not only to measurement precision but also to user experience, particularly in pediatric and adolescent populations. By doing so, it ensures that technology functions as a genuinely facilitative and accessible instrument for cognitive evaluation and intervention, rather than as a source of additional cognitive or emotional burden.

While the establishment of user acceptance is an essential step in the development of any digital assessment tool, demonstrating psychometric robustness remains a fundamental requirement. The third study addressed this need by providing a comprehensive, multi-method validation of MatriKS, the computerized adaptive test designed to assess fluid intelligence through Raven-like matrices. This validation adopted an integrative methodological framework that combined Classical Test Theory (CTT), Rasch modeling, and Knowledge Space Theory (KST), allowing a multifaceted evaluation of the test's reliability, validity, and diagnostic precision. The use of multiple analytic perspectives

represents a significant methodological advancement in the development of psychological and neuropsychological tests. Each approach offers distinct yet complementary insights: CTT assesses internal consistency and construct validity; Rasch modeling examines the precision of measurement across ability levels and the fit between item difficulty and respondent ability; and KST provides a fine-grained, person-centered analysis of cognitive structures and developmental trajectories. Together, these methods offer a triangulated and comprehensive assessment of psychometric soundness. The results confirmed that MatriKS performs well across all three analytical frameworks. Under CTT, the test demonstrated satisfactory reliability and convergent validity, supporting its measurement consistency and theoretical coherence. Rasch modeling revealed good item fit and adequate targeting across ability levels, ensuring that the adaptive algorithm could accurately adjust item presentation in response to performance. Finally, KST-based analysis provided a more nuanced view of how individuals acquire and integrate cognitive skills underlying fluid reasoning, illustrating the system's potential for detailed diagnostic feedback. Developmental trends observed in the MatriKS data further supported its external validity. Average scores on fluid intelligence increased steadily between ages 4 and 11, while score variability decreased with age, reflecting developmental stabilization consistent with previous literature (Horn, 2007; Schroeders et al., 2015). The convergence of evidence across CTT, Rasch, and KST frameworks therefore strengthens confidence in MatriKS as a valid and reliable instrument for the assessment of fluid intelligence, particularly within developmental populations. Beyond its psychometric performance, MatriKS demonstrated significant potential for clinical and educational applications. Its capacity to adapt to the individual's ability level ensures that each examinee receives an appropriately challenging set of items, minimizing frustration and disengagement while optimizing measurement precision. Moreover, the KST-based analysis produces individualized feedback profiles that highlight both cognitive strengths and areas for improvement. This information can support clinicians and educators in designing

personalized interventions, monitoring progress, and tailoring cognitive training to each child's unique developmental trajectory.

The study also outlined several promising directions for future research. Extending MatriKS to populations with diverse neurodevelopmental or learning profiles—such as Specific Learning Disorders, ADHD, or autism spectrum conditions—would test its diagnostic sensitivity and ecological validity across clinical groups. Furthermore, developing a fully adaptive version capable of dynamically selecting items based on real-time performance would further enhance both efficiency and measurement accuracy. Integrating MatriKS into intervention programs could also enable longitudinal monitoring of cognitive changes, allowing practitioners to evaluate the effectiveness of training and to observe how reasoning strategies evolve over time.

In summary, the third study confirmed that MatriKS is a psychometrically sound, developmentally sensitive, and conceptually innovative instrument. Its methodological triangulation not only strengthens its scientific foundations but also exemplifies a rigorous approach that future test developers can adopt. By combining traditional psychometric models with probabilistic and structural frameworks such as Rasch and KST, this study demonstrates how digital adaptive assessment can achieve both measurement precision and interpretive depth, offering clinicians richer and more actionable insights into cognitive functioning.

The fourth study represented a crucial step toward a process-oriented understanding of cognitive assessment, focusing on the analysis of planning abilities through Markov modeling applied to performance data from the Adap-ToL task. Whereas traditional approaches to neuropsychological evaluation often rely on static outcome measures—such as total scores or number of moves—this study sought to capture the dynamic structure of problem-solving behavior, modeling how individuals plan, adjust, and execute actions over time. The Tower of London paradigm, from which Adap-ToL derives, has long been recognized as a sensitive measure of planning and executive functioning.

However, its diagnostic potential is often limited by the reliance on aggregate performance indices, which fail to reflect the temporal organization of problem-solving processes. By adopting Markov models, the study introduced a novel analytic framework capable of describing cognitive operations as sequences of probabilistic transitions between different states of the problem space. This approach allows for a richer and more dynamic interpretation of planning behavior, emphasizing how individuals reach a solution rather than merely whether they succeed. The findings provided several important insights. Markov modeling revealed clear differences in transition probabilities across difficulty levels, reflecting the varying cognitive demands of tasks with increasing complexity. As expected, higher task difficulty was associated with longer solution times, an increased number of intermediate moves, and greater variability in transition patterns. These results suggest that as problems become more complex, participants engage in more exploratory and less stereotyped strategies, reflecting a shift from routine procedural behavior to flexible, goal-directed reasoning. A particularly noteworthy outcome of the analysis was the identification of distinct problem-solving profiles, which emerged from differences in the organization and efficiency of transitions among problem states. Some participants demonstrated highly structured and efficient strategies characterized by a small number of states and predictable transitions, while others displayed more dispersed and less stable patterns, indicating trial-and-error approaches. These differences are not merely behavioral but reflect underlying variations in cognitive control, working memory, and anticipatory planning—core components of executive functioning.

The process-oriented methodology adopted in this study carries significant implications for both theory and practice. From a theoretical standpoint, it supports the view that planning should be conceptualized as a dynamic system, governed by continuous feedback between internal representations of goals and external action outcomes. This aligns with contemporary models of executive control emphasizing adaptive regulation and self-monitoring rather than static hierarchical

command structures. From a practical perspective, the application of Markov modeling allows clinicians and researchers to go beyond global performance indices and to examine individual differences in problem-solving processes. Such analyses can reveal subtle impairments that may remain undetected in traditional scoring systems, providing a more precise and ecologically valid picture of executive functioning.

Moreover, the integration of Markov modeling within the Adap-ToL system opens new possibilities for adaptive assessment and intervention. By identifying transition patterns indicative of suboptimal planning, the system could, in future iterations, adapt task presentation in real time to target specific weaknesses, thus transforming assessment into a form of interactive cognitive training. This dynamic adaptation would further personalize the testing experience, aligning with the broader objectives of PsycAssist to combine assessment precision with individualized feedback and cognitive support.

Despite its innovative contributions, the study also highlighted methodological challenges. Markov modeling requires large datasets to ensure stability of transition estimates and careful definition of state spaces to avoid oversimplification. Future research could address these challenges by employing more complex hierarchical models, integrating behavioral and temporal data, and exploring cross-task generalization of transition patterns. Additionally, combining Markov modeling with eye-tracking or neural measures could provide a multimodal understanding of planning, linking cognitive states with attentional and neural dynamics.

In sum, this fourth study demonstrated that process-oriented analysis, when grounded in formal probabilistic modeling, offers a powerful framework for understanding the mechanisms underlying executive functions. By revealing the temporal architecture of planning behavior, Markov models move beyond outcome-based assessment, capturing the unfolding of cognitive control in real time. This represents a meaningful step toward a new generation of neuropsychological tools capable not only of measuring performance but also of explaining the processes that produce it.

The fifth study extended the scope of this research by applying the full standard version of MatriKS to a clinical population of children and adolescents with Specific Learning Disorders (SLD). The main aim was to investigate whether MatriKS, originally validated in typically developing populations, could effectively capture the cognitive profiles and reasoning processes characteristic of individuals with SLD, with a particular focus on error patterns. This study adopted a qualitative, process-oriented approach, emphasizing error analysis rather than overall performance, in order to better understand how children with SLD approach complex reasoning tasks. The specific objectives included: (1) comparing total accuracy scores in MatriKS between children with SLD and typically developing peers, while also observing the overall distribution of errors; (2) examining potential differences in the proportion of error types (Repetition, Wrong Principle, Difference, Incomplete Correlate) between SLD and control groups; (3) investigating the effect of age on these measures, including age as a covariate to account for developmental influences; and (4) examining the relationship between fluid intelligence, as measured by MatriKS, and learning skills (reading, writing, mathematics, and comprehension), as well as domain-general cognitive abilities (such as attention, working memory, and logical reasoning), assessed using standardized test batteries with the most relevant subtests selected for each cognitive domain. This approach enables the detection of underlying reasoning processes, which are often hidden in traditional aggregated scoring systems. Although overall accuracy scores do not show significant differences between children with SLD and typically developing peers, error pattern analysis suggests the potential presence of inefficiencies in reasoning processes related to fluid intelligence, which may contribute to their academic difficulties. From a clinical perspective, error analysis can reveal specific cognitive vulnerabilities highly relevant for intervention planning. For example, children with frequent Incomplete Correlate errors might benefit from interventions targeting visuospatial working memory, while those with frequent Repetition errors might benefit from training in systematic, holistic problem-solving approaches.

Additionally, the digital format of MatriKS offers distinct advantages through automated and precise recording of response patterns, reducing human scoring errors and opening possibilities for adaptive testing paradigms and real-time tracking of problem-solving processes.

It is important to note that data collection for this study is still ongoing, and the current sample is relatively small, which limits statistical power and generalizability. The SLD group was heterogeneous, comprising children with different primary areas of academic difficulty, which may have obscured subgroup-specific patterns. As more participants are included, the study will allow for more robust analyses. Future investigations could further benefit from dividing the SLD sample into subgroups, such as reading/writing versus mathematics difficulties, or more specifically dyslexia, dysorthographia, and dyscalculia, to better delineate the cognitive profiles underlying different diagnostic presentations and examine whether different SLD subtypes show distinct error patterns. For instance, children with dyscalculia might show particularly elevated Incomplete Correlate errors due to greater visuospatial working memory deficits, while children with dyslexia might show different patterns reflecting their verbal working memory profiles. Additionally, expanding the administration of MatriKS to other clinical populations, such as ADHD, autism spectrum disorder, or intellectual disability, would provide the opportunity to evaluate its utility in capturing diverse reasoning patterns and error profiles across developmental disorders. demonstrating highlights the methodological and clinical value of extending digital assessment tools to clinical populations, demonstrating that error pattern analysis can enrich the understanding of fluid reasoning abilities and the cognitive processes involved in SLD, while paving the way for broader applications in developmental neuropsychology.

Taken together, these five studies presented in this dissertation trace a coherent and progressively integrated path toward a new paradigm in neuropsychological assessment one that harmonizes technological innovation, psychometric rigor, and process-oriented understanding of cognition. Each

study contributes a distinct yet complementary piece to this vision: establishing the technological infrastructure and confirming user acceptance (Study 1), developing valid tools to measure usability and acceptability in young populations (Study 2), ensuring psychometric robustness of adaptive reasoning assessment (Study 3), revealing the temporal dynamics of planning through formal modeling (Study 4), and demonstrating clinical applicability through error pattern analysis in SLD (Study 5). PsycAssist, in its current implementation, hosts two principal instruments, Adap-ToL and MatriKS, enabling the assessment of planning abilities and fluid intelligence through adaptive, algorithmically guided procedures. However, the platform's architecture was deliberately designed as a modular and scalable ecosystem, with the prospect of including additional tests for the assessment of other cognitive functions, as well as personalized intervention programs, and cognitive training such as serious games, in the future. Such expansion would transform PsycAssist from a specialized assessment platform into a comprehensive digital environment supporting the entire cycle of neuropsychological practice—from initial evaluation through targeted intervention and longitudinal monitoring. This work has demonstrated that it is possible to capture not only performance outcomes but also the strategies and cognitive mechanisms underlying behavior, opening new perspectives for targeted educational and clinical interventions. The digitization and adaptivity of the tests represent a conceptual leap compared to traditional assessment, transforming technology from a mere delivery medium into a methodological tool capable of observing the mind in action. By combining adaptive test administration with formal cognitive modeling and detailed process analysis, digital assessment tools become instruments capable of observing cognition as it unfolds in real time, rather than merely documenting its outcomes. This represents a fundamental shift: technology is no longer simply a delivery medium but an active methodological component that enables new forms of measurement and understanding.

It is important to emphasize that this project represents the foundation of a longer trajectory rather than its culmination. Data collection is still ongoing, and clinical samples need to be expanded to enable more powerful statistical analyses and firmer conclusions. Dividing participants into diagnostic subgroups will allow a clearer delineation of specific cognitive profiles associated with different neurodevelopmental conditions. At the same time, extending the platform to other clinical populations, including ADHD, autism spectrum disorder, intellectual disability, and psychiatric disorders, will test the generalizability and clinical utility of these instruments across diverse populations. Moreover, integrating multimodal data, behavioral, physiological, and neural could offer extraordinary opportunities to push the boundaries of cognitive assessment even further, providing converging evidence linking cognitive processes to their neural substrates and enriching both theoretical models and clinical interpretation. Looking forward, the platform could incorporate real-time adaptive intervention, where assessment seamlessly transitions into personalized cognitive training based on identified weaknesses. Machine learning algorithms could identify subtle patterns in assessment data that predict intervention response or developmental trajectories, supporting more precise prognosis and treatment planning. Longitudinal tracking capabilities could document cognitive change over time, whether due to development, intervention, or neurological condition, providing clinicians with dynamic rather than static snapshots of functioning.

In conclusion, this dissertation establishes both the feasibility and value of digitally mediated, adaptive, process-oriented neuropsychological assessment. The convergence of psychometric sophistication, computational modeling, and user-centered design principles creates assessment tools that are simultaneously more rigorous, more informative, and more engaging than their traditional counterparts. This work lays the groundwork for a new generation of digital neuropsychological tools, capable of combining precision, interpretive depth, and personalization. The transition from static measurement to adaptive, process-sensitive evaluation marks a genuine evolution in how cognitive

abilities are assessed and understood. As the neuropsychological community continues to embrace technological innovation while maintaining scientific rigor, the collective efforts of researchers, clinicians, and developers can shape a future where cognitive assessment provides not merely scores but genuine insight into the mechanisms of mind—insight that directly informs more effective, personalized support for individuals across the lifespan and across the full spectrum of cognitive abilities. While this work addresses one portion of the vast landscape of cognitive assessment, it demonstrates that the vision of truly personalized, process-oriented, technologically sophisticated neuropsychological assessment is not merely aspirational but achievable, opening a promising path toward a future where the understanding and support of cognitive processes will be more dynamic, integrated, and truly centered on the individual.

References

Ackerman, P. L. (2008). Knowledge and cognitive aging. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (3rd ed., pp. 445–489). Psychology Press.

Ahola, K., Vilkki, J., & Servo, A. (1996). Frontal tests do not detect frontal infarctions after ruptured intracranial aneurysm. *Brain and Cognition, 31*(1), 1–16. https://doi.org/10.1006/brcg.1996.0021

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Academiai Kiado.

Akyürek, G. (2018). *The effect of cognitive therapy on executive functions and occupational routines in children with dyslexia*. Archives of Physical Medicine and Rehabilitation.

Albert, D., & Steinberg, L. (2011). Age differences in strategic planning as indexed by the Tower of London. *Child Development, 82*(5), 1501–1517.https://doi.org/10.1111/j.1467-8624.2011.01613.x

Alderson, R. M., Rapport, M. D., Hudec, K. L., Sarver, D. E., & Kofler, M. J. (2010). Competing core processes in attention-deficit/hyperactivity disorder (ADHD): Do working memory deficiencies underlie behavioral inhibition deficits? *Journal of Abnormal Child Psychology, 38*(4), 497–507. https://doi.org/10.1007/s10802-010-9387-0

AlGhannam, B. A., Albustan, S. A., Al-Hassan, A. A., & Albustan, L. A. (2017). Towards a standard Arabic system usability scale: Psychometric evaluation using communication disorder app. *International Journal of Human–Computer Interaction, 34*(9), 799–804. https://doi.org/10.1080/10447318.2017.1388099

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology, 8*(2), 71–82. https://doi.org/10.1076/chin.8.2.71.8724

Anderson, P., Anderson, V., & Lajoie, G. (1996). The tower of London test: Validation and standardization for pediatric populations. *The Clinical Neuropsychologist, 10*(1), 54–65. https://doi.org/10.1080/13854049608406663

Anderson, P., & Reidy, N. (2012). Assessing executive function in preschoolers. *Neuropsychology Review, 22*, 345–360. https://doi.org/10.1007/s11065-012-9220-3

Andrès, P., & Van der Linden, M. (2000). Age-related differences in supervisory attentional system functions. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 55*(6), P373–P380. https://doi.org/10.1093/geronb/55.6.P373

Andrès, P., & Van der Linden, M. (2001). Supervisory attentional system in patients with focal frontal lesions. *Journal of Clinical and Experimental Neuropsychology, 23*(2), 225–239. https://doi.org/10.1076/jcen.23.2.225.1212

Andrés, P., & Van der Linden, M. (2002). Are central executive functions working in patients with focal frontal lesions? *Neuropsychologia, 40*(7), 835–845. https://doi.org/10.1016/s0028-3932(01)00182-8

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(5), 571–585. https://doi.org/10.1586/erp.11.59

Angelini, A. L. (1969). *Um novo método para avaliar a motivação humana: Teste de apercepção temática (TAT) de H. A. Murray*. Editora da Universidade de São Paulo.

Anselmi, P., Robusto, E., & Stefanutti, L. (2012). Uncovering the Best Skill Multimap by Constraining the Error Probabilities of the Gain-Loss Model. *Psychometrika, 77*(4), 763–781. doi:10.1007/s11336-012-9286-0

Anselmi, P., Robusto, E., Stefanutti, L., & de Chiusole, D. (2016). An Upgrading Procedure for Adaptive Assessment of Knowledge. *Psychometrika*, *81*(2), 461–482. doi:10.1007/s11336-016-9498-9

Anselmi, P., Stefanutti, L., de Chiusole, D., & Robusto, E. (2017). The assessment of knowledge and learning in competence spaces: The gain–loss model for dependent skills. *British Journal of Mathematical and Statistical Psychology*, *70*(3), 457–479. https://doi.org/10.1111/bmsp.12095

Anselmi, P., Stefanutti, L., Chiusole, D., & Robusto, E. (2021). Modeling learning in knowledge space theory through bivariate Markov processes. *Journal of Mathematical Psychology, 103*, 102549. https://doi.org/10.1016/j.jmp.2021.102549

Antonucci, G., Spitoni, G., Orsini, A., D'Olimpio, F., & Cantagallo, A. (2014). *Behavioural Assessment of the Dysexecutive Syndrome: La batteria per la valutazione dei deficit delle funzioni esecutive*.

Appollonio, I., Leone, M., Isella, V., Piamarta, F., Consoli, T., Villa, M. L., Forapani, E., Russo, A., & Nichelli, P. (2005). The Frontal Assessment Battery (FAB): Normative values in an Italian population sample. *Neurological Sciences, 26*(2), 108–116. https://doi.org/10.1007/s10072-005-0443-4

Arce-Ferrer, A. J., & Martinez Guzman, E. (2009). Studying the equivalence of computer-delivered and paper-based administrations of the Raven Standard Progressive Matrices test. *Educational and Psychological Measurement, 69*(6), 855–867. https://doi.org/10.1177/0013164409332219

Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement, 54*(2), 394–403. https://doi.org/10.1177/0013164494054002013

Asano, D., Takeda, M., Nobusako, S., & Morioka, S. (2023). Error analysis of Raven's Coloured Progressive Matrices in children and adolescents with cerebral palsy. *Journal of Intellectual Disability Research: JIDR*, *67*(7), 655–667. https://doi.org/10.1111/jir.13034

Asato, M. R., Sweeney, J. A., & Luna, B. (2006). Cognitive processes in the development of TOL performance. *Neuropsychologia, 44*(12), 2259–2269. https://doi.org/10.1016/j.neuropsychologia.2006.05.010

Atance, C. M., & Meltzoff, A. N. (2006). Preschoolers' current desires warp their choices for the future. *Psychological Science, 17*(7), 583–587. https://doi.org/10.1111/j.1467-9280.2006.01748.x

Axelrod, B. N. (2017). Validity of the Wechsler Abbreviated Scale of Intelligence and other very short forms of estimating intellectual functioning. *Assessment*, *9*(1), 17–23. https://doi.org/10.1177/1073191102009001003

Babcock, R. L. (2002). Analysis of age differences in types of errors on the Raven's Advanced Progressive Matrices. *Intelligence, 30*(5), 485–503. https://doi.org/10.1016/S0160-2896(02)00124-1

Baddeley, A., Della Sala, S., Gray, C., Papagno, C., & Spinnler, H. (1997). Testing central executive functioning with a pencil-and-paper test. In P. Rabbitt (Ed.), *Methodology of frontal and executive function* (pp. 61–80). Psychology Press.

Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience, 19*(12), 2082–2099. https://doi.org/10.1162/jocn.2007.19.12.2082

Bacherini, A., Anselmi, P., Havercamp, S., & Balboni, G. (2024). Psychometric properties of the Beliefs About Adults with ID Scale in American physicians: Application of classical test and Rasch measurement theories. *Journal of Intellectual Disability Research, 68*(8), 913–926. https://doi.org/10.1111/jir.13143

Baggetta, P., & Alessandro, P. A. (2016). Conceptualization and operationalization of executive function. *Mind, Brain, and Education, 10*(1), 10–33. https://doi.org/10.1111/mbe.12100

Balboni, G., Naglieri, J. A., & Cubelli, R. (2010). Concurrent and predictive validity of the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test. *Journal of Psychoeducational Assessment, 28*(3), 222–235. https://doi.org/10.1177/0734282909343763

Baltes, P. B. (1987). Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology, 23*(5), 611–626. https://doi.org/10.1037/0012-1649.23.5.611

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology, 50*, 471–507. https://doi.org/10.1146/annurev.psych.50.1.471

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction, 24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*(3), 361–370. https://doi.org/10.1177/00131640021970592

Bauer, C. M., Cabral, H. J., & Killiany, R. J. (2018). Multimodal discrimination between normal aging, mild cognitive impairment and Alzheimer's disease and prediction of cognitive decline. *Diagnostics, 8*(1), 14. https://doi.org/10.3390/diagnostics8010014

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., & Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *The Clinical Neuropsychologist*, *26*(2), 177–196. https://doi.org/10.1080/13854046.2012.663001

Bayliss, D. M., Jarrold, C., Baddeley, A. D., & Leigh, E. (2005). Differential constraints on the working memory and reading abilities of individuals with learning difficulties and typically developing children. *Journal of Experimental Child Psychology, 92*(1), 76–99. https://doi.org/10.1016/j.jecp.2005.04.002

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain, 123*(11), 2189–2202. https://doi.org/10.1093/brain/123.11.2189

Bechger, T. M., Maris, G., Verstralen, H. H., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27*(5), 319–334. https://doi.org/10.1177/0146621603257518

Belacchi, C., Carretti, B., & Lanfranchi, S. (2012). Raven Coloured Matrices: An analysis of the performance of Down syndrome and typical development individuals. *Psicologia Clinica dello Sviluppo, 16* (1), 221–229.

Belacchi, C., Scalisi, T. G., Cannoni, E., & Cornoldi, C. (2008). *CPM: Coloured Progressive Matrices. Standardizzazione Italiana.* Giunti O. S. Organizzazioni Speciali.

Belelli, F., Muzio, C., & Belacchi, C. (2010, October 14–16). *Analisi della tipologia degli errori nelle Matrici Progressive Colore in popolazioni cliniche (DSL-DSA-FIL-DI) in età evolutiva* [Conference presentation]. XIX Congresso Nazionale Airipa – I disturbi dell'apprendimento, Ivrea, Italy.

Bender, L. (1938). *A visual motor gestalt test and its clinical use* (Research Monographs No. 3). American Orthopsychiatric Association.

Bender, H. A., Martín García, A., & Barr, W. B. (2010). An interdisciplinary approach to neuropsychological test construction: perspectives from translation studies. *Journal of the International Neuropsychological Society : JINS*, *16*(2), 227–232. https://doi.org/10.1017/S1355617709991378

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benton, A. L., & Hamsher, K. (1989). *Multilingual Aphasia Examination*. AJA Associates.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *The Journal of General Psychology, 39*(1), 15–22. https://doi.org/10.1080/00221309.1948.9918159

Berg, W. K., & Byrd, D. L. (2002). The Tower of London spatial problem-solving task: Enhancing clinical and research implementation. *Journal of Clinical and Experimental Neuropsychology, 24*(5), 586–604. https://doi.org/10.1076/jcen.24.5.586.1006

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development, 81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Biancardi, A., Nicoletti, C., & Bachmann, C. (2016). *BDE-2: Batteria per la Discalculia Evolutiva: Test per la diagnosi dei disturbi dell'elaborazione numerica e del calcolo in età evolutiva (8–13 anni).* Edizioni Centro Studi Erickson.

Biederman, J., Petty, C. R., Doyle, A. E., Spencer, T., Henderson, C. S., Marion, B., Fried, R., & Faraone, S. V. (2007). Stability of executive function deficits in girls with ADHD: A prospective longitudinal followup study into adolescence. *Developmental Neuropsychology, 33*(1), 44–61. https://doi.org/10.1080/87565640701729755

Bilder, R. M. (2011). Neuropsychology 3.0: Evidence-based science and practice. *Journal of the International Neuropsychological Society*, *17*(1), 7–13. https://doi.org/10.1017/S1355617710001396

Bilder, R. M., Postal, K. S., Barisa, M., Aase, D. M., Cullum, C. M., Gillaspy, S. R., Harder, L. H., Kanter, G., Lanca, M., Lechuga, D. M., Morgan, J. E., Most, R. B., Puente, A. E., Salinas, C. M., Woodhouse, J., & Members of the IOPC Tele-Neuropsychology Task Force. (2020). Interorganizational practice committee recommendations/guidance for teleneuropsychology (TeleNP) in response to the COVID-19 pandemic. *The Clinical Neuropsychologist, 34*(7–8), 1314–1334. https://doi.org/10.1080/13854046.2020.1767214

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's Standard Progressive Matrices test. *Assessment, 19*(3), 354–369. https://doi.org/10.1177/1073191112446655

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences, 29*(2), 109–125. https://doi.org/10.1017/S0140525X06009034

Blair, C., & Ursache, A. (2011). A bidirectional model of executive functions and self-regulation. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (2nd ed., pp. 300–320). Guilford Press.

Blakemore, S. J., & Choudhury, S. (2006). Development of the adolescent brain: Implications for executive function and social cognition. *Journal of Child Psychology and Psychiatry, 47*(3–4), 296–312. https://doi.org/10.1111/j.1469-7610.2006.01611.x

Blažica, B., & Lewis, J. R. (2015). A Slovene translation of the system usability scale: The SUS-SI. *International Journal of Human–Computer Interaction, 31*(2), 112–117. https://doi.org/10.1080/10447318.2014.986634

Bolster, B., Marshall, W., Bow, J., Chalmers, N., & Stubel, M. (1986). Visual selective attention and impulsivity in learning-disabled children. Developmental Neuropsychology, 2(1), 25–40. https://doi.org/10.1080/87565648609540325

Borkowska, A., & Jach, K. (2017). Pre-testing of Polish translation of system usability scale (SUS). In *Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology – ISAT 2016 – Part I* (pp. 143–153). Springer International Publishing. https://doi.org/10.1007/978-3-319-46583-8_12

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*(3), 382–398. https://doi.org/10.1177/0013164498058003002

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the system usability scale: A test of alternative measurement models. *Cognitive Processing, 10*(3), 193–197. https://doi.org/10.1007/s10339-009-0268-9

Bottesi, G., Spoto, A., Freeston, M. H., Sanavio, E., & Vidotto, G. (2015). Beyond the score: Clinical evaluation through formal psychological assessment. *Journal of Personality Assessment, 97*(3), 252–260. https://doi.org/10.1080/00223891.2014.958846

Bradley-Johnson, S. (1998). Test reviews. *Psychology in the Schools, 35*(4), 409–415.

Brancaccio, A., de Chiusole, D., & Stefanutti, L. (2023). Algorithms for the adaptive assessment of procedural knowledge and skills. *Behavior Research Methods, 55*(7), 3929–3951. https://doi.org/10.3758/s13428-022-01998-y

Brancaccio, A., de Chiusole, D., & Wickelmaier, F. (2024). Software. In *Knowledge structures: Recent developments in theory and application* (Vol. 7). World Scientific.

Brancaccio, A., Epifania, O. M., & de Chiusole, D. (2023). *matRiks: Generates Raven-like matrices according to rules* (R package version 0.1.1). https://CRAN.R-project.org/package=matRiks

Brancaccio, A., de Chiusole, D., Epifania, O. M., Anselmi, P., Spinoso, M., Mazzoni, N., Bacherini, A., Orsoni, M., Giovagnoli, S., Pierluigi, I., Benassi, M., Balboni, G., & Stefanutti, L. (2025, in). Two Markov solution process models for the assessment of planning in problem-solving. *Psychometrika, 1-31*.https://doi.org/10.1017/psy.2025.10042

Brearly, T. W., Shura, R. D., Martindale, S. L., Lazowski, R. A., Luxton, D. D., Shenal, B. V., & Rowland, J. A. (2017). Neuropsychological test administration by videoconference: A systematic review and meta-analysis. *Neuropsychology Review*, *27*(2), 174–186. https://doi.org/10.1007/s11065-017-9349-1

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Brosnan, M., Demetre, J., Hamill, S., Robson, K., Shepherd, H., & Cody, G. (2002). Executive functioning in adults and children with developmental dyslexia. *Neuropsychologia, 40*(12), 2144–2155. https://doi.org/10.1016/S0028-3932(02)00046-5

Brydges, C. R., Fox, A. M., Reid, C. L., & Anderson, M. (2012). A unitary executive function predicts intelligence in children. *Intelligence, 40*(5), 458–469. https://doi.org/10.1016/j.intell.2012.05.006

Burgess, P. W., & Shallice, T. (1996a). Confabulation and the control of recollection. *Memory*, *4*(4), 359–411. https://doi.org/10.1080/096582196388906

Burgess, P. W., & Shallice, T. (1996b). Response suppression, initiation and strategy use following frontal lobe lesions. *Neuropsychologia, 34*(4), 263–273. https://doi.org/10.1016/0028-3932(95)00104-2

Burgess, P. W., & Simons, J. S. (2005). Theories of frontal lobe executive function: Clinical applications. In P. W. Halligan & D. T. Wade (Eds.), *Effectiveness of rehabilitation for cognitive deficits* (pp. 211–231). Oxford University Press.

Burke, H. R. (1958). Raven's Progressive Matrices: A review and critical evaluation. *The Journal of Genetic Psychology*, *93*(2), 199–228. https://doi.org/10.1080/00221325.1958.10532420

Busch, R. M., McBride, A., Curtiss, G., & Vanderploeg, R. D. (2005). The components of executive functioning in traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 27*(8), 1022–1032. https://doi.org/10.1080/13803390490919263

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology*, *20*(4), 419–426. https://doi.org/10.1016/j.acn.2005.02.002

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*(Número 4), 872–882.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge. https://doi.org/10.4324/9780203807644

Caffarra, P., Vezzadini, G., Zonato, F., Copelli, S., & Venneri, A. (2003). A normative study of a shorter version of Raven's Progressive Matrices 1938. *Neurological Sciences*, *24*(5), 336–339. https://doi.org/10.1007/s10072-003-0185-0

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *The Journal of Special Education*, *41*(4), 254–270. https://doi.org/10.1177/0022466907310366

Calvert, E., & Waterfall, R. (1982). A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. *International Journal of Man-Machine Studies*, *17*(3), 305–310. https://doi.org/10.1016/S0020-7373(82)80032-1

Cambridge Cognition Ltd. (2012). *CANTAB Eclipse test administration guide*. Cambridge Cognition Ltd.

Canini, M., Battista, P., Della Rosa, P. A., Catricalà, E., Salvatore, C., Gilardi, M. C., & Castiglioni, I. (2014). Computerized neuropsychological assessment in aging: Testing efficacy and clinical ecology of different interfaces. *Computational and Mathematical Methods in Medicine*, *2014*, 804723. https://doi.org/10.1155/2014/804723

Carlin, D., Bonerba, J., Phipps, M., Alexander, G., Shapiro, M., & Grafman, J. (2000). Planning impairments in frontal lobe dementia and frontal lobe lesion patients. *Neuropsychologia, 38*(5), 655–665. https://doi.org/10.1016/S0028-3932(99)00102-5

Carone, D. A. (Ed.). (2024). *A compendium of neuropsychological tests: Fundamentals of neuropsychological assessment and test reviews for clinical practice* (E. M. S. Sherman, J. E. Tan, & M. Hrabok, Eds.). Oxford University Press.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological assessment: Past and future. *Journal of the International Neuropsychological Society, 23*(9–10), 778–790. https://doi.org/10.1017/S1355617717001060

Castellanos, F. X., & Proal, E. (2012). Large-scale brain systems in ADHD: Beyond the prefrontal–striatal model. *Trends in Cognitive Sciences, 16*(1), 17–26. https://doi.org/10.1016/j.tics.2011.11.007

Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology, 31*(3), 161–179. https://doi.org/10.1037/h0059043

Cattell, R. B. (1949). *Culture free intelligence test, scale 1, handbook*. Institute of Personality and Ability Testing.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22. https://doi.org/10.1037/h0046743

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Houghton Mifflin.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Elsevier.

Cattell, R., & Cattell, A. K. S. (1981). *Manual for the culture fair intelligence test (Scales 2 and 3)*. Institute for Personality and Ability Testing.

Cattell, R. B., & Cattell, A. K. S. (1987). *Measuring intelligence with the Culture Fair Tests*. Institute for Personality and Ability Testing.

Caviola, S., Colling, L. J., Mammarella, I. C., & Szűcs, D. (2020). Predictors of mathematics in primary school: Magnitude comparison, verbal and spatial working memory measures. *Developmental Science, 23*(6), Article e12957. https://doi.org/10.1111/desc.12957

Cazzato, V., Basso, D., Cutini, S., & Bisiacchi, P. (2010). Gender differences in visuospatial planning: An eye movements study. *Behavioural Brain Research, 206*(2), 177-183.https://doi.org/10.1016/j.bbr.2009.09.010

CC-ISS (Consensus Conference dell'Istituto Superiore di Sanità). (2010). *Disturbi specifici dell'apprendimento*. Sistema Nazionale Linee Guida.

CC-ISS (Consensus Conference dell'Istituto Superiore di Sanità). (2022). *Linea Guida sulla gestione dei Disturbi Specifici dell'Apprendimento*. Sistema Nazionale Linee Guida.

Cernich, A. N., Brennana, D. M., Barker, L. M., & Bleiberg, J. (2007). Sources of error in computerized neuropsychological assessment. *Archives of Clinical Neuropsychology*, *22*(Suppl. 1), S39–S48. https://doi.org/10.1016/j.acn.2006.10.004

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learning and Individual Differences, 22*(3), 390–396. https://doi.org/10.1016/j.lindif.2011.12.007

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1998). A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. In A. C. Roberts, T. W. Robbins, & L. Weiskrantz (Eds.), *The prefrontal cortex: Executive and cognitive functions* (pp. 195–220). Oxford University Press.

Collette, F., & Van der Linden, M. (2002). Brain imaging of the central executive component of working memory. *Neuroscience & Biobehavioral Reviews, 26*(2), 105–125. https://doi.org/10.1016/S0149-7634(01)00063-X

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLOS Biology, 10*(3), Article e1001293. https://doi.org/10.1371/journal.pbio.1001293

Colom, R., Karama, S., Jung, R. E., & Haier, R. J. (2010). Human intelligence and brain networks. *Dialogues in Clinical Neuroscience, 12*(4), 489–501. https://doi.org/10.31887/DCNS.2010.12.4/rcolom

Cornoldi, C. (2007). *Difficoltà e disturbi dell'apprendimento*. Il Mulino.

Cornoldi, C., Antonucci, A. M., Bertolo, L., Brembati, F., Frinco, M., Giofrè, D., Giorgetti, G., Miliozzi, M., Pezzuti, L., Ramanzini, E., Sironi, E., Stoppa, E., Vio, C., & Toffalini, E. (2019a). Sintesi dei risultati principali ottenuti con la banca dati AIRIPA di più di 1.800 casi di DSA valutati con la WISC-IV. *Dislessia, Giornale Italiano di Ricerca Clinica e Applicativa, 16*(3), 249–263. https://doi.org/doi:10.14605/DIS1631901

Cornoldi, C., Pra Baldi, A., Giofrè, D., Albano, D., Friso, G., & Morelli, E. (2017). *Prove MT Avanzate-3 Clinica: La valutazione delle abilità di lettura, comprensione, scrittura e matematica per il biennio della scuola secondaria di II grado.* Giunti Edu.

Cornoldi, C., & Candela, M. (2015). *Prove di lettura e scrittura MT-16-19: Batteria per la verifica degli apprendimenti e la diagnosi di dislessia e disortografia – classi terza, quarta, quinta della scuola secondaria di secondo grado.* Edizioni Centro Studi Erickson.

Cornoldi, C., & Carretti, B. (2016). *Prove MT-3 Clinica: La valutazione delle abilità di lettura e comprensione per la scuola primaria e secondaria di I grado.* Giunti Edu.

Cornoldi, C., Di Caprio, R., De Francesco, G., & Toffalini, E. (2019b). The discrepancy between verbal and visuoperceptual IQ in children with a specific learning disorder: An analysis of 1624 cases. *Research in Developmental Disabilities*, *87*, 64–72. https://doi.org/10.1016/j.ridd.2019.02.002

Cornoldi, C., Ferrara, R., & Re, A. M. (2022). *BVSCO-3: Batteria per la valutazione clinica della scrittura e della competenza ortografica*. Giunti Psychometrics.

Cornoldi, C., Giofrè, D., Orsini, A., & Pezzuti, L. (2014). Differences in the intellectual profile of children with intellectual vs. Learning disability. *Research in Developmental Disabilities*, *35*(9), 2224–2230. https://doi.org/10.1016/j.ridd.2014.05.013

Cornoldi, C., & Tressoldi, P. E. (2014). I disturbi specifici di apprendimento. In C. Cornoldi (Ed.), *I disturbi dell'apprendimento* (pp. 17–44). Il Mulino.

Cortés Pascual, A., Moyano Muñoz, N., & Quílez Robres, A. (2019). The relationship between executive functions and academic performance in primary education: Review and meta-analysis. *Frontiers in Psychology*, *10*, Article 1582. https://doi.org/10.3389/fpsyg.2019.01582

Costa, A., Bagoj, E., Monaco, M., Zabberoni, S., De Rosa, S., Papantonio, A. M., Mundi, C., Caltagirone, C., & Carlesimo, G. A. (2014). Standardization and normative data obtained in the Italian population for a new verbal fluency instrument, the phonemic/semantic alternate fluency test. *Neurological Sciences*, *35*(3), 365–372. https://doi.org/10.1007/s10072-013-1520-8

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(7), 1–9. https://doi.org/10.7275/jyj1-4868

Cowey, A., & Green, S. (1996). The hippocampus: A "working memory" structure? The effect of hippocampal sclerosis on working memory. *Memory, 4*(1), 19–30. https://doi.org/10.1080/741940668

Craig, F., Margari, F., Legrottaglie, A. R., Palumbi, R., De Giambattista, C., & Margari, L. (2016). A review of executive function deficits in autism spectrum disorder and attention-deficit/hyperactivity disorder. *Neuropsychiatric Disease and Treatment, 12*, 1191–1202. https://doi.org/10.2147/NDT.S104620

Crawford, J. R., Bryan, J., Luszcz, M. A., Obonsawin, M. C., & Stewart, L. (2000). The executive decline hypothesis of cognitive aging: Do executive deficits qualify as differential deficits and do they mediate age-related memory decline? *Aging, Neuropsychology, and Cognition, 7*(1), 9–31. https://doi.org/10.1076/anec.7.1.9.806

Crisci, G., Caviola, S., Cardillo, R., & Mammarella, I. C. (2021). Executive functions in neurodevelopmental disorders: Comorbidity overlaps between attention deficit and hyperactivity disorder and specific learning disorders. *Frontiers in Human Neuroscience, 15*, Article 594234. https://doi.org/10.3389/fnhum.2021.594234

Cristofori, I., Cohen-Zimerman, S., & Grafman, J. (2019). Executive functions. *Handbook of Clinical Neurology, 163*, 197–219. https://doi.org/10.1016/B978-0-12-804281-6.00011-2

Crone, E. A., & Steinbeis, N. (2017). Neural perspectives on cognitive control development during childhood and adolescence. *Trends in Cognitive Sciences, 21*(3), 205–215. https://doi.org/10.1016/j.tics.2017.01.003

Crosbie, J., Arnold, P., Paterson, A., Swanson, J., Dupuis, A., Li, X., Shan, J., Goodale, T., Tam, C., Strug, L. J., & Schachar, R. J. (2013). Response inhibition and ADHD traits: Correlates and heritability in a community sample. *Journal of Abnormal Child Psychology, 41*(3), 497–507. https://doi.org/10.1007/s10802-012-9693-9

Culbertson, W. C., & Zillmer, E. A. (1998). The Tower of London^DX: A standardized approach to assessing executive functioning in children. *Archives of Clinical Neuropsychology, 13*(3), 285–301. https://doi.org/10.1016/S0887-6177(97)00033-4

Das, J. P. (1989). Cognitive plans and performance: A theory of cognitive remediation. In W. L. Pickering (Ed.), *Intelligence and cognition: Contemporary frames of reference* (pp. 117–158). Erlbaum.

Davidson, M., Keating, J. L., & Eyres, S. (2004). A low back-specific version of the SF-36 physical functioning scale. *Spine, 29*(5), 586–594. 10.1097/01.brs.0000103346.38557.73

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982–1003. https://doi.org/10.1287/mnsc.35.8.982

de Chiusole, D., Anselmi, P., Stefanutti, L., & Robusto, E. (2013). The gain–loss model: Bias and variance of the parameter estimates. *Electronic Notes in Discrete Mathematics, 42*, 33–40. https://doi.org/10.1016/j.endm.2013.05.143

de Chiusole, D., & Stefanutti, L. (2013). Modeling skill dependence in probabilistic competence structures. *Electronic Notes in Discrete Mathematics, 42*, 41–48. https://doi.org/10.1016/j.endm.2013.05.144

de Chiusole, D., Stefanutti, L., Anselmi, P., & Robusto, E. (2013). Assessing parameter invariance in the BLIM: Bipartition models. *Psychometrika, 78*(4), 710–724. https://doi.org/10.1007/s11336-013-9325-5

de Chiusole, D., Stefanutti, L., Anselmi, P., & Robusto, E. (2015). Modeling missing data in knowledge space theory. *Psychological Methods, 20*(4), 506–522. 10.1037/met0000050

de Chiusole, D., Stefanutti, L., Anselmi, P., & Robusto, E. (2020). Stat-Knowlab. Assessment and learning of statistics with competence-based knowledge space theory. *International Journal of Artificial Intelligence in Education*, *30*(4), 668–700. https://doi.org/10.1007/s40593-020-00223-1

de Chiusole, D., Stefanutti, L., & Spoto, A. (2017). A class of k-modes algorithms for extracting knowledge structures from data. *Behavior Research Methods*, *49*(4), 1212–1226. https://doi.org/10.3758/s13428-016-0780-7

de Chiusole, D., Spoto, A., & Stefanutti, L. (2020). Extracting partially ordered clusters from ordinal polytomous data. *Behavior Research Methods*, *52*, 503–520. https://doi.org/10.3758/s13428-019-01248-8

de Chiusole, D., Spinoso, M., Anselmi, P., Bacherini, A., Balboni, G., Mazzoni, N., Brancaccio, A., Epifania, O. M., Orsoni, M., Giovagnoli, S., Garofalo, S., Benassi, M., Robusto, E., Stefanutti, L., & Pierluigi, I. (2024). PsycAssist: A web-based artificial intelligence system designed for adaptive neuropsychological assessment and training. *Brain Sciences*, *14*(2), 122. https://doi.org/10.3390/brainsci14020122

De Clercq-Quaegebeur, M., Casalis, S., Lemaitre, M.-P., Bourgois, B., Getto, M., & Vallée, L. (2010). Neuropsychological profile on the WISC-IV of French children with dyslexia. *Journal of Learning Disabilities*, *43*(6), 563–574. https://doi.org/10.1177/0022219410375000

De Luca, C. R., Wood, S. J., Anderson, V., Buchanan, J. A., Proffitt, T. M., Mahony, K., & Pantelis, C. (2003). Normative data from the CANTAB. I: Development of executive function over the lifespan. *Journal of Clinical and Experimental Neuropsychology*, *25*(2), 242–254. https://doi.org/10.1076/jcen.25.2.242.13639

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Deguchi, K., Kono, S., Deguchi, S., Morimoto, N., Kurata, T., Ikeda, Y., & Abe, K. (2013). A novel useful tool of computerized touch panel–type screening test for evaluating cognitive function of chronic ischemic stroke

patients. *Journal of Stroke and Cerebrovascular Diseases, 22*(7), e197–e206. https://doi.org/10.1016/j.jstrokecerebrovasdis.2012.11.011

Degreef, E., Doignon, J. P., Ducamp, A., & Falmagne, J. C. (1986). Languages for the assessment of knowledge. *Journal of Mathematical Psychology, 30*(3), 243–256. https://doi.org/10.1016/0022-2496(86)90032-5

Dehn, M. J. (2017). How working memory enables fluid reasoning. *Applied Neuropsychology: Child, 6*(3), 245–247. https://doi.org/10.1080/21622965.2017.1317490

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System*. Pearson. https://doi.org/10.1037/t15082-000

Della Sala, S., MacPherson, S. E., Phillips, L. H., Sacco, L., & Spinnler, H. (2003). How many camels are there in Italy? Cognitive estimates standardised on the Italian population. *Neurological Sciences, 24*(1), 10–15. https://doi.org/10.1007/s100720300015

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

DeShon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence, 21*(2), 135–155. https://doi.org/10.1016/0160-2896(95)90023-3

D'Esposito, M., & Grossman, M. (1996). The physiological basis of executive function and working memory. *The Neuroscientist, 2*(6), 345–352. https://doi.org/10.1177/107385849600200612

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39*(2), 155–164. https://doi.org/10.1111/j.1547-5069.2007.00161.x

De Weerdt, F., Desoete, A., & Roeyers, H. (2013). Behavioral inhibition in children with learning disabilities. *Research in Developmental Disabilities, 34*(6), 1998–2007. https://doi.org/10.1016/j.ridd.2013.02.020

Diamond, A. (2000). Close interrelation of motor development and cognitive development and of the cerebellum and prefrontal cortex. *Child Development, 71*(1), 44–56. https://doi.org/10.1111/1467-8624.00117

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science, 333*(6045), 959–964. https://doi.org/10.1126/science.1204529

Dianat, I., Ghanbari, Z., & AsghariJafarabadi, M. (2014). Psychometric properties of the Persian language version of the system usability scale. *Health Promotion Perspectives, 4*(1), 82–87. https://doi.org/10.5681/hpp.2014.011

Dizon, G. (2016). Measuring Japanese EFL student perceptions of Internet-based tests with the technology acceptance model. *TESL-EJ, 20*(2), n2.

Díaz, A., Martín Lobo, P., Vergara Moragues, E., García, E., Jiménez Rodríguez, J. E., & Hernández, S. (2012). Torre de Hanoi: Datos normativos y desarrollo evolutivo de la planificación. *European Journal of Education and Psychology, 5*(1), 79–91. https://doi.org/10.30552/ejep.v5i1.81

Doignon, J. P. (1994). Knowledge spaces and skill assignments. In *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 111-121). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-4308-3_8

Doignon, J. P., & Falmagne, J. C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies, 23*(2), 175–196. https://doi.org/10.1016/S0020-7373(85)80031-6

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Springer. https://doi.org/10.1007/978-3-642-58625-5

Donadello, I., Spoto, A., Sambo, F., Badaloni, S., Granziol, U., & Vidotto, G. (2017). ATS-PD: An adaptive testing system for psychological disorders. *Educational and Psychological Measurement, 77*(5), 792–815. https://doi.org/10.1177/0013164416652188

Donfrancesco, R., Mugnaini, D., & Dell'Uomo, A. (2005). Cognitive impulsivity in specific learning disabilities. E*uropean Child & Adolescent Psychiatry, 14*(5), 270-275. https://link.springer.com/article/10.1007/s00787-005-0472-9

Dowling, C. E. (1993). Applying the basis of a knowledge space for controlling the questioning of an expert. *Journal of Mathematical Psychology, 37*(1), 21–48. https://doi.org/10.1006/jmps.1993.1002

Dowling, C. E., & Hockemeyer, C. (2001). Automata for the assessment of knowledge. *IEEE Transactions on Knowledge and Data Engineering, 13*(3), 451–461. 10.1109/69.929902

Droder, S., Mano, Q., Guerin, J., Becker, S., Epstein, J., & Tamm, L. (2024). The shifting role of fluid reasoning in reading among children evaluated for ADHD. *Applied Neuropsychology. Child, 13*(4), 325–333. https://doi.org/10.1080/21622965.2023.2178922

Druin, A. (2002). The role of children in the design of new technology. *Behaviour and Information Technology, 21*(1), 1–25. https://doi.org/10.1080/01449290110108659

Dubois, B., Slachevsky, A., Litvan, I., & Pillon, B. (2000). The FAB: A Frontal Assessment Battery at bedside. *Neurology, 55*(11), 1621–1626. https://doi.org/10.1212/WNL.55.11.1621

Dumontheil, I. (2014). Development of abstract thinking during childhood and adolescence: The role of rostrolateral prefrontal cortex. *Developmental Cognitive Neuroscience, 10*, 57–76. https://doi.org/10.1016/j.dcn.2014.07.009

Duncan, J. E. (1985). *The heuristics utilized by fifth grade students in solving verbal mathematics problems in a small group setting* [Doctoral dissertation, State University of New York at Albany]. ProQuest One Academic.

https://www.proquest.com/dissertations-theses/heuristics-utilized-fifth-grade-students-solving/docview/303388607/se-2

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences, 23*(10), 475–483. https://doi.org/10.1016/S0166-2236(00)01633-7

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Dünsch, I., & Gediga, G. (1995). Skills and knowledge structures. *British Journal of Mathematical and Statistical Psychology, 48*, 9–27. https://doi.org/10.1111/j.2044-8317.1995.tb01047.x

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1–26. https://doi.org/10.1214/aos/1176344552

Eid, M. E., & Diener, E. E. (2006). *Handbook of multimethod measurement in psychology*. American Psychological Association. https://doi.org/10.1037/11383-000

Eluwa, O. I., Eluwa, A. N., & Abang, B. K. (2011). Evaluation of mathematics achievement test: A comparison between classical test theory (CTT) and item response theory (IRT). *Journal of Educational and Social Research, 1*(4), 99–106.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Espy, K. A. (2004). Using developmental, cognitive, and neuroscience approaches to understand executive control in young children. *Developmental Neuropsychology, 26*(1), 379–384. https://doi.org/10.1207/s15326942dn2601_1

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., Frans, Ö., Gremigni, P., Halama, P., Iliescu, D., Laak, J. t., Matesic, K., Nystad, W., Roe, R. A., & Urbánek, T. (2012). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist*, *17*(4), 300–319. https://doi.org/10.1027/1016-9040/a000102

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199734177.001.0001

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Falmagne, J.-C., Albert, D., Doble, C., Eppstein, D., & Hu, X. (2013). *Knowledge spaces: Applications in education*. Springer Science & Business Media. https://doi.org/10.1007/978-3-642-35329-1

Falmagne, J.-C., & Doignon, J.-P. (1988a). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, *41*(1), 1–23. https://doi.org/10.1111/j.2044-8317.1988.tb00884.x

Falmagne, J.-C., & Doignon, J.-P. (1988b). A Markovian procedure for assessing the state of a system. *Journal of Mathematical Psychology*, *32*(3), 232–258. https://doi.org/10.1016/0022-2496(88)90011-9

Falmagne, J. C., & Doignon, J. P. (2010). *Learning spaces: Interdisciplinary applied mathematics*. Springer Science & Business Media. https://doi.org/10.1007/978-3-642-01039-2

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P., & Johanessen, L. (1990). Introduction to knowledge spaces: How to build, test and search them. *Psychological Review*, *97*(2), 204–224. https://doi.org/10.1037/0033-295X.97.2.201

Fancello, G. S. (2021). *Torre di Londra – ToL*. Erickson.

Fancello, G. S., Vio, C., & Cianchetti, C. (2006). *TOL. Torre di Londra. Test di valutazione delle funzioni esecutive (pianificazione e problem solving). Con CD-ROM*. Edizioni Erickson.

Ferrer, E., O'Hare, E. D., & Bunge, S. A. (2009). Fluid reasoning and the developing brain. *Frontiers in Neuroscience*, *3*(1), 46–51. https://doi.org/10.3389/neuro.01.003.2009

Filippi, R., Ceccolini, A., Periche-Tomas, E., & Bright, P. (2020). Developmental trajectories of metacognitive processing and executive function from childhood to older age. *Quarterly Journal of Experimental Psychology*, *73*(11), 1757–1773. https://doi.org/10.1177/1747021820931096

Fluss, R., Reiser, B., Faraggi, D., & Rotnitzky, A. (2009). Estimation of the ROC curve under verification bias. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *51*(3), 475–490. https://doi.org/10.1002/bimj.200800128

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*(4), 237–241. https://doi.org/10.1111/j.1467-9280.1996.tb00366.x

Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology*, *54*(1), 1–34. https://doi.org/10.1016/S0301-0511(00)00051-X

Fuster, J. M. (1993). Frontal lobes. *Current Opinion in Neurobiology*, *3*(2), 160–165. https://doi.org/10.1016/0959-4388(93)90204-C

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. Basic Books.

Gediga, G., & Düntsch, I. (2002). Skill set analysis in knowledge structures. *British Journal of Mathematical and Statistical Psychology*, *55*, 361–384. https://doi.org/10.1348/000711002760554516

Georgiou, G. K., Li, J., & Das, J. P. (2017). Tower of London: What level of planning does it measure? *Psychological Studies, 62,* 261–267. https://doi.org/10.1007/s12646-017-0416-8

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare* (pp. 295–336). Academic Press. https://doi.org/10.1016/B978-0-12-818438-7.00012-5

Germine, L., Reinecke, K., & Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist, 33*(2), 271–286. https://doi.org/10.1080/13854046.2018.1535662

Geurts, H. M., van den Bergh, S. F., & Ruzzano, L. (2014). Prepotent response inhibition and interference control in autism spectrum disorders: Two meta-analyses. *Autism Research, 7*(4), 407–420. https://doi.org/10.1002/aur.1369

Gialluisi, A., Andlauer, T. F. M., Mirza-Schreiber, N., Moll, K., Becker, J., Hoffmann, P., Ludwig, K. U., Czamara, D., St Pourcain, B., Honbolygó, F., Tóth, D., Csépe, V., Huguet, G., Morris, A. P., Hulslander, J., Willcutt, E. G., DeFries, J. C., Olson, R. K., Smith, S. D., . . . Schulte-Körne, G. (2021). Genome-wide association study reveals new insights into the heritability and genetic correlates of developmental dyslexia. *Molecular Psychiatry*, *26*(6), 3004–3017. https://doi.org/10.1038/s41380-020-00898-x

Gilbert, S. J., & Burgess, P. W. (2008). Executive function. *Current Biology, 18*(3), R110–R114. https://doi.org/10.1016/j.cub.2007.12.014

Gilberstadt, H., Lushene, R., & Buegel, B. (1976). Automated assessment of intelligence: The TAPAC test battery and computerized report writing. *Perceptual and Motor Skills, 43*(2), 627–635. https://doi.org/10.2466/pms.1976.43.2.627

Giofrè, D., & Cornoldi, C. (2015). The structure of intelligence in children with specific learning disabilities is different as compared to typically developing children. *Intelligence*, *52*, 36–43. https://doi.org/10.1016/j.intell.2015.07.002

Giofrè, D., Mammarella, I. C., & Cornoldi, C. (2014). The relationship among geometry, working memory, and intelligence in children. *Journal of Experimental Child Psychology*, *123*, 112–128. https://doi.org/10.1016/j.jecp.2014.01.002

Giofrè, D., Stoppa, E., Ferioli, P., Pezzuti, L., & Cornoldi, C. (2016). Forward and backward digit span difficulties in children with specific learning disorder. *Journal of Clinical and Experimental Neuropsychology*, *38*(4), 478–486. https://doi.org/10.1080/13803395.2015.1125454

Goel, V., & Grafman, J. (1995). Are the frontal lobes implicated in "planning" functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia*, *33*(5), 623–642. https://doi.org/10.1016/0028-3932(95)90866-P

Goharpey, N., Crewther, D. P., & Crewther, S. G. (2013). Problem solving ability in children with intellectual disability as measured by the Raven's Colored Progressive Matrices. *Research in Developmental Disabilities*, *34*(12), 4366–4374. https://doi.org/10.1016/j.ridd.2013.09.013

Goldberg, E., & Podell, K. (2000). Adaptive decision making, ecological validity, and the frontal lobes. *Journal of Clinical and Experimental Neuropsychology*, *22*(1), 56–68. https://doi.org/10.1076/1380-3395(200002)22:1;1-8;FT056

Golden, C., Freshwater, S. M., & Golden, Z. (2002). *Stroop Color and Word Test*. APA PsycTests. https://doi.org/10.1037/t06065-000

González-de Paz, L., Kostov, B., López-Pina, J. A., Solans-Julián, P., Navarro-Rubio, M. D., & Sisó-Almirall, A. (2015). A Rasch analysis of patients' opinions of primary health care professionals' ethical behaviour with respect to communication issues. *Family Practice*, *32*(2), 237–243. https://doi.org/10.1093/fampra/cmu073

Goswami, U. (1992). *Analogical reasoning in children*. Lawrence Erlbaum Associates.

Gottfredson, L. S. (1997a). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8

Gottfredson, L. S. (1997b). Why g matters: The complexity of everyday life. *Intelligence, 24*(1), 79–132. https://doi.org/10.1016/S0160-2896(97)90014-3

Grafman, J. (1995). Similarities and distinctions among current models of prefrontal cortical functions. *Annals of the New York Academy of Sciences, 769*(1), 337–368. https://doi.org/10.1111/j.1749-6632.1995.tb38149.x

Grafman, J. (2002). The structured event complex and the human prefrontal cortex. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 292–310). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195134971.003.0019

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology, 38*(4), 404–411. https://doi.org/10.1037/h0059831

Green, C. T., Bunge, S. A., Briones Chiongbian, V., Barrow, M., & Ferrer, E. (2017). Fluid reasoning predicts future mathematical performance among children and adolescents. *Journal of Experimental Child Psychology*, *157*, 125–143. https://doi.org/10.1016/j.jecp.2016.12.005

Gronier, G., & Baudet, A. (2021). Psychometric evaluation of the F-SUS: Creation and validation of the French version of the system usability scale. *International Journal of Human–Computer Interaction, 37*(16), 1571–1582. https://doi.org/10.1080/10447318.2021.1898828

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Gunn, D. M., & Jarrold, C. (2004). Raven's matrices performance in Down syndrome: Evidence of unusual errors. *Research in Developmental Disabilities, 25*(5), 443–457. https://doi.org/10.1016/j.ridd.2003.07.004

Guthke, J., Beckmann, J. F., Stein, H., Vahle, H., & Rittner, S. (1995). *Adaptive computergestützte Intelligenz-Lerntestbatterie (ACIL)* [The adaptive, computer-based learning test battery].https://durham-repository.worktribe.com/output/1124715

Hackman, D. A., Gallop, R., Evans, G. W., & Farah, M. J. (2015). Socioeconomic status and executive function: Developmental trajectories and mediation. *Developmental Science, 18*(5), 686–702. https://doi.org/10.1111/desc.12246

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Pearson Education Limited.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Pearson Prentice Hall.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications* (Vol. 7). Springer Science & Business Media.

Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement, 66*(6), 1039–1046. https://doi.org/10.1177/0013164406288169

Hanks, R. A., Rapport, L. J., Millis, S. R., & Deshpande, S. A. (1999). Measures of executive functioning as predictors of functional ability and social integration in a rehabilitation sample. *Archives of Physical Medicine and Rehabilitation, 80*(9), 1030–1037. https://doi.org/10.1016/S0003-9993(99)90056-4

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hansen, J. A. (2019). Development and psychometric evaluation of the Hansen Research Services Matrix Adaptive Test: A measure of nonverbal IQ. *Journal of Autism and Developmental Disorders*, *49*(7), 2721–2732. https://doi.org/10.1007/s10803-016-2932-0

Hanscom, A. J. (2016). *Balanced and barefoot: How unrestricted outdoor play makes for strong, confident, and capable children*. New Harbinger Publications.

Happé, F. (2013). Fluid intelligence. In *Encyclopedia of autism spectrum disorders* (pp. 1310–1311). Springer. https://doi.org/10.1007/978-1-4419-1698-3_1686

Happé, F., Booth, R., Charlton, R., & Hughes, C. (2006). Executive function deficits in autism spectrum disorders and attention-deficit/hyperactivity disorder: Examining profiles across domains and ages. *Brain and Cognition, 61*(1), 25–39. https://doi.org/10.1016/j.bandc.2006.03.004

Harniss, M., Amtmann, D., Cook, D., & Johnson, K. (2007). Considerations for developing interfaces for collecting patient-reported outcomes that allow the inclusion of individuals with disabilities. *Medical Care, 45*(5), S48–S54. 10.1097/01.mlr.0000250822.41093.ca

Harris, C., Tang, Y., Birnbaum, E., Cherian, C., Mendhe, D., & Chen, M. H. (2024). Digital neuropsychology beyond computerized cognitive assessment: Applications of novel digital technologies. *Archives of Clinical Neuropsychology*, *39*(3), 290–304. https://doi.org/10.1093/arclin/acae016

Hartanto, A., & Yang, H. (2016). Is the smartphone a smart choice? The effect of smartphone separation on executive functions. *Computers in Human Behavior, 64*, 329–336. https://doi.org/10.1016/j.chb.2016.07.002

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193–225). Academic Press. https://doi.org/10.1016/S0079-7421(08)60041-9

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science, 3*(4), 275–310. https://doi.org/10.1207/s15516709cog0304_1

Haynes, S. N., Richard, D. C., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hebben, N., & Milberg, W. (2009). *Essentials of neuropsychological assessment*. John Wiley & Sons.

Helland, T., & Asbjørnsen, A. (2000). Executive functions in dyslexia. *Child Neuropsychology, 6*(1), 37–48. https://doi.org/10.1076/0929-7049(200003)6:1;1-B;FT037

Heller, J. (2017). Identifiability in probabilistic knowledge structures. *Journal of Mathematical Psychology, 77*, 46–57. https://doi.org/10.1016/j.jmp.2016.07.008

Heller, J. (2021). Generalizing quasi-ordinal knowledge spaces to polytomous items. *Journal of Mathematical Psychology, 101*, 102515. https://doi.org/10.1016/j.jmp.2021.102515

Heller, J., Anselmi, P., Stefanutti, L., & Robusto, E. (2017). A necessary and sufficient condition for unique skill assessment. *Journal of Mathematical Psychology, 79*, 23–28. https://doi.org/10.1016/j.jmp.2017.05.004

Heller, J., Augustin, T., Hockemeyer, C., Stefanutti, L., & Albert, D. (2013). Recent developments in competence-based knowledge space theory. In *Knowledge spaces* (pp. 243–286). Springer. https://doi.org/10.1007/978-3-642-35329-1_12

Heller, J., Stefanutti, L., Anselmi, P., & Robusto, E. (2015). On the link between cognitive diagnostic models and knowledge space theory. *Psychometrika, 80*(4), 995–1019. https://doi.org/10.1007/s11336-015-9457-x

Heller, J., Stefanutti, L., Anselmi, P., & Robusto, E. (2016). Erratum to: On the link between cognitive diagnostic models and knowledge space theory. *Psychometrika, 81*(1), 250–251. https://doi.org/10.1007/s11336-015-9494-5

Heller, J., Ünlü, A., & Albert, D. (2013). Skills, competencies and knowledge structures. In *Knowledge spaces: Applications in education* (pp. 229–242). Springer. https://doi.org/10.1007/978-3-642-35329-1_11

Heller, J., & Wickelmaier, F. (2013). Minimum discrepancy estimation in probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics, 42*, 49–56. https://doi.org/10.1016/j.endm.2013.05.145

Hill, E. L. (2004). Executive dysfunction in autism. *Trends in Cognitive Sciences, 8*(1), 26–32. https://doi.org/10.1016/j.tics.2003.11.003

Hirschfeld, G., & von Brachel, R. (2014). Multiple-group confirmatory factor analysis in R—A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation, 19*(7). https://doi.org/10.7275/qazy-2946

Hockemeyer, C. (2002). A comparison of non-deterministic procedures for the adaptive assessment of knowledge. *Psychological Test and Assessment Modeling, 44*(4), 495–510.

Hoffmann, M. (2013). The human frontal lobes and frontal network systems: An evolutionary, clinical, and treatment perspective. *ISRN Neurology, 2013,* Article 892459. https://doi.org/10.1155/2013/892459

Holden, R. J., & Karsh, B. T. (2010). The technology acceptance model: Its past and its future in health care. *Journal of Biomedical Informatics, 43*(1), 159–172. https://doi.org/10.1016/j.jbi.2009.07.002

Holmes, J., Gathercole, S. E., Place, M., Alloway, T. P., Elliott, J. G., & Hilton, K. A. (2010). The diagnostic utility of executive function assessments in the identification of ADHD in children. *Child and Adolescent Mental Health, 15*(1), 37–43. https://doi.org/10.1111/j.1475-3588.2009.00536.x

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253–270. https://doi.org/10.1037/h0023816

Horn, J. L. (2007). Spearman, g, expertise, and the nature of human cognitive capability. In *Extending intelligence* (pp. 173–208). Routledge.

Hornke, L. F. (1999). Computerized adaptive testing. In F. W. Rudiger & W. Guenter (Eds.), *Advances in psychological assessment* (pp. 115–140). Hogrefe & Huber.

Hornke, L. F., & Habon, T. (1986). *AIL—Adaptive Intelligenz Diagnostikum* [Adaptive intelligence test]. Beltz Test.

Horowitz-Kraus, T. (2015). Differential effect of cognitive training on executive functions and reading abilities in children with ADHD and in children with ADHD comorbid with reading difficulties. *Journal of Attention Disorders, 19*(6), 515–526. https://doi.org/10.1177/1087054713502079

Hourcade, J. P. (2008). Interaction design and children. *Foundations and Trends in Human–Computer Interaction, 1*(4), 277–392. https://doi.org/10.1561/1100000006

Howieson, D. (2019). Current limitations of neuropsychological tests and assessment procedures. *The Clinical Neuropsychologist, 33*(2), 200–208. https://doi.org/10.1080/13854046.2018.1552762

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hughes, C., Ensor, R., Wilson, A., & Graham, A. (2009). Tracking executive function across the transition to school: A latent variable approach. *Developmental Neuropsychology, 35*(1), 20–36. https://doi.org/10.1080/87565640903325691

Huizinga, M., Dolan, C. V., & Van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia, 44*(11), 2017–2036. https://doi.org/10.1016/j.neuropsychologia.2006.01.010

Humes, G. E., Welsh, M. C., Retzlaff, P., & Cookson, N. (1997). Towers of Hanoi and London: Reliability and validity of two executive function tasks. *Assessment, 4*(3), 249–257. https://doi.org/10.1177/107319119700400305

Huteau, M., & Lautrey, J. (2000). *Evaluer l'intelligence: Psychométrie cognitive*. Presses Universitaires de France.

Hvidt, J. C. S., Christensen, L. F., Sibbersen, C., Helweg-Jørgensen, S., Hansen, J. P., & Lichtenstein, M. B. (2020). Translation and validation of the System Usability Scale in a Danish mental health setting using digital technologies in treatment interventions. *International Journal of Human–Computer Interaction, 36*(8), 709–716. 10.1080/10447318.2019.1680922

Iaia, M., Vizzi, F., Carlino, M. D., Marinelli, C. V., Angelelli, P., & Turi, M. (2025). WAIS-IV Cognitive Profiles in Italian University Students with Dyslexia. *Journal of Intelligence, 13*(8), 100. https://doi.org/10.3390/jintelligence13080100

IBM Corp. (2017). *IBM SPSS Statistics for Windows, version 25.0*. IBM Corp.

Injoque Ricle, I., Barreyro, J. P., Formoso, J., & Jaichenco, V. I. (2014). Tower of London: Planning assessment in children from 6 to 13 years of age. *Spanish Journal of Psychology, 17*, Article E43. https://doi.org/10.1017/sjp.2014.83

Irwin, L. N., Kofler, M. J., Soto, E. F., & Groves, N. B. (2019). Do children with attention-deficit/hyperactivity disorder (ADHD) have set switching deficits? *Neuropsychology, 33*(4), 470–481. https://doi.org/10.1037/neu0000546

JASP Team. (2024). *JASP* (Version 0.18.3.0) [Computer software]. https://jasp-stats.org/

Jeffries, R., & Desurvire, H. W. (1992). Usability testing vs. heuristic evaluation: Was there a contest? *SIGCHI Bulletin, 24*(4), 39–41. https://doi.org/10.1145/142167.142179

Jensen, A. R. (1974). *Abilities: Their structure, growth, and action*. Houghton Mifflin.

Johann, V., Könen, T., & Karbach, J. (2020). The unique contribution of working memory, inhibition, cognitive flexibility, and intelligence to reading comprehension and reading speed. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *26*(3), 324–344. https://doi.org/10.1080/09297049.2019.1649381

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Johnco, C., Wuthrich, V. M., & Rapee, R. M. (2014). The influence of cognitive flexibility on treatment outcome and cognitive restructuring skill acquisition during cognitive behavioural treatment for anxiety and depression in older adults: Results of a pilot study. *Behaviour Research and Therapy, 57*, 55–64. https://doi.org/10.1016/j.brat.2014.04.005

Jones, H. E., & Conrad, H. S. (1933). The growth and decline of intelligence: A study of a homogeneous group between the ages of ten and sixty. *Genetic Psychology Monographs, 13*(3), 223–298.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183–202. https://doi.org/10.1007/BF02289343

Jurado, M. B., & Rosselli, M. (2007). The elusive nature of executive functions: A review of our current understanding. *Neuropsychology Review, 17*(3), 213–233. https://doi.org/10.1007/s11065-007-9040-z

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401–415. https://doi.org/10.1007/BF02291817

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark iv. *Educational and Psychological Measurement, 34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kaller, C. P., Debelak, R., Köstering, L., Egle, J., Rahm, B., Wild, P. S., Blettner, M., Beutel, M. E., & Unterrainer, J. M. (2016). Assessing planning ability across the adult life span: Population-representative and age-adjusted reliability estimates for the Tower of London (TOL-F). *Archives of Clinical Neuropsychology, 31*(2), 148–164. https://doi.org/10.1093/arclin/acv088

Kaller, C. P., Rahm, B., Spreer, J., Mader, I., & Unterrainer, J. M. (2008). Thinking around the corner: The development of planning abilities. *Brain and Cognition, 67*(3), 360–370. https://doi.org/10.1016/j.bandc.2008.02.003

Kaller, C. P., Unterrainer, J. M., Rahm, B., & Halsband, U. (2004). The impact of problem structure on planning: Insights from the Tower of London task. *Cognitive Brain Research, 20*(3), 462–472. https://doi.org/10.1016/j.cogbrainres.2004.04.002

Kaller, C. P., Unterrainer, J. M., & Stahl, C. (2012). Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. *Psychological Assessment, 24*(1), 46–61. https://doi.org/10.1037/a0025174

Kessels, R. P. (2019). Improving precision in neuropsychological assessment: Bridging the gap between classic paper-and-pencil tests and paradigms from cognitive neuroscience. *The Clinical Neuropsychologist, 33*(2), 357–368. https://doi.org/10.1080/13854046.2018.1518489

Khosrorad, R., & Soltani Kohbani, S. (2015). The relationship between executive function and theory of mind in students with mathematics learning's disorder. *Journal of Sabzevar University of Medical Sciences, 21*(6), 1152–1162.

Klahr, D. (1978). Goal formation, planning, and learning by pre-school problem solvers or: "My socks are in the dryer". In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 181–212). Lawrence Erlbaum Associates.

Klenberg, L., Korkman, M., & Lahti-Nuuttila, P. (2001). Differential development of attention and executive functions in 3- to 12-year-old Finnish children. *Developmental Neuropsychology, 20*(1), 407–428. https://doi.org/10.1207/S15326942DN2001_6

Kline, P. (2014). *An easy guide to factor analysis*. Routledge. https://doi.org/10.4324/9781315788135

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kłosowska, J. (1979). Zarys metodologii badań psychologicznych. *Państwowe Wydawnictwo Naukowe*, 153–168.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences, 11*(6), 229–235. https://doi.org/10.1016/j.tics.2007.04.005

Koppen, M. (1993). Extracting human expertise for constructing knowledge spaces: An algorithm. *Journal of Mathematical Psychology, 37*(1), 1–20. https://doi.org/10.1006/jmps.1993.1001

Korkman, M. (1998). *NEPSY. A developmental neuropsychological assessment*. The Psychological Corporation.

Korkman, M., Kemp, S. L., & Kirk, U. (2001). Effects of age on neurocognitive measures of children ages 5 to 12: A cross-sectional study on 800 children from the United States. *Developmental Neuropsychology, 20*(1), 331–354. https://doi.org/10.1207/S15326942DN2001_2

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY-II: Clinical and interpretive manual.* The Psychological Corporation.

Korossy, K. (1997). Extending the theory of knowledge spaces: A competence-performance approach. *Zeitschrift für Psychologie, 205*, 53–82.

Korossy, K. (1999). Modeling knowledge as competence and performance. In D. Albert & J. Lukas (Eds.), *Knowledge spaces: Theories, empirical research, applications* (pp. 103–132). Lawrence Erlbaum Associates.

Kortum, P., & Bangor, A. (2013). Usability ratings for everyday products measured with the system usability scale. *International Journal of Human–Computer Interaction, 29*(2), 67–76. https://doi.org/10.1080/10447318.2012.681221

Kortum, P., Acemyan, C. Z., & Oswald, F. L. (2021). Is it time to go positive? Assessing the positively worded System Usability Scale (SUS). *Human Factors, 63*(6), 987–998. https://doi.org/10.1177/0018720819881556

Köstering, L., McKinlay, A., Stahl, C., & Kaller, C. P. (2012). Differential patterns of planning impairments in Parkinson's disease and sub-clinical signs of dementia? A latent-class model-based approach. *PLoS ONE, 7*(6), e38855. https://doi.org/10.1371/journal.pone.0038855

Kramer, A. W., & Huizenga, H. M. (2023). Raven's Standard Progressive Matrices for adolescents: A case for a shortened version. *Journal of Intelligence, 11*(4), 72. https://doi.org/10.3390/jintelligence11040072

Krikorian, R., Bartok, J., & Gay, N. (1994). Tower of London procedure: A standard method and developmental data. *Journal of Clinical and Experimental Neuropsychology, 16*(6), 840–850. https://doi.org/10.1080/01688639408402697

Kubinger, K. D., Formann, A. K., & Farkas, M. G. (1991). Psychometric shortcomings of Raven's Standard Progressive Matrices, in particular for computerized testing. *European Review of Applied Psychology/Revue Européenne de Psychologie Appliquée, 41*(4), 295–300.

Kunda, M., Soulières, I., Rozga, A., & Goel, A. K. (2016). Error patterns on the Raven's Standard Progressive Matrices Test. *Intelligence, 59*, 181–198. https://doi.org/10.1016/j.intell.2016.09.004

Kuniavsky, M. (2003). *Observing the user experience: A practitioner's guide to user research*. Elsevier.

Kyriazos, T. A. (2018). Applied psychometrics: The 3-faced construct validation method, a routine for evaluating a factor structure. *Psychology, 9*(8), 2044–2072. https://doi.org/10.4236/psych.2018.98117

La Porta, F., Franceschini, M., Caselli, S., Cavallini, P., Susassi, S., & Tennant, A. (2011). Unified balance scale: An activity-based, bed to community, and aetiology-independent measure of balance calibrated with Rasch analysis. *Journal of Rehabilitation Medicine, 43*(5), 435–444. https://doi.org/10.2340/16501977-0797

Lancioni, G. E., O'Reilly, M. F., Cuvo, A. J., Singh, N. N., Sigafoos, J., & Didden, R. (2007). PECS and VOCAs to enable students with developmental disabilities to make requests: An overview of the literature. *Research in Developmental Disabilities, 28*(5), 468–488. https://doi.org/10.1016/j.ridd.2006.06.003

Landerl, K., Bevan, A., & Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: A study of 8–9-year-old students. *Cognition, 93*(2), 99–125. https://doi.org/10.1016/j.cognition.2003.11.004

Lange, K. W., Tucha, O., Alders, G. L., Preier, M., Csoti, I., Merz, B., Mark, G., Herting, B., Fornadi, F., Reichmann, H., Vieregge, P., & Reiners, K. (2003). Differentiation of parkinsonian syndromes according to differences in

executive functions. *Journal of Neural Transmission, 110*(9), 983–995. https://doi.org/10.1007/s00702-003-0011-0

Langener, A. M., Kramer, A. W., van den Bos, W., & Huizenga, H. M. (2022). A shortened version of Raven's Standard Progressive Matrices for children and adolescents. *British Journal of Developmental Psychology, 40*(1), 35–45. https://doi.org/10.1111/bjdp.12381

Lanyon, R. I., & Goodstein, L. D. (1997). *Personality assessment* (Vol. 3). John Wiley & Sons.

Lazeron, R. H., Rombouts, S. A., Machielsen, W. C., Scheltens, P., Witter, M. P., Uylings, H. B., & Barkhof, F. (2000). Visualizing brain activation during planning: The tower of London test adapted for functional MR imaging. *AJNR. American Journal of Neuroradiology, 21*(8), 1407–1414.

Lee, G. J., Do, C., & Suhr, J. (2023). Noncredible presentations of symptoms and functional impairment in the assessment of adult attention-deficit/hyperactivity disorder. *Psychology & Neuroscience, 16*(3), 284.https://doi.org/10.1037/pne0000319

Lee, Y., Kozar, K. A., & Larsen, K. R. (2003). The technology acceptance model: Past, present, and future. *Communications of the Association for Information Systems, 12*(1), 50. https://doi.org/10.17705/1CAIS.01250

Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology, 21*(1), 59–80. https://doi.org/10.1348/026151003321164627

Leontjev, A. N. (1978). *Activity, consciousness, and personality*. Prentice-Hall.

Levine, B., Stuss, D. T., Milberg, W. P., Alexander, M. P., Schwartz, M., & Macdonald, R. (1998). The effects of focal and diffuse brain damage on strategy application: Evidence from focal lesions, traumatic brain injury and normal aging. *Journal of the International Neuropsychological Society, 4*(3), 247–264. https://doi.org/10.1017/S1355617798002471

Lewis, J. R. (1992). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society Annual Meeting*, *36*(16), 1259–1260. SAGE Publications. https://doi.org/10.1177/154193129203601617

Lewis, J. R. (2018). The System Usability Scale: Past, present, and future. *International Journal of Human-Computer Interaction*, *34*(7), 577–590. https://doi.org/10.1080/10447318.2018.1455307

Lewis, J. R. (2019). Comparison of four TAM item formats: Effect of response option labels and order. *Journal of Usability Studies*, *14*(4), 224–236.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In *Human Centered Design: First International Conference, HCD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19–24, 2009, Proceedings* (pp. 94–103). Springer. https://doi.org/10.1007/978-3-642-02806-9_12

Lewis, J. R., & Sauro, J. (2017). Can I leave this one out? The effect of dropping an item from the SUS. *Journal of Usability Studies*, *13*(1), 38–46.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099–2102). ACM. https://doi.org/10.1145/2470654.2481287

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505. https://doi.org/10.1080/10447318.2015.1064654

Lezak, M. D. (1982). The problem of assessing executive functions. *International Journal of Psychology*, *17*(1–4), 281–297. https://doi.org/10.1080/00207598208247445

Lezak, M. D. (1983). *Neuropsychological assessment* (2nd ed.). Oxford University Press.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). Oxford University Press.

Lezak, M. D. (2004). *Neuropsychological assessment*. Oxford University Press.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). Neuropsychological assessment (5th ed). NewYork: Oxford University Press.

Li, S. C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science, 15*(3), 155–163. https://doi.org/10.1111/j.0956-7976.2004.01503003.x

Lifshitz, H., Shtein, S., Weiss, I., & Vakil, E. (2021). Meta-analysis of explicit memory studies in individuals with intellectual disability. *European Journal of Special Needs Education, 36*(1), 57–74. https://doi.org/10.1080/08856257.2011.543535

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 1–55.

Lilley, M., Barker, T., & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education, 43*(1–2), 109–123. https://doi.org/10.1016/j.compedu.2003.12.008

Lima, R. F., Azoni, C. A. S., & Ciasca, S. M. (2011). Attentional performance and executive functions in children with learning difficulties. *Psicologia: Reflexão e Crítica, 24*(4), 685–691. https://doi.org/10.1590/S0102-79722011000400008

Lin, J. J., Mamykina, L., Lindtner, S., Delajoux, G., & Strub, H. B. (2006). Fish'n'Steps: Encouraging physical activity with an interactive computer game. In *International Conference on Ubiquitous Computing* (pp. 261–278). Springer. https://doi.org/10.1007/11853565_16

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*(2), 350.

Lindenberger, U., Scherer, H., & Baltes, P. B. (2001). The strong connection between sensory and cognitive performance in old age: Not due to sensory acuity reductions operating during cognitive assessment. *Psychology and Aging*, *16*(2), 196–205. https://doi.org/10.1037//0882-7974.16.2.196

Lindgren, K. A., Stachowiak, J., & Smith, D. B. (2003). The neuropsychology of traumatic brain injury. In M. D. Stiers (Ed.), *Neuropsychology handbook* (pp. 167–198). Springer.

Lipina, S. J., Martelli, M. I., Vuelta, B., Injoque-Ricle, I., & Colombo, J. A. (2004). Pobreza y desempeño ejecutivo en alumnos preescolares de la ciudad de Buenos Aires (República Argentina). *Interdisciplinaria*, *21*(2), 153–193.

Loring, D. W., Bauer, R. M., Cavanagh, L., Drane, D. L., Enriquez, K. D., Reise, S. P., … Bilder, R. M. (2022). Rationale and Design of the National Neuropsychology Network. *Journal of the International Neuropsychological Society*, *28*(1), 1–11. https://doi.org/10.1017/s1355617721000199

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Loumidis, K. S., & Shropshire, J. M. (1997). Effects of waiting time on appointment attendance with clinical psychologists and length of treatment. *Irish Journal of Psychological Medicine*, *14*(2), 49–54. https://doi.org/10.1017/S0790966700002986

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*(2), 73–79. https://doi.org/10.1027/1614-2241.4.2.73

Luciana, M., Collins, P. F., Olson, E. A., & Schissel, A. M. (2009). Tower of London performance in healthy adolescents: The development of planning skills and associations with self-reported inattention and impulsivity. *Developmental Neuropsychology*, *34*(4), 461–475. https://doi.org/10.1080/87565640902964540

Luciana, M., & Nelson, C. A. (1998). The functional emergence of prefrontally-guided working memory systems in 4–8-year-old children. *Neuropsychologia, 36*(3), 273–293. https://doi.org/10.1016/S0028-3932(97)00109-7

Luciana, M., & Nelson, C. A. (2002). Assessment of neuropsychological function in children using the Cambridge Neuropsychological Testing Automated Battery (CANTAB): Performance in 4–12 year-olds. *Developmental Neuropsychology, 22*(3), 595–623. https://doi.org/10.1207/S15326942DN2203_3

Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A., & Sweeney, J. A. (2004). Maturation of cognitive processes from late childhood to adulthood. *Child Development, 75*(5), 1357–1372. https://doi.org/10.1111/j.1467-8624.2004.00745.x

Lunt, L., Bramham, J., Morris, R. G., Bullock, P. R., Selway, R. P., Xenitidis, K., & David, A. S. (2012). Prefrontal cortex dysfunction and 'Jumping to Conclusions': Bias or deficit? *Journal of Neuropsychology, 6*(1), 65–78. https://doi.org/10.1111/j.1748-6653.2011.02005.x

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology*. Basic Books.

Luxton, D. D. (2016). An introduction to artificial intelligence in behavioral and mental health care. In *Artificial Intelligence in Behavioral and Mental Health Care* (pp. 1–26). Academic Press. https://doi.org/10.1016/B978-0-12-420248-1.00001-5

Luxton, D. D., Kayl, R. A., & Mishkind, M. C. (2014). mHealth data security: The need for HIPAA-compliant standardization. *Telemedicine and e-Health*, *18*(4), 284–288. https://doi.org/10.1089/tmj.2011.0180

Lyytinen, H., Erskine, J., Hämäläinen, J., Torppa, M., & Ronimus, M. (2015). Dyslexia—Early identification and prevention: Highlights from the Jyväskylä Longitudinal Study of Dyslexia. *Current Developmental Disorders Reports, 2*(4), 330–338. https://doi.org/10.1007/s40474-015-0067-1

Maheu, M. M., Drude, K. P., Hertlein, K. M., Lipschutz, R., Wall, K., Long, R., Hilty, D. M., Mockenhaupt, R. E., & Hawkins, A. (2018). An interdisciplinary framework for telebehavioral health competencies. *Journal of Technology in Behavioral Science*, *3*(2), 108–140. https://doi.org/10.1007/s41347-017-0038-y

Mahone, E. M., Hagelthorn, K. M., Cutting, L. E., Schuerholz, L. J., Pelletier, S. F., Rawlins, C., Singer, H. S., & Denckla, M. B. (2002). Effects of IQ on executive function measures in children with ADHD. *Child Neuropsychology, 8*(1), 52–65. https://doi.org/10.1076/chin.8.1.52.8719

Malloy-Diniz, L. F., Cardoso-Martins, C., Nassif, E. P., Levy, A. M., Leite, W. B., & Fuentes, D. (2008). Planning abilities of children aged 4 years and 9 months to 8½ years: Effects of age, fluid intelligence and school type on performance in the Tower of London test. *Dementia & Neuropsychologia*, *2*(1), 26–30. https://doi.org/10.1590/S1980-57642009DN20100006

Mammarella, I. C., Caviola, S., Giofrè, D., & Szűcs, D. (2018). The underlying structure of visuospatial working memory in children with mathematical learning disability. *British Journal of Developmental Psychology, 36*(2), 220–235. https://doi.org/10.1111/bjdp.12202

Mammarella, I. C., Giofrè, D., Ferrara, R., & Cornoldi, C. (2013). Intuitive geometry and visuospatial working memory in children showing symptoms of nonverbal learning disabilities. *Child Neuropsychology, 19*(3), 235–249. https://doi.org/10.1080/09297049.2011.640931

Mano, Q. R., Jastrowski Mano, K. E., Guerin, J. M., Gibler, R. C., Becker, S. P., Denton, C. A., Tamm, L., & Epstein, J. N. (2018). Fluid reasoning and reading difficulties among children with ADHD. *Applied Neuropsychology: Child, 8*(4), 307–318. https://doi.org/10.1080/21622965.2018.1466706

Marcopulos, B., & Lojek, E. (2019). Introduction to the special issue: Are modern neuropsychological assessment methods really "modern"? Reflections on the current neuropsychological test armamentarium. *The Clinical Neuropsychologist, 33*(2), 187–199. https://doi.org/10.1080/13854046.2018.1560502

Markopoulos, P., Read, J. C., MacFarlane, S., & Hoysniemi, J. (2008). *Evaluating children's interactive products: Principles and practices for interaction designers*. Elsevier.

Martins, A. I., Rosa, A. F., Queirós, A., Silva, A., & Rocha, N. P. (2015). European Portuguese validation of the system usability scale (SUS). *Procedia Computer Science, 67*, 293–300. https://doi.org/10.1016/j.procs.2015.09.273

Martinussen, R., & Tannock, R. (2006). Working memory impairments in children with attention-deficit hyperactivity disorder with and without comorbid language learning disorders. *Journal of Clinical and Experimental Neuropsychology, 28*(7), 1073–1094. https://doi.org/10.1080/13803390500205700

Maruff, P., Thomas, E., Cysique, L., Brew, B., Collie, A., Snyder, P., & Pietrzak, R. H. (2009). Validity of the CogState brief battery: Relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. *Archives of Clinical Neuropsychology, 24*(2), 165–178. https://doi.org/10.1093/arclin/acp010

Marzocchi, G. M., Oosterlaan, J., Zuddas, A., Cavolina, P., Geurts, H., Redigolo, D., Vio, C., & Sergeant, J. A. (2008). Contrasting deficits on executive functions between ADHD and reading disabled children. *Journal of Child Psychology and Psychiatry, 49*(5), 543–552. https://doi.org/10.1111/j.1469-7610.2007.01859.x

Marzuki, M. F. M., Yaacob, N. A., & Yaacob, N. M. (2018). Translation, cross-cultural adaptation, and validation of the Malay version of the system usability scale questionnaire for the assessment of mobile apps. *JMIR Human Factors, 5*(2), e10308. https://doi.org/10.2196/10308

Matar, E., Shine, J. M., Halliday, G. M., & Lewis, S. J. (2020). Cognitive fluctuations in Lewy body dementia: Towards a pathophysiological framework. *Brain, 143*(1), 31–46. https://doi.org/10.1093/brain/awz311

Mather, N., & Schneider, D. (2023). The Use of Cognitive Tests in the Assessment of Dyslexia. *Journal of Intelligence, 11*(5), 79. https://doi.org/10.3390/jintelligence11050079

Mattison, R. E., & Mayes, S. D. (2012). Relationships between learning disability, executive function, and psychopathology in children with ADHD. *Journal of Attention Disorders*, *16*(2), 138–146. https://psycnet.apa.org/doi/10.1177/1087054710380188

Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., & Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods, 42*(2), 525–541. https://doi.org/10.3758/BRM.42.2.525

Mazzocco, M. M. M., Hagerman, R. J., Cronister-Silverman, A., & Pennington, B. F. (1992). Specific frontal lobe deficits among women with the fragile X gene. *Journal of the American Academy of Child & Adolescent Psychiatry*, *31*(6), 1141–1148. https://doi.org/10.1097/00004583-199211000-00025

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, *38*(1), 115–142. https://doi.org/10.1037/0012-1649.38.1.115

McCormack, T., & Atance, C. M. (2011). Planning in young children: A review and synthesis. *Developmental Review*, *31*(1), 1–31. https://doi.org/10.1016/j.dr.2011.02.002

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

McGarry, L. M., & Carter, A. G. (2017). Prefrontal cortex drives distinct projection neurons in the basolateral amygdala. *Cell Reports*, *21*(6), 1426–1433. https://doi.org/10.1016/j.celrep.2017.10.046

McGrath, L. M., Peterson, R. L., & Pennington, B. F. (2019). The Multiple Deficit Model: Progress, Problems, and Prospects. *Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading*, *24*(1), 7–13. https://doi.org/10.1080/10888438.2019.1706180

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–179). Guilford Press.

McHenry, M. S., Mukherjee, D., Bhavnani, S., Kirolos, A., Piper, J. D., Crespo-Llado, M. M., & Gladstone, M. J. (2023). The current landscape and future of tablet-based cognitive assessments for children in low-resourced settings. *PLOS Digital Health, 2*(2), e0000196. https://doi.org/10.1371/journal.pdig.0000196

McKinlay, A., & McLellan, T. (2011). Does mode of presentation affect performance on the Tower of London task? *Clinical Psychologist, 15*(2), 63–68. https://psycnet.apa.org/doi/10.1111/j.1742-9552.2011.00021.x

McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetic learning difficulties. *Journal of Experimental Child Psychology, 74*(3), 240–260. https://doi.org/10.1006/jecp.1999.2516

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Meltzer, L. (Ed.). (2007). *Executive function in education: From theory to practice*. Guilford Press.

Menghini, D., Finzi, A., Benassi, M., Bolzani, R., Facoetti, A., Giovagnoli, S., Ruffino, M., & Vicari, S. (2010). Different underlying neurocognitive deficits in developmental dyslexia: A comparative study. *Neuropsychologia*, *48*(4), 863–872. https://doi.org/10.1016/j.neuropsychologia.2009.11.003

Meyers, J. E., & Meyers, K. R. (1995). Rey complex figure test under four different administration procedures. *The Clinical Neuropsychologist, 9*(1), 63–67. https://doi.org/10.1080/13854049508402059

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*(3), 287–304. https://doi.org/10.1207/s15324818ame0904_1

Ministero dell'Istruzione. (2020). *Gli alunni con Disturbi Specifici dell'Apprendimento (DSA) nell'a.s. 2018/2019*. Ufficio Gestione Patrimonio Informativo e Statistica.

Mirmehdi, R., Alizadeh, H., & Seifnaraghi, M. (2009). The efficacy of executive functions training on mathematics and reading performance of children with learning disabilities. *J. Res. Except. Child, 1*, 1–12.

Mitsopoulos, C., Mareschal, D., & Cooper, R. P. (2015). Model-based analysis of the Tower of London task. In J. Pineau & P. Dayan (Eds.), *Reinforcement learning and decision making 2015* (pp. 198–202).

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Mohai, K., Kálózi-Szabó, C., Jakab, Z., Fecht, S. D., Domonkos, M., & Botzheim, J. (2022). Development of an adaptive computer-aided soft sensor diagnosis system for assessment of executive functions. *Sensors, 22*(16), 5880. https://doi.org/10.3390/s22155880

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. De Gruyter Mouton. https://doi.org/10.1515/9783110813203

Moll, K., Göbel, S. M., Gooch, D., Landerl, K., & Snowling, M. J. (2016). Cognitive risk factors for specific learning disorder: Processing speed, temporal processing, and working memory. *Journal of Learning Disabilities, 49*(3), 272–281. https://doi.org/10.1177/0022219414547221

Montague, M. (2008). Self-regulation strategies to improve mathematical problem solving for students with learning disabilities. *Learning Disability Quarterly, 31*(1), 37–44. https://doi.org/10.2307/30035524

Morrison, G. E., Simone, C. M., Ng, N. F., & Hardy, J. L. (2015). Reliability and validity of the NeuroCognitive Performance Test, a web-based neuropsychological assessment. *Frontiers in Psychology, 6*, Article 1652. https://doi.org/10.3389/fpsyg.2015.01652

Morris, R. G. (1994). Working memory in Alzheimer-type dementia. *Neuropsychology, 8*(4), 544–554. https://doi.org/10.1037/0894-4105.8.4.544

Morris, R. G., Ahmed, S., Syed, G. M., & Toone, B. K. (1993). Neural correlates of planning ability: Frontal lobe activation during the Tower of London task. *Neuropsychologia, 31*(12), 1367–1378. https://doi.org/10.1016/0028-3932(93)90104-8

Morris, R. G., Feigenbaum, J. D., Bullock, P., Polkey, C. E., & Shoqeirat, M. (1997). Planning ability after frontal and temporal lobe lesions in humans: The effects of selection equivocation and working memory load. *Cognitive Neuropsychology, 14*(7), 1007–1027. https://psycnet.apa.org/doi/10.1080/026432997381330

Morris, R. G., Gick, M. L., & Craik, F. I. (1988). Processing resources and age differences in working memory. *Memory & Cognition, 16*(4), 362–366. https://doi.org/10.3758/BF03197047

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(3), 191–205. https://doi.org/10.1177/001316444700700201

Mountain, M. A., & Snow, W. G. (1993). Wisconsin Card Sorting Test as a measure of frontal pathology: A review. *The Clinical Neuropsychologist, 7*(1), 108–118. https://psycnet.apa.org/doi/10.1080/13854049308401893

Moura, O., Simões, M. R., & Pereira, M. (2014). Executive functioning in children with developmental dyslexia. *The Clinical Neuropsychologist, 28*(Suppl. 1), S20–S41. https://doi.org/10.1080/13854046.2014.964326

Mulder, H., & Cragg, L. (2014). Executive function and academic achievement: Current theories and measures. In S. Goldstein & J. A. Naglieri (Eds.), *Handbook of executive functioning* (pp. 405–418). Springer. https://doi.org/10.1002/icd.1836

Mungkhetklang, C., Crewther, S. G., Bavin, E. L., Goharpey, N., & Parsons, C. (2016). Comparison of Measures of Ability in Adolescents with Intellectual Disability. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00683

Nader, A.-M., Courchesne, V., Dawson, M., & Soulières, I. (2016). Does WISC-IV Underestimate the Intelligence of Autistic Children? *Journal of Autism and Developmental Disorders*, *46*(5), 1582–1589. https://doi.org/10.1007/s10803-014-2270-z

Navarro, J. J., & Mourgues-Codern, C. V. (2018). Dynamic assessment and computerized adaptive tests in reading processes. *Journal of Cognitive Education & Psychology*, *17*(1), https://psycnet.apa.org/doi/10.1891/1945-8959.17.1.70

Nebeker, C., Torous, J., & Bartlett Ellis, R. J. (2019). Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Medicine*, *17*(1), 137. https://doi.org/10.1186/s12916-019-1377-7

Nelson, T. D., Nelson, J. M., James, T. D., Clark, C. A. C., Kidwell, K. M., & Espy, K. A. (2017). Executive control goes to school: Implications of preschool executive performance for observed elementary classroom learning engagement. *Developmental Psychology, 53*(5), 836–844. https://doi.org/10.1037/dev0000296

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104). Prentice-Hall.

Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 249–256). ACM. https://doi.org/10.1145/97243.97281

Nikolaus, S., Bode, C., Taal, E., Vonkeman, H. E., Glas, C. A., & van de Laar, M. A. (2014). Acceptance of new technology: A usability test of a computerized adaptive test for fatigue in rheumatoid arthritis. *JMIR Human Factors, 1*(1), e4. https://doi.org/10.2196/humanfactors.3424

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130–159. https://doi.org/10.1037/a0026699

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 1–18). Plenum Press. https://doi.org/10.1007/978-1-4757-0629-1_1

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Oades, R. D., & Christiansen, H. (2008). Cognitive switching processes in young people with attention-deficit/hyperactivity disorder. *Archives of Clinical Neuropsychology, 23*(1), 21–32. https://doi.org/10.1016/j.acn.2007.09.002

O'Brien, J. W., Dowell, L. R., Mostofsky, S. H., Denckla, M. B., & Mahone, E. M. (2010). Neuropsychological profile of executive function in girls with attention-deficit/hyperactivity disorder. *Archives of Clinical Neuropsychology, 25*(7), 656–670. https://doi.org/10.1093/arclin/acq050

Odeh, A., & Obaidat, O. (2013). The effectiveness of computerized adaptive testing in estimating mental ability using Raven's Matrices. *Dirasat: Educational Sciences, 40*(1).

Ongür, D., Ferry, A. T., & Price, J. L. (2003). Architectonic subdivision of the human orbital and medial prefrontal cortex. *Journal of Comparative Neurology, 460*(3), 425–449. https://doi.org/10.1002/cne.10609

Orsini, A., & Pezzuti, L. (2013). *WAIS-IV. In Contributo alla taratura italiana (16–69 anni).* Giunti O.S.

Orsini A., Pezzuti, L., Picone, L. (2012). *WISC-IV: Wechsler Intelligence Scale for Children-IV – Quarta Edizione. Manuale di somministrazione e scoring.* Giunti O.S.

Owen, A. M. (1997). Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Progress in Neurobiology*, *53*(4), 431–450. https://doi.org/10.1016/S0301-0082(97)00042-7

Owen, A. M., Doyon, J., Petrides, M., & Evans, A. C. (1996). Planning and spatial working memory: A positron emission tomography study in humans. *European Journal of Neuroscience*, *8*(2), 353–364. https://doi.org/10.1111/j.1460-9568.1996.tb01219.x

Owen, A. M., Downes, J. D., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia*, *28*(10), 1021–1034. https://doi.org/10.1016/0028-3932(90)90137-D

Owen, A. M., Morris, R. G., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1996). Double dissociations of memory and executive functions in working memory tasks following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Brain*, *119*(5), 1597–1615. https://doi.org/10.1093/brain/119.5.1597

Overton, M., Pihlsgård, M., & Elmståhl, S. (2016). Test administrator effects on cognitive performance in a longitudinal study of ageing. *Cogent Psychology*, *3*(1), 1260237. https://doi.org/10.1080/23311908.2016.1260237

Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia*, *28*(10), 1021–1034. https://doi.org/10.1016/0028-3932(90)90137-D

Owen, A. M., Sahakian, B. J., Hodges, J. R., Summers, B. A., Polkey, C. E., & Robbins, T. W. (1995). Dopamine-dependent frontostriatal planning deficits in early Parkinson's disease. *Neuropsychology, 9*(1), 126–140. https://doi.org/10.1037/0894-4105.9.1.126

Panek, P. E., & Stoner, S. B. (1980). Age differences on Raven's Coloured Progressive Matrices. *Perceptual and Motor Skills, 50*, 977–978.

Papagno, C., Cappa, S. F., Garavaglia, G., Capitani, E., Laiacona, M., Lucchelli, F., & Vallar, G. (2007). *Batteria per la valutazione del linguaggio in soggetti con deficit neurologico*. Raffaello Cortina Editore.

Parsons, T. D. (2016). Neuropsychological Assessment 3.0. In *Clinical neuropsychology and technology: What's new and how we can use it* (pp. 65–96). Springer International Publishing. https://doi.org/10.1007/978-3-319-31075-6_5

Parsons, T. D., Carlew, A. R., Magtoto, J., & Stonecipher, K. (2019). The potential of function-led technologies for assessment and rehabilitation of executive functions. *Rehabilitation Psychology*, *64*(4), 387–399. https://doi.org/10.1080/09602011.2015.1109524

Parsey, C. M., & Schmitter-Edgecombe, M. (2013). Applications of technology in neuropsychological assessment. *The Clinical Neuropsychologist, 27*(8), 1328–1361. https://doi.org/10.1080/13854046.2013.834971

Passolunghi, M. C., De Blas, G. D., Carretti, B., Gomez-Veiga, I., Doz, E., & Garcia-Madruga, J. A. (2022). The role of working memory updating, inhibition, fluid intelligence, and reading comprehension in explaining differences between consistent and inconsistent arithmetic word-problem-solving performance. *Journal of Experimental Child Psychology*, *224*, 105512. https://doi.org/10.1016/j.jecp.2022.105512

Passolunghi, M. C., & Cornoldi, C. (2008). Working memory failures in children with arithmetical difficulties. *Child Neuropsychology, 14*(5), 387–400. https://doi.org/10.1080/09297040701566662

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48–76. https://doi.org/10.1037/bul0000124

Peng, P., & Fuchs, D. (2016). A meta-analysis of working memory deficits in children with learning difficulties: Is there a difference between verbal domain and numerical domain? *Journal of Learning Disabilities*, *49*(1), 3–20. https://doi.org/10.1177/0022219414521667

Peng, P., Wang, T., Wang, C., & Lin, X. (2019). A meta-analysis on the relation between fluid intelligence and reading/mathematics: Effects of tasks, age, and social economics status. *Psychological Bulletin*, *145*(2), 189–236. https://doi.org/10.1037/bul0000182

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition*, *101*(2), 385–413. https://doi.org/10.1016/j.cognition.2006.04.008

Pennington, B. F., Santerre-Lemmon, L., Rosenberg, J., MacDonald, B., Boada, R., Friend, A., Leopold, D. R., Samuelsson, S., Byrne, B., Willcutt, E. G., & Olson, R. K. (2012). Individual prediction of dyslexia by single versus multiple deficit models. *Journal of Abnormal Psychology*, *121*(1), 212–224. https://doi.org/10.1037/a0025823

Perret, E. (1974). The left frontal lobe of man and the suppression of habitual responses in verbal categorical behaviour. *Neuropsychologia*, *12*(3), 323–330. https://doi.org/10.1016/0028-3932(74)90047-5

Perrig, S. A., & von Felten, N. (2024). Development and validation of a positively worded German version of the System Usability Scale: First study. *International Journal of Human–Computer Interaction*. https://doi.org/10.1080/10447318.2024.2434720

Phillips, L. H., Lawrie, L., Schaefer, A., Tan, C. Y., & Yong, M. H. (2021). The effects of adult ageing and culture on the Tower of London task. *Frontiers in Psychology*, *12*, 631458. https://doi.org/10.3389/fpsyg.2021.631458

Phillips, L. H., Wynn, V., Gilhooly, K. J., Della Sala, S., & Logie, R. H. (1999). The role of memory in the Tower of London task. *Memory (Hove, England)*, *7*(2), 209–231. https://doi.org/10.1080/741944066

Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the Tower of London task. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, *54*(2), 579–597. https://doi.org/10.1080/713755977

Piaget, J. (1970). Piaget's theory. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology* (Vol. 1, pp. 703–732). Wiley.

Piaget, J. (1976). Piaget's theory. In B. Inhelder, H. H. Chipman, & C. Zwingmann (Eds.), *Piaget and his school* (pp. 11–23). Springer Study Edition. Springer. https://doi.org/10.1007/978-3-642-46323-5_2

Piaget, J. (1978). *Piaget's theory of intelligence*. Prentice Hall.

Pickering, S. J. (2006). Working memory in dyslexia. In T. P. Alloway & S. E. Gathercole (Eds.), *Working memory and neurodevelopmental disorders* (pp. 7–40). Psychology Press.

Picone, L., Orsini, A., & Pezzuti, L. (2018). *Le Matrici Progressive di Raven: Contributo alla taratura italiana per soggetti tra i 6 ei 18 anni*. Giunti Psychometrics Srl-Firenze.

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50*(3), 603–610. https://doi.org/10.1177/0013164490503016

Poletti, M. (2014). *Le funzioni esecutive in età evolutiva: Modelli neuropsicologici, strumenti diagnostici, interventi riabilitativi*. FrancoAngeli.

Poletti, M. (2016). WISC-IV Intellectual Profiles in Italian Children With Specific Learning Disorder and Related Impairments in Reading, Written Expression, and Mathematics. *Journal of Learning Disabilities*, *49*(3), 320–335. https://doi.org/10.1177/0022219414555416

Polson, P. G., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies, 36*(5), 741–773. https://doi.org/10.1016/0020-7373(92)90039-N

Porteus, S. D. (1924). *A guide to Porteus Maze Test*. Training School at Vineland.

Prifitera, A., Saklofske, D. H., Weiss, L. G., & Rolfhus, E. (2008). The WISC-IV in the clinical assessment context. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical assessment and intervention* (2nd ed., pp. 3–32). Academic Press. https://doi.org/10.1016/B978-012564931-5/50002-5

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*(1), 41–70. https://doi.org/10.1016/S0160-2896(01)00067-8

Prince, M. (2017). Does active play in natural environments affect children's executive function and self-regulation? *Frontiers in Public Health, 5*, Article 346.

Prins, P. J. M., Dovis, S., Ponsioen, A., ten Brink, E., & van der Oord, S. (2011). Does computerized working memory training with game elements enhance motivation and training efficacy in children with ADHD? *Cyberpsychology, Behavior, and Social Networking*, *14*(3), 115–122. https://doi.org/10.1089/cyber.2009.0206

Pueyo, R., Junqué, C., Vendrell, P., Narberhaus, A., & Segarra, D. (2008). Raven's Coloured Progressive Matrices as a measure of cognitive functioning in Cerebral Palsy. *Journal of Intellectual Disability Research: JIDR*, *52*(Pt 5), 437–445. https://doi.org/10.1111/j.1365-2788.2008.01045.x

Purvis, K. L., & Tannock, R. (2000). Phonological processing, not inhibitory control, differentiates ADHD and reading disability. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(4), 485-494. https://doi.org/10.1097/00004583-200004000-00018

Quílez-Robres, A., Moyano, N., & Cortés-Pascual, A. (2021). Motivational, emotional, and social factors explain academic achievement in children aged 6–12 years: A meta-analysis. *Education Sciences*, *11*(9), Article 513. https://psycnet.apa.org/doi/10.3390/educsci11090513

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology*, *31*(3), 206–230. https://doi.org/10.1093/arclin/acw007

Raita, E., & Oulasvirta, A. (2011). Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers*, *23*(4), 363–371. https://doi.org/10.1016/j.intcom.2011.04.002

Rajendran, K., Rindskopf, D., O'Neill, S., Marks, D. J., Nomura, Y., & Halperin, J. M. (2013). Neuropsychological functioning and severity of ADHD in early childhood: A four-year cross-lagged study. *Journal of Abnormal Psychology*, *122*(4), 1179–1188. https://doi.org/10.1037/a0034237

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The University of Chicago Press.

Ratna, P. A., & Mehra, S. (2015). Exploring the acceptance for e-learning using the technology acceptance model among university students in India. *International Journal of Process Management and Benchmarking*, *5*(2), 194–210. https://doi.org/10.1504/IJPMB.2015.068667

Raven, J. (1980). *Parents, teachers, and children: A study of an educational home visiting scheme*. Scottish Council for Research in Education.

Raven, J. C. (1938). *Raven standard progressive matrices*. Psychological Corporation.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology, 19*(1), 137–150. https://doi.org/10.1111/j.2044-8341.1941.tb00316.x

Raven, J. C. (1947). *Coloured Progressive Matrices*. H. K. Lewis.

Raven, J. C. (1954). *Progressive Matrices 1947. Series A, AB, B*. H. R. Lewis & Co.

Raven, J. C. (1962). *Coloured Progressive Matrices, Sets A, AB, B*. HK Lewis.

Raven, J. C. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement, 26*(1), 1–16. https://doi.org/10.1111/j.1745-3984.1989.tb00314.x

Raven, J. C. (2009). *Standard Progressive Matrices*. Pearson.

Raven, J., Court, J. H., & Raven, J. C. (1990). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 3, Standard Progressive Matrices*. Oxford Psychologists Press.

Raven, J. C., Court, J. H., & Raven, J. (1989). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 2, Coloured Progressive Matrices*. H. K. Lewis.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 2, The Coloured Progressive Matrices*. Oxford Psychologists Press.

Raven, J., & Raven, J. (2003). *Raven's Progressive Matrices*. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 223–237). Springer.

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., & Acker, J. D. (2005). Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. *Cerebral Cortex, 15*(11), 1676–1689. https://doi.org/10.1093/cercor/bhi044

Read, J. C., & MacFarlane, S. (2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 Conference on Interaction Design and Children* (pp. 81–88). ACM. https://doi.org/10.1145/1139073.1139096

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills, 8*(3), 271–276. https://doi.org/10.2466/pms.1958.8.3.271

Reiter, A., Tucha, O., & Lange, K. W. (2005). Executive functions in children with dyslexia. *Dyslexia, 11*(2), 116–131. https://doi.org/10.1002/dys.289

Renzulli, J. S. (2000). The identification and development of giftedness as a paradigm for school reform. *Journal of Science Education and Technology, 9*(2), 95–114. https://doi.org/10.1023/A:1009429218821

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (R package version 2.3.9). https://CRAN.R-project.org/package=psych

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de Psychologie, 28*, 215–285.

Reynolds, C. R., & Horton, A. M., Jr. (2014). *Test of Verbal Conceptualization and Fluency*. Pro-Ed.

Rhodes, M. G. (2004). Age-related differences in performance on the Wisconsin card sorting test: A meta-analytic review. *Psychology and Aging, 19*(3), 482–494. https://doi.org/10.1037/0882-7974.19.3.482

Riccio, C. A., Wolfe, M. E., Romine, C., Davis, B., & Sullivan, J. R. (2004). The Tower of London and neuropsychological assessment of ADHD in adults. *Archives of Clinical Neuropsychology, 19*(5), 661–671. https://doi.org/10.1016/j.acn.2003.09.001

Rivolta, L., Lang, M., & Michelotti, C. (2010). Un nuovo modello di intelligenza: La teoria di Cattell-Horn-Carroll (CHC). *Items–La Newsletter del Testing Psicologico, 17*.

Robbins, T. W., James, M., Owen, A. M., Sahakian, B. J., McInnes, L., & Rabbitt, P. (1994). Cambridge Neuropsychological Test Automated Battery (CANTAB): A factor analytic study of a large sample of normal elderly volunteers. *Dementia, 5*(5), 266–281. https://doi.org/10.1159/000106735

Robinson, S. J., & Brewer, G. (2016). Performance on the traditional and the touch screen, tablet versions of the Corsi Block and the Tower of Hanoi tasks. *Computers in Human Behavior, 60*, 29–34. https://doi.org/10.1016/j.chb.2016.02.047

Robinson, S., Goddard, L., Dritschel, B., Wisley, M., & Howlin, P. (2009). Executive functions in children with autism spectrum disorders. *Brain and Cognition, 71*(3), 362–368. https://doi.org/10.1016/j.bandc.2009.06.007

Robusto, E., Stefanutti, L., & Anselmi, P. (2010). The gain-loss model: A probabilistic skill multimap model for assessing learning processes. *Journal of Educational Measurement, 47*(4), 373–394. https://doi.org/10.1111/j.1745-3984.2010.00119.x

Rock, D. L., & Nolen, P. A. (1982). Comparison of the standard and computerized versions of the Raven Coloured Progressive Matrices Test. *Perceptual and Motor Skills, 54*(1), 40–42. https://doi.org/10.2466/pms.1982.54.1.40

Roid, G. H., & Koch, C. (2017). *Leiter-3: Leiter International Performance Scale* (3rd ed.). Western Psychological Services.

Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale-Revised*. Stoelting Co.

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2016). Infant cognitive abilities: Potential building blocks of later executive functions. In J. A. Griffin, P. McCardle, & L. S. Freund (Eds.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research* (pp. 139–156). American Psychological Association. https://doi.org/10.1037/14797-007

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition, 52*(3), 326–333. https://doi.org/10.1016/S0278-2626(03)00170-2

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. John Wiley & Sons. https://doi.org/10.1002/0471725382

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules* (R package version 4.1-4). https://CRAN.R-project.org/package=TAM

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd ed.). Wiley.

Ruocco, A. C., Lam, J., & McMain, S. F. (2014). Mood-dependent memory in borderline personality disorder: Influence of state affect on recall. *Journal of Psychiatry and Neuroscience, 39*(4), 282–288. https://doi.org/10.1097/hrp.0000000000000123

Saggino, A., & Tommasi, M. (2019). *CFT 20-R: Test di Cultura Fair – Revisione* [Adattamento italiano]. Hogrefe.

Salaffi, F., Gasparini, S., Ciapetti, A., Gutierrez, M., & Grassi, W. (2013). Usability of an innovative and interactive electronic system for collection of patient-reported data in axial spondyloarthritis: comparison with the traditional paper-administered format. *Rheumatology*, *52*(11), 2062-2070. https://doi.org/10.1093/rheumatology/ket276

Salkind, N. J. (2010). Face validity. In *Encyclopedia of research design* (p. 2455). SAGE Publications.

Salvia, J., & Ysseldyke, J. E. (1991). *Assessment* (5th ed.). Houghton Mifflin.

Sannio Fancello, G., Vio, C., & Ciacchetti, C. (2021). *TOL – Torre di Londra. Test di valutazione delle funzioni esecutive (pianificazione e problem solving)*. Erickson.

Sartori, G., Job, R., & Tressoldi, P. E. (2007). *DDE-2. Batteria per la Valutazione della Dislessia e Disortografia Evolutiva-2*. Giunti O. S.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2215–2224). ACM. https://doi.org/10.1145/1978942.1979266

Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research* (1st ed.). Morgan Kaufmann.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Morgan Kaufmann.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23–74.

Schmidt, K. M., & Embretson, S. E. (2003). Item response theory and measuring abilities. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 429–444). John Wiley & Sons.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162–173. https://doi.org/10.1037/0022-3514.86.1.162

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373. https://doi.org/10.1177/014662168500900405

Schnirman, G. M., Welsh, M. C., & Retzlaff, P. D. (1998). Development of the Tower of London-Revised. *Assessment*, *5*(4), 355–360. https://doi.org/10.1177/107319119800500404

Scholnick, E. K., & Friedman, S. L. (1987). The planning construct in the psychological literature. In S. L. Friedman, E. K. Scholnick, & R. R. Cocking (Eds.), *Blueprints for thinking: The role of planning in cognitive development* (pp. 3–38). Cambridge University Press.

Schreiber, J. E., Possin, K. L., Girard, J. M., & Rey-Casserly, C. (2014). Executive function in children with attention deficit/hyperactivity disorder: The NIH EXAMINER battery. *Journal of the International Neuropsychological Society*, *20*(1), 41–51. https://doi.org/10.1017/S1355617713001100

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, *41*(4), 1101–1114. https://doi.org/10.1177/001316448104100420

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2015). Age-related changes in the mean and covariance structure of fluid and crystallized intelligence in childhood and adolescence. *Intelligence*, *48*, 15–29. https://doi.org/10.1016/j.intell.2014.10.006

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Scorza, M., Gontkovsky, S. T., Puddu, M., Ciaramidaro, A., Termine, C., Simeoni, L., Mauro, M., & Benassi, E. (2023). Cognitive Profile Discrepancies among Typical University Students and Those with Dyslexia and Mixed-Type Learning Disorder. *Journal of Clinical Medicine*, *12*(22), 7113. https://doi.org/10.3390/jcm12227113

Seidman, L. J., Biederman, J., Faraone, S. V., Weber, W., & Ouellette, C. (1997). Toward defining a neuropsychology of attention deficit-hyperactivity disorder: Performance of children and adolescents from a large clinically referred sample. *Journal of Consulting and Clinical Psychology*, *65*(1), 150–160. https://doi.org/10.1037/0022-006X.65.1.150

Seidman, L. J., Biederman, J., Monuteaux, M. C., Doyle, A. E., & Faraone, S. V. (2001). Learning disabilities and executive dysfunction in boys with attention-deficit/hyperactivity disorder. *Neuropsychology, 15*(4), 544–556. https://doi.org/10.1037/0894-4105.15.4.544

Semrud-Clikeman, M. (2012). The role of inattention on academics, fluid reasoning, and visual-spatial functioning in two subtypes of ADHD. *Applied Neuropsychology. Child, 1*(1), 18–29. https://doi.org/10.1080/21622965.2012.665766

Semendeferi, K., Lu, A., Schenker, N., & Damasio, H. (2002). Humans and great apes share a large frontal cortex. *Nature Neuroscience, 5*(3), 272–276. https://doi.org/10.1038/nn814

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *298*(1089), 199–209. https://doi.org/10.1098/rstb.1982.0082

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain, 114*(2), 727–741. https://doi.org/10.1093/brain/114.2.727

Shallice, T., & Evans, M. E. (1978). The involvement of the frontal lobes in cognitive estimation. *Cortex, 14*(2), 294–303. https://doi.org/10.1016/S0010-9452(78)80055-0

Shapiro, E. S., & Gebhardt, S. N. (2012). Comparing computer-adaptive and curriculum-based measurement methods of assessment. *School Psychology Review, 41*(3), 295–305. https://doi.org/10.1145/344949.345077

Shimoni, M., Engel-Yeger, B., & Tirosh, E. (2012). Executive dysfunctions among boys with attention deficit hyperactivity disorder (ADHD): Performance-based test and parents report. *Research in Developmental Disabilities, 33*(3), 858–865. https://doi.org/10.1016/j.ridd.2011.12.014

Shneiderman, B. (2000). Creating creativity: User interfaces for supporting innovation. *ACM Transactions on Computer-Human Interaction (TOCHI), 7*(1), 114–138. https://doi.org/10.1145/344949.345077

Simkins-Bullock, J., Brown, G. G., Greiffenstein, M. F., Malik, G. M., & McGillicuddy, J. E. (1994). Neuropsychological correlates of short-term memory distractor tasks among patients with surgical repair of anterior communicating artery aneurysms. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *7*(4), 246–252. https://doi.org/10.1037/0894-4105.8.2.246

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology*, *7*(2), 268–288. https://doi.org/10.1016/0010-0285(75)90012-2

Snowling, M. J., & Hulme, C. (2012). Annual research review: The nature and classification of reading disorders – A commentary on proposals for DSM-5. *Journal of Child Psychology and Psychiatry*, *53*(5), 593–607. https://doi.org/10.1111/j.1469-7610.2011.02495.x

Soĉan, G. (2000). Assessment of reliability when test items are not essentially tau-equivalent. *Advances in Methodology and Statistics, 15*, 23–35.

Sohlberg, M. M., & Mateer, C. A. (1987). Effectiveness of an attention-training program. *Journal of Clinical and Experimental Neuropsychology*, *9*(2), 117–130. https://doi.org/10.1080/01688638708405352

Sohlberg, M. M., & Mateer, C. A. (2001). *Cognitive rehabilitation: An integrative neuropsychological approach*. Guilford Press.

Sonuga-Barke, E. J., Dalen, L., & Remington, B. (2003). Do executive deficits and delay aversion make independent contributions to preschool attention-deficit/hyperactivity disorder symptoms? *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(11), 1335–1342. https://doi.org/10.1097/01.chi.0000087564.34977.21

Sorel, O., & Pennequin, V. (2008). Aging of the planning process: The role of executive functioning. *Brain and Cognition*, *66*(2), 196–201. https://doi.org/10.1016/j.bandc.2007.07.006

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–292. https://doi.org/10.2307/1412107

Spearman, C. (1923). *The nature of "intelligence" and the principles of cognition*. Macmillan.

Spinnler, H., & Tognoni, G. (1987). Standardizzazione e taratura italiana di test neuropsicologici. *Italian Journal of Neurological Sciences, 8*(Suppl. 6), 1–120.

Spinoso, M., Benassi, M., Mazzoni, N., Orsoni, M., Stefanutti, L., Anselmi, P., de Chiusole, D., Bacherini, A., Pierluigi, I., Garofalo, S., Balboni, G., & Giovagnoli, S. (2025). Psychometric properties of the Usability and Attitude Scale (UAS) for evaluating digital tools in children and adolescent users [*Manuscript under review*].

Spoto, A., Stefanutti, L., & Vidotto, G. (2010). Knowledge space theory, formal concept analysis, and computerized psychological assessment. *Behavior Research Methods, 42*(1), 342–350. https://doi.org/10.3758/BRM.42.1.342

Spoto, A., Stefanutti, L., & Vidotto, G. (2012). On the unidentifiability of a certain class of skill multi map based probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics, 56*, 248–255. https://doi.org/10.1016/j.jmp.2012.05.001

Spoto, A., Stefanutti, L., & Vidotto, G. (2013). Considerations about the identification of forward-and backward-graded knowledge structures. *Journal of Mathematical Psychology, 57*(5), 249–254. https://doi.org/10.1016/j.jmp.2013.09.002

Spoto, A., Stefanutti, L., & Vidotto, G. (2016). An iterative procedure for extracting skill maps from data. *Behavior Research Methods, 48*(2), 729–741. https://doi.org/10.3758/s13428-015-0609-9

Stefanutti, L. (2019). On the assessment of procedural knowledge: From problem spaces to knowledge spaces. *British Journal of Mathematical and Statistical Psychology, 72*(2), 185–218. https://doi.org/10.1111/bmsp.12139

Stefanutti, L., & Albert, D. (2003). Skill assessment in problem solving and simulated learning environments. *Journal of Universal Computer Science*, *9*(12), 1455–1468.

Stefanutti, L., Anselmi, P., de Chiusole, D., & Spoto, A. (2020). On the polytomous generalization of knowledge space theory. *Journal of Mathematical Psychology*, *94*, 102306. https://doi.org/10.1016/j.jmp.2019.102306

Stefanutti, L., Anselmi, P., & Robusto, E. (2011). Assessing learning processes with the gain-loss model. *Behavior Research Methods*, *43*(1), 66–76. https://doi.org/10.3758/s13428-010-0036-x

Stefanutti, L., & de Chiusole, D. (2017). On the assessment of learning in competence-based knowledge space theory. *Journal of Mathematical Psychology*, *80*, 22–32. https://doi.org/10.1016/j.jmp.2017.08.003

Stefanutti, L., de Chiusole, D., Anselmi, P., & Spoto, A. (2020). Extending the basic local independence model to polytomous data. *Psychometrika*, *85*(3), 684–715. https://doi.org/10.1007/s11336-020-09722-5

Stefanutti, L., de Chiusole, D., & Brancaccio, A. (2021). Markov solution processes: Modeling human problem solving with procedural knowledge space theory. *Journal of Mathematical Psychology*, *103*, 102552. https://doi.org/10.1016/j.jmp.2021.102552

Stefanutti, L., de Chiusole, D., Gondan, M., & Maurer, A. (2020). Modeling misconceptions in knowledge space theory. *Journal of Mathematical Psychology*, *99*, 102435. https://doi.org/10.1016/j.jmp.2020.102435

Stefanutti, L., de Chiusole, D., Spoto, A., & Anselmi, P. (2023). Towards a competence-based polytomous knowledge structure theory. *Journal of Mathematical Psychology*, *115*, 102781. https://doi.org/10.1016/j.jmp.2023.102781

Stefanutti, L., Heller, J., Anselmi, P., & Robusto, E. (2012). Assessing local identifiability of probabilistic knowledge structures. *Behavior Research Methods*, *44*(4), 1197–1211. https://doi.org/10.3758/s13428-012-0187-z

Stefanutti, L., & Koppen, M. (2003). A procedure for the incremental construction of a knowledge space. *Journal of Mathematical Psychology*, *47*(3), 265–277. https://psycnet.apa.org/doi/10.1016/S0022-2496(02)00022-6

Stefanutti, L., & Robusto, E. (2009). Recovering a probabilistic knowledge structure by constraining its parameter space. *Psychometrika, 74*(1), 83–96. https://doi.org/10.1007/s11336-008-9095-7

Stefanutti, L., Spoto, A., & Vidotto, G. (2018). Detecting and explaining BLIM's unidentifiability: Forward and backward parameter transformation groups. *Journal of Mathematical Psychology, 82*, 38–51. https://doi.org/10.1016/j.jmp.2017.11.001

Steinberg, L., Albert, D., Cauffman, E., Banich, M., Graham, S., & Woolard, J. (2008). Age differences in sensation seeking and impulsivity as indexed by behavior and self-report: Evidence for a dual systems model. *Developmental Psychology, 44*(6), 1764–1778. https://doi.org/10.1037/a0012955

Sterba, S. K., & Pek, J. (2012). Individual influence on model selection. *Psychological Methods, 17*(4), 582–599. https://doi.org/10.1037/a0029253

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge University Press.

Sternberg, R. J., Jarvin, L., & Grigorenko, E. L. (2011). *Explorations in giftedness*. Cambridge University Press. https://doi.org/10.1017/CBO9780511778049

Stewart, T. J., & Frye, A. W. (2004). Investigating the use of negatively phrased survey items in medical education settings: Common wisdom or common mistake? *Academic Medicine, 79*(10), S18–S20. https://doi.org/10.1097/00001888-200410001-00006

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199231881.001.0001

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Stuss, D. T. (1991). Self, awareness, and the frontal lobes: A neuropsychological perspective. In J. Strauss & G. R. Goethals (Eds.), *The self: Interdisciplinary approaches* (pp. 255–278). Springer-Verlag. https://doi.org/10.1007/978-1-4684-8264-5_13

Stuss, D. T. (1992). Biological and psychological development of executive functions. *Brain and Cognition, 20*(1), 8–23. https://doi.org/10.1016/0278-2626(92)90059-U

Stuss, D. T., & Benson, D. F. (1986). *The frontal lobes*. Raven Press.

Sullivan, J. R., Riccio, C. A., & Castillo, C. L. (2009). Concurrent validity of the Tower tasks as measures of executive function in adults: A meta-analysis. *Applied Neuropsychology, 16*(1), 62–75. https://doi.org/10.1080/09084280802644243

Swanson, H. L. (1994). Short-term memory and working memory: Do both contribute to our understanding of academic achievement in children and adults with learning disabilities? *Journal of Learning Disabilities, 27*(1), 34–50. https://doi.org/10.1177/002221949402700107

Sweet, J. J., Benson, L. M., Nelson, N. W., & Moberg, P. J. (2015). The American Academy of Clinical Neuropsychology, National Academy of Neuropsychology, and Society for Clinical Neuropsychology (APA Division 40) 2015 *TCN* Professional Practice and 'Salary Survey': Professional Practices, Beliefs, and Incomes of U.S. Neuropsychologists. *The Clinical Neuropsychologist*, *29*(8), 1069–1162. https://doi.org/10.1080/13854046.2016.1140228

Styles, I., & Andrich, D. (1993). Linking the standard and advanced forms of the Raven's Progressive Matrices in both the pencil-and-paper and computer-adaptive-testing formats. *Educational and Psychological Measurement, 53*(4), 905–925. https://doi.org/10.1177/0013164493053004004

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Tamm, L., & Juranek, J. (2012). Fluid reasoning deficits in children with ADHD: Evidence from fMRI. *Brain Research*, *1465*, 48–56. https://doi.org/10.1016/j.brainres.2012.05.021

Taylor, S. J., Barker, L. A., Heavey, L., & McHale, S. (2015). The longitudinal development of social and executive functions in late adolescence and early adulthood. *Frontiers in Behavioral Neuroscience*, *9*, Article 252. https://doi.org/10.3389/fnbeh.2015.00252

Tchanturia, K., Davies, H., Roberts, M., Harrison, A., Nakazato, M., Schmidt, U., Treasure, J., & Morris, R. (2012). Poor cognitive flexibility in eating disorders: Examining the evidence using the Wisconsin Card Sorting Task. *PLOS ONE*, *7*(1), Article e28331. https://doi.org/10.1371/journal.pone.0028331

Teo, T., Lee, C. B., & Chai, C. S. (2008). Understanding pre-service teachers' computer attitudes: Applying and extending the technology acceptance model. *Journal of Computer Assisted Learning*, *24*(2), 128–143. https://doi.org/10.1111/j.1365-2729.2007.00247.x

The International Organization for Standardization. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs)* (ISO 9241-11:1998).

Thorell, L. B. (2007). Do delay aversion and executive function deficits make distinct contributions to the functional impact of ADHD symptoms? A study of early academic skill deficits. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *48*(11), 1061–1070. https://doi.org/10.1111/j.1469-7610.2007.01777.x

Thurstone, L. L. (1938). *Primary mental abilities*. University of Chicago Press.

Toffalini, E., Giofrè, D., & Cornoldi, C. (2017). Strengths and Weaknesses in the Intellectual Profile of Different Subtypes of Specific Learning Disorder: A Study on 1,049 Diagnosed Children. *Clinical Psychological Science*, *5*(2), 402–409. https://doi.org/10.1177/2167702616672038

Toll, S. W. M., Van der Ven, S. H. G., Kroesbergen, E. H., & Van Luit, J. E. H. (2011). Executive functions as predictors of math learning disabilities. *Journal of Learning Disabilities, 44*(6), 521–532. https://doi.org/10.1177/0022219410387302

Tombaugh, T. N. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology, 19*(2), 203–214. https://doi.org/10.1016/S0887-6177(03)00039-8

Tretti, M. L., Vio, C., Terreni, A., & Marzocchi, G. M. (2014). Epidemiologia dei disturbi specifici dell'apprendimento. In C. Cornoldi (Ed.), *I disturbi dell'apprendimento* (pp. 45–70). Il Mulino.

Tun, P. A., & Lachman, M. E. (2010). The association between computer use and cognition across adulthood: Use it so you won't lose it? *Psychology and Aging*, *25*(3), 560–568. https://doi.org/10.1037/a0019543

Urgesi, C., Campanella, F., & Fabbro, F. (2011). *NEPSY-II. La batteria di test per la valutazione dello sviluppo neuropsicologico in età evolutiva.* Giunti Psychometrics.

Ünlü, A., Schrepp, M., Heller, J., Hockemeyer, C., Wesiak, G., & Albert, D. (2013). Recent developments in performance-based knowledge space theory. In *Knowledge spaces* (pp. 147–192). Springer. https://doi.org/10.1007/978-3-642-35329-1_9

Unterrainer, J. M., & Owen, A. M. (2006). Planning and problem solving: From neuropsychology to functional neuroimaging. *Journal of Physiology-Paris, 99*(4–6), 308–317. https://doi.org/10.1016/j.jphysparis.2006.03.014

Unterrainer, J., Rahm, B., Halsband, U., & Kaller, C. (2005). What is in a name: Comparing the Tower of London with one of its variants. *Cognitive Brain Research, 23*(2–3), 418–428. https://psycnet.apa.org/doi/10.1016/j.cogbrainres.2004.11.013

Unterrainer, J. M., Rahm, B., Leonhart, R., Ruff, C. C., & Halsband, U. (2003). The Tower of London: The impact of instructions, cueing, and learning on planning abilities. *Cognitive Brain Research*, *17*(3), 675–683. https://doi.org/10.1016/S0926-6410(03)00191-5

Van den Haak, M., de Jong, M., & Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*(5), 339–351. https://doi.org/10.1080/0044929031000

Van der Elst, W., Ouwehand, C., van Rijn, P., Lee, N., Van Boxtel, M., & Jolles, J. (2013). The shortened Raven Standard Progressive Matrices: Item response theory–based psychometric analyses and normative data. *Assessment, 20*(1), 48–59. https://doi.org/10.1177/1073191111415999

Van der Linden, M., Collette, F., Salmon, E., Delfiore, G., Degueldre, C., Luxen, A., & Franck, G. (2000). The neural correlates of updating of information in verbal working memory. *Memory, 8*(2), 85–95. https://doi.org/10.1080/096582199387742

Van Herwegen, J., Farran, E., & Annaz, D. (2011). Item and error analysis on Raven's Coloured Progressive Matrices in Williams syndrome. *Research in Developmental Disabilities, 32*(1), 93–99. https://doi.org/10.1016/j.ridd.2010.09.005

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

Varvara, P., Varuzza, C., Sorrentino, A. C. P., Vicari, S., & Menghini, D. (2014). Executive functions in developmental dyslexia. *Frontiers in Human Neuroscience, 8*, Article 120. https://doi.org/10.3389/fnhum.2014.00120

Venkatesan, S., & Lokesh, L. (2019). Impulsivity in students with specific learning disabilities. International Journal of Indian Psychology, 7(4), 37-47.

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a research agenda on interventions. *Decision Sciences, 39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Verhaeghen, P. (2011). Aging and executive control: Reports of a demise greatly exaggerated. *Current Directions in Psychological Science, 20*(3), 174–180. https://doi.org/10.1177/0963721411408772

Vernon, P. E. (1971). *The structure of human abilities*. Methuen.

Vilkki, J., Servo, A., & Surma-aho, O. (1996). Word list learning and prediction of recall after frontal lobe lesions. *Neuropsychology, 12*(2), 268–277. https://doi.org/10.1037/0894-4105.12.2.268

Villardita, C. (1985). Raven's Colored Progressive Matrices and intellectual impairment in patients with focal brain damage. *Cortex, 21*(4), 627–635. https://doi.org/10.1016/S0010-9452(58)80010-6

Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory: A meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience, 3*(4), 255–274. https://doi.org/10.3758/CABN.3.4.255

Walton, K., Krahn, G. L., Buck, A., Andridge, R., Lecavalier, L., Hollway, J. A., Davies, D. K., Arnold, L. E., & Havercamp, S. M. (2022). Putting "ME" into measurement: Adapting self-report health measures for use with individuals with intellectual disability. *Research in Developmental Disabilities, 128*, 104298. https://doi.org/10.1016/j.ridd.2022.104298

Wang, Y., Lei, T., & Liu, X. (2020). Chinese system usability scale: Translation, revision, psychological measurement. *International Journal of Human–Computer Interaction, 36*(10), 953–963. https://doi.org/10.1080/10447318.2019.1700644

Ward, G., & Allport, A. (1997). Planning and problem solving using the five disc Tower of London task. *The Quarterly Journal of Experimental Psychology Section A, 50*(1), 49–78. https://doi.org/10.1080/713755681

Ward, G., & Morris, R. (2005). Introduction to the psychology of planning. In R. Morris & G. Ward (Eds.), *The cognitive psychology of planning* (pp. 1–34). Psychology Press.

Watkins, L., Rogers, R., Lawrence, A. D., Sahakian, B., Rosser, A. E., & Robbins, T. (2000). Impaired planning but intact decision making in early Huntington's disease: Implications for specific fronto-striatal pathology. *Neuropsychologia, 38*(8), 1112–1125. https://doi.org/10.1016/S0028-3932(00)00028-2

Watts, K., Baddeley, A., & Williams, M. (1982). Automated tailored testing using Raven's Matrices and the Mill Hill Vocabulary tests: A comparison with manual administration. *International Journal of Man-Machine Studies, 17*(3), 331–344. https://psycnet.apa.org/doi/10.1016/S0020-7373(82)80035-7

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). Psychological Corporation.

Wechsler, D. (2008). *WAIS-IV administration and scoring manual.* The Psychological Corporation.

Weinberger, D. R. (1993). A connectionist approach to the prefrontal cortex. *Journal of Neuropsychiatry and Clinical Neurosciences, 5*(3), 241–253. https://doi.org/10.1176/jnp.5.3.241

Weiss, R. H. (2019). *CFT 20-R: Grundintelligenztest Skala 2 – Revision*. Hogrefe.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 105–140). Wiley.

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology, 108*(3), 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Wild, K., Howieson, D., Webbe, F., Seelye, A., & Kaye, J. (2008). Status of computerized cognitive testing in aging: A systematic review. *Alzheimer's & Dementia, 4*(6), 428–437. https://doi.org/10.1016/j.jalz.2008.07.003

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry, 57*(11), 1336–1346. https://doi.org/10.1016/j.biopsych.2005.02.006

Willcutt, E. G., Petrill, S. A., Wu, S., Boada, R., DeFries, J. C., Olson, R. K., & Pennington, B. F. (2013). Comorbidity between reading disability and math disability: Concurrent psychopathology, functional impairment, and neuropsychological functioning. *Journal of Learning Disabilities, 46*(6), 500–516. https://doi.org/10.1177/0022219413477476

Williams, J. R. (2008). The Declaration of Helsinki and public health. *Bulletin of the World Health Organization, 86*(8), 650–652. https://doi.org/10.2471/BLT.08.050955

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior, 22*(5), 791–800. https://doi.org/10.1016/j.chb.2004.03.005

Willoughby, M. T., Blair, C. B., Wirth, R. J., Greenberg, M., & Family Life Project Investigators. (2012). The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment, 24*(1), 226–239. https://doi.org/10.1037/a0025361

Witt, J. A., Alpherts, W., & Helmstaedter, C. (2013). Computerized neuropsychological testing in epilepsy: Overview of available tools. *Seizure, 22*(6), 416–423. https://doi.org/10.1016/j.seizure.2013.04.004

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research, 30*(1), 72–91. https://doi.org/10.1086/374697

World Health Organization. (1992). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical descriptions and diagnostic guidelines*. https://www.who.int/publications/i/item/9241544228

Wytek, R., Opgenoorth, E., & Presslich, O. (1984). Development of a new shortened version of Raven's Matrices test for application rough assessment of present intellectual capacity within psychopathological investigation. *Psychopathology, 17*(1), 49–58. https://doi.org/10.1159/000284003

Xiang, Y., Li, Y., Shu, C., Liu, Z., Wang, H., & Wang, G. (2021). Prefrontal cortex activation during verbal fluency task and Tower of London task in schizophrenia and major depressive disorder. *Frontiers in Psychiatry, 12*, 709875. https://doi.org/10.3389/fpsyt.2021.709875

Ye, S., Sun, K., Huynh, D., Phi, H. Q., Ko, B., Huang, B., & Ghomi, R. H. (2022). A computerized cognitive test battery for detection of dementia and mild cognitive impairment: Instrument validation study. *JMIR Aging*, *5*(2), Article e36825. https://doi.org/10.2196/36825

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yeniad, N., Malda, M., Mesman, J., van IJzendoorn, M. H., & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical study. *Learning and Individual Differences, 23*, 1–9. https://doi.org/10.1016/j.lindif.2012.10.004

Younger, J. W., O'Laughlin, K. D., Anguera, J. A., Bunge, S. A., Ferrer, E. E., Hoeft, F., McCandliss, B. D., Mishra, J., Rosenberg-Lee, M., Gazzaley, A., Schöner, G., & Uncapher, M. R. (2023). Better together: Novel methods for measuring and modeling development of executive function diversity while accounting for unity. *Frontiers in Human Neuroscience, 17*, 1195013. https://doi.org/10.3389/fnhum.2023.1195013

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Conway, K. P., Gershon, R., & Weintraub, S. (2014). NIH Toolbox Cognition Battery (CB): Validation of executive function measures in adults. *Journal of the International Neuropsychological Society, 20*(6), 620–629. https://doi.org/10.1017/S1355617714000472

Zhang, L., & Luo, W. (2019). Application of exploratory factor analysis in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment Volume I: Fundamental techniques* (pp. 243–261). Routledge. 10.4324/9781315187815-12

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3

Zygouris, S., & Tsolaki, M. (2015). Computerized cognitive testing for older adults: A review. *American Journal of Alzheimer's Disease & Other Dementias, 30*(1), 13–28. https://doi.org/10.1177/1533317514522852