DOCTORAL (PHD) DISSERTATION

Hanif Akhtar

Assessing Fluid Reasoning with Computerized Adaptive Testing: Psychometric and Psychological Aspects

2024

EÖTVÖS LORÁND UNIVERSITY
FACULTY OF EDUCATION AND PSYCHOLOGY

Hanif Akhtar

Assessing Fluid Reasoning with Computerized Adaptive Testing: Psychometric and Psychological Aspects

DOI ID: 10.15476/ELTE.2024.097

Doctoral School of Psychology
Head of the doctoral school: Prof. Dr. Róbert Urbán

Cognitive Psychology Programme
Head of the doctoral programme: Prof. Dr. Ildikó Király

Supervisor: Dr. Kristóf Kovács

Budapest, 2024

Table of Content

List of Tables

List of Figures

Acknowledgement

I would like to express my deepest gratitude to a number of people whose support was invaluable in the completion of this dissertation.

Firstly, my gratitude goes to my supervisor, Dr. Kristof Kovacs, for his knowledgeable guidance, patience, and helpful advice during my research. Similarly, I am thankful to Dr. Szilvia Fodor and Dr. Attila Pásztor for their constructive feedback on my dissertation.

Secondly, I owe a great deal of gratitude to my family. My wife and son have constantly provided me with strength and motivation. Their understanding and affection, especially during the periods I was dedicated to my research, have been invaluable. In the same vein, my parents and sisters have been a continual source of encouragement throughout my academic endeavors, for which I am equally thankful.

I extend my sincere gratitude to all my colleagues, especially the lecturers and staff at my home-based university, the University of Muhammadiyah Malang, Indonesia. Your extensive support, not only throughout the research process but also beyond, has been nothing short of remarkable. Additionally, I extend my thanks to my friends, colleagues, and landlord for their considerable help with many facets of my daily life in Hungary.

Finally, I would like to acknowledge the anonymous participants who volunteered their time and information for this study. Your participation was essential to the success of this research, and I am immensely thankful for your contribution.

This dissertation would not have been possible without the support and collaboration of each one of you.

Abstract

Due to its central role in the structure of cognitive abilities and its near identity with *g*, a fluid reasoning (Gf) test is the best candidate to predict overall IQ. Several Gf tests have been developed, but most of them have limitations: 1) they only measure a single narrow ability, and 2) they are developed as fixed-item tests (FIT) with a limited measurement range. There is a need for Gf tests that are flexible, efficient, and entirely free for non-commercial use. Computerized adaptive testing (CAT) can address these issues. My dissertation aimed twofold: first, to develop a new multidimensional CAT measuring two narrow abilities of Gf, and second, to compare the psychometric and psychological aspects between CAT and FIT. This dissertation was divided into four separate but related studies.

First, I performed a systematic review and meta-analysis (Chapter 2) to synthesize previous research regarding the psychological impact of CAT over FIT. Articles were eligible if they employed an empirical study that directly compared motivation and/or anxiety between CAT and FIT. The review and meta-analysis suggest that, overall, there is no effect of test type on anxiety and motivation. However, easier CAT (i.e., a CAT targeted at higher success rate), demonstrates a positive effect compared with a FIT.

The next two studies (Chapter 3) were conducted to develop a new multidimensional CAT, named the Multidimensional Induction-Deduction Computerized Adaptive Test (MID-CAT), which measures two narrow abilities of Gf. We created 530 items, administered them to a sample of N = 2247 Indonesians, and calibrated them using the Rasch model. The results indicate that the final item pool has a wide range of difficulty levels that can precisely measure a broad range of abilities. The simulation study also indicates that MID-CAT provides greater measurement efficiency than either separate unidimensional CAT or FIT.

Finally, the last study was conducted to investigate the psychometric and psychological impact of CAT compared to FIT. Participants (N = 286) were randomly assigned to one of two conditions, varying in test types (CAT vs. FIT). We employed two different measurement approaches to evaluate participants' experiences: self-report and time-based measures. The results showed that CAT outperforms FIT in terms of measurement precision but had a minimal impact on the test-taking experience. At the same time, FIT resulted in a more optimistic evaluation: participants believed they had answered more items correctly when completing FIT than when completing CAT. Nevertheless, an analysis of response times revealed that participants invested more effort when engaging with the CAT.

The studies in this dissertation suggest that while CAT offers distinct psychometric benefits, it may not lead to significantly different test experiences for most examinees. Given that examinees do not perceive a CAT to be markedly different from a FIT, the appeal of using adaptive testing as a more efficient alternative to traditional fixed-item testing is likely to increase.

Keywords: fluid reasoning, multidimensional CAT, test-taking experience, motivation, anxiety

Chapter 1: Introduction

Fluid reasoning, also known as fluid intelligence (Gf), has been considered crucial to solving a wide range of novel real-world problems. Gf contributes to many aspects of human life, such as job performance (Schmidt, 2002), educational achievement (Roth et al., 2015), and health (Gottfredson & Deary, 2004). In past research employing hierarchical factor analyses, Gf showed a nearly perfect correlation with the *g* factor (Gustafsson, 1984; Weiss et al., 2013), even though this result has been challenged (Matzke et al., 2010). Therefore, if a single measure is required, a Gf test is the best option to predict overall IQ due to its central role in the structure of abilities and its near-identity with *g* (Kovacs & Conway, 2019).

Currently, several Gf tests are available, most of which are standardized, copyright-protected commercial tests. Even though standardized commercial tests provide clear benefits in clinical and industrial settings, researchers often dislike them due to their inflexibility and increased research costs. Therefore, several tests have been developed for research purposes, such as the series tests (e.g., Kyllonen et al., 2019) and matrices tests (e.g., Chierchia et al., 2019; Heydasch et al., 2013; Koch et al., 2022). However, there is room for further development. First, most tests measure only one narrow ability of Gf (i.e., inductive reasoning). Second, most tests are developed as fixed-item tests (FITs). FITs typically have a limited measurement range and are mostly designed for examinees with average ability levels to increase overall measurement precision. However, using items that are too easy or too difficult for many examinees can result in floor or ceiling effects, which introduce significant measurement errors.

Computerized adaptive testing (CAT) can address the limitations of FIT. CAT represents an advancement in computer-based tests, wherein items are selected adaptively based on previous responses so that the presented items match the examinees' ability levels. From a psychometric perspective, CAT is considered the gold standard in measurement due to its ability to provide more accurate measurements with fewer items for all ability levels (Wainer, 1993). However, in practice, CAT has not been well implemented. Apart from the complexity of its development process, its equivalence with FIT regarding test-taking experience is also debatable. For instance, Betz and Weiss (1976) discovered that students reported higher motivation levels in CAT than in FIT but also reported higher anxiety. In addition, several features of CAT, such as the inability to skip items and return to them later are disliked by examinees (Tonidandel & Quiñones, 2000)

My research is motivated by two factors: 1) the lack of Gf tests that are flexible, efficient, and entirely free for non-commercial use, and 2) the lack of evidence indicating equivalence between CAT and FIT, especially from the psychological aspects of test-takers. Therefore, my dissertation aims are twofold. *First*, I aim to develop a new CAT for measuring Gf. More specifically, I have developed a multidimensional CAT (MCAT) since it measures two narrow factors of Gf: inductive and deductive reasoning. *Second*, I aim to compare the psychometric and psychological aspects between CAT and FIT. The following sections will cover relevant literature regarding Gf, CAT, and the psychological aspects of test-taking.

1.1. Fluid reasoning (Gf)[1]

1.1.1. Gf in intelligence theory

Gf (as *gf*) was first proposed by Cattell (1963) along with the companion construct crystallized intelligence (*gc*) to represent two distinct dimensions of Spearman's (1904) General intelligence (*g*). Fluid intelligence was described as the ability to solve novel problems using reasoning and was hypothesized to be primarily biologically determined, while crystallized intelligence, in contrast, was defined as a knowledge-based ability and was assumed to be learned through education and experience. This model was then expanded further by his student, Horn (1968), who found more than two broad abilities. He proposed using the capital 'G' (i.e., Gf, Gc) to denote the abilities in the extended fluid-crystallized theory. According to Cattell and Horn, intelligent behaviour is best characterized by fluid reasoning (Kent, 2017; see also Cattell, 1963, Horn, 1968).

John Carroll (1993) published his book *Human cognitive abilities: A survey of factor-analytic studies* in which he summarized and reanalyzed factor-analytic studies of human cognitive abilities. The main strength of Carroll's meta-factor analysis was to provide an empirically based taxonomy of human cognitive ability in a single structured framework for the first time ever. The result of this work is an extensive theory of hierarchy called the "three-

[1] This section contains exact copies of some segments of the following paper:
Akhtar, H. (2022) Measuring Fluid Reasoning and Its Cultural Issues: A Review in the Indonesian Context. *Buletin Psikologi*. 30(2), 348-260. https://doi.org/10.22146/buletinpsikologi.74475

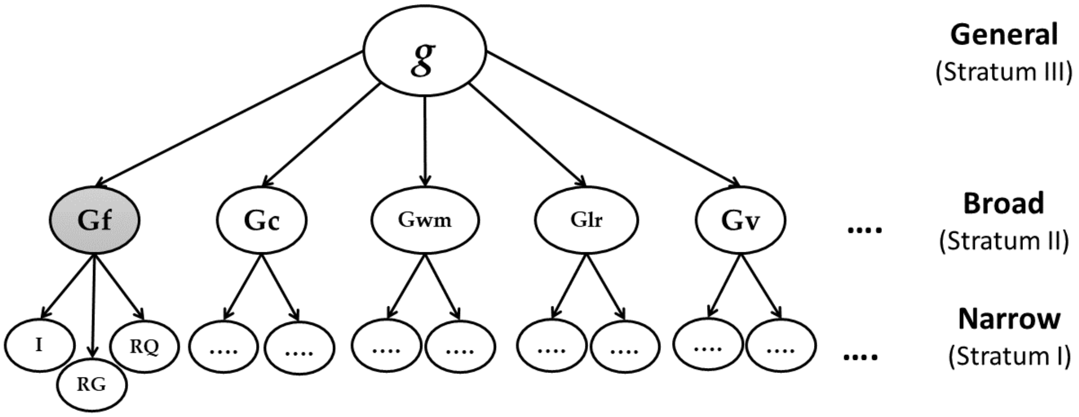
stratum model". At the first stratum of his hierarchy there are several narrow abilities. At the second stratum are eight abilities, and *g* is at the top of the structure as the third stratum. The Gf-Gc theory developed by Cattell and Horn is closely linked to Carroll's model. The major difference between the three-stratum theory of Carroll and the Gf-Gc theory of Horn and Cattell is that the latter does not postulate a general factor (g). According to the Cattell-Horn model the nature of the *g* is determined by the composition of the test battery.

An integrated Cattell-Horn and Carroll model proposed by McGrew (2009) was the first attempt to create a single taxonomy (Figure 1). Both John Horn and John Carroll accepted the unified model proposed by McGrew and colleagues and thus it became known as the Cattell-Horn-Carroll (CHC) theory. It is the most comprehensive and empirically supported psychometric theory of the structure of cognitive abilities (Flanagan & Dixon, 2014). Gf forms a major part of the CHC structure of cognitive abilities (Schneider & McGrew, 2012). Recently, Schneider & McGrew (2018) conducted a comprehensive review of the CHC theory and updated the model. The last model of CHC theory includes 18 broad cognitive abilities subsumed by more narrow abilities. Given the breadth of empirical support for the CHC intelligence structure, it offers one of the most valuable frameworks for designing and evaluating psychoeducational testing.

Within the CHC framework, Gf is classified as a broad ability encompassing three specific narrow abilities: induction, general sequential reasoning (deduction), and quantitative reasoning (Schneider & McGrew, 2012, 2018). Induction, or rule inference, is the ability to observe a phenomenon and discover the underlying principles. Deduction, or rule application, pertains to logical reasoning based on established rules and premises. While Quantitative reasoning is the skill of logical thinking with numerical data. While Gf is just one among various broad abilities in the CHC theory, it holds significant relevance owing to its strong association with the general intelligence factor, *g*.

Figure 1

Cattell-Horn-Carroll (CHC) Model of Human Cognitive Abilities



Note. Gf = Fluid Reasoning, Gc = Comprehension Knowledge, Gwm = Short-Term Working Memory, Glr = Long-Term Memory, Gv = Visual Processing, I = Induction, RG = General Sequential (Deductive) Reasoning, RQ = Quantitative Reasoning. Figure was modified from Schneider and McGrew (2012). For simplicity, I only focus on fluid reasoning; thus, not all abilities are displayed.

Although Gf is only considered one of several abilities in the Cattell-Horn, Carroll, and CHC model, they all share a similar conclusion that Gf is the most important human cognition ability. Gustafsson (1984) found that the second-order factor of Gf is statistically identical to the *g*-factor; thus, they should be considered the same factor.

The investment theory (Cattell, 1987) postulates that there is initially a single, general ability (gf) in the development of the individual which is related to the brain's maturation. Through practice and experience individuals develop abilities, and these developed abilities (i.e., gc) are influenced by gf and by struggle, motivation, and interest. Gf develops into gc as it influences the acquisition of knowledge and skills in different domains. For instance, most acquired knowledge comes from the inductive inference of partial knowledge found in different contexts (Landauer & Dumais, 1997). Indeed, gf seems to be involved in every aspect of the learning process, providing further support for Cattell’s investment theory (Kvist & Gustafsson, 2008).

1.1.2. Measuring Gf

Intelligence test score interpretation has a long history, and the varying conceptualizations of intelligence shape measurement approaches and vice versa. Boring suggested that intelligence could and should be defined operationally as that which intelligence tests measure (Boring, 1923). Kamphaus et al. (2018) provide a comprehensive review of the history of intelligence test interpretation, starting from the quantification of the general level, clinical profile analysis, psychometric profile analysis, and the modern factor-analytic model in test development. This development is interconnected with the evolution of intelligence theory. For instance, following Spearman's research, which identified a central ability influencing performance across various tasks, practitioners have commonly focused on interpreting a single general intelligence score. Nowadays, the focus has shifted to interpreting score profiles in relation to academic and achievement-related outcomes. The advent of factor-analytic research identifying Stratum II and III factors has refined these interpretations, providing a more robust framework for abilities grouping. As Kamphaus et al. (2018) noted, the gap between intelligence theory and test development has decreased, leading to more validity evidence for score interpretations.

A test should be designed a priori with a strong theoretical foundation and supported by considerable validity evidence in order to measure a particular construct (Kamphaus et al., 2018). This dissertation focusses on the CHC model. Currently, the CHC model is extensively used as the foundation for developing and interpreting tests of cognitive abilities (Flanagan et al., 2013), even though this model is not without criticism either (e.g., Canivez & Youngstrom, 2019). Using the CHC model as a guiding framework allows the positioning of tests within a recognized and established taxonomy of cognitive abilities, which, in turn, aids in the interpretation of test results. Particularly, this dissertation focuses on Gf, as this ability is important in the structure of human cognitive abilities (Kent, 2017).

There are two traditions of conceptualizing Gf: process factors and content factors. These two conceptualizations are rooted on the same basis: factorial analysis. The process factors were classified based on the reasoning processes involved: inductive (rule inference) and deductive (rule application). The content factors were classified based on the stimulus content of the test: figural, numeric/quantitative, and verbal. This conceptualization consequently affects how to measure fluid reasoning. Process factors such as inductive and deductive reasoning suggest that tests should evaluate reasoning through scenarios that challenge these cognitive processes. Meanwhile, for content factors, test designs must include tasks that assess reasoning across various domains. Lakin and Gambrell (2012) argued that measuring Gf across multiple content domains could prevent construct underrepresentation and construct-irrelevant variance. Addressing this concern, Schneider and McGrew (2018) contend that verbal, figural, and numeric content facets in Gf merely reflect factor impurities from Comprehension-knowledge (Gc), Visual processing (Gv), and Quantitative knowledge (Gq), respectively. Despite these issues, it is clear that Gf tests cluster by content, and these content-based clusters are distinct in their predictive validity (Gustafsson & Wolff, 2015). Therefore, both process and content factors aro both important.

Carroll's (1993) summary of factor analytic studies found three factors of Gf: inductive reasoning, sequential (deductive) reasoning, and quantitative reasoning. These factors also become the narrow abilities under Gf in the CHC model (Schneider & McGrew, 2018). In addition, he suggested that inductive reasoning, which is mostly measured by the figural test, has the highest loading factor on Gf.

However, Wilhelm (2005) has a different argument on how to conceptualize Gf. He argued that introducing a distinction between inductive and deductive reasoning is unnecessary since they are perfectly correlated. It can be best interpreted as content rather than process factors, with verbal, figural, and numerical content factors determining Gf. In addition, he suggested that if researchers want to measure *g* with a single task, they should select a figural reasoning test since it has the highest loading on Gf. Although two perspectives exist on how Gf should be conceptualized, both agree that figural tests exhibit the closest relationship to *g* (Carroll, 1993; Wilhelm, 2005).

According to the CHC framework, broad abilities are sufficiently represented when assessed using a minimum of two distinct tests of narrow abilities (Flanagan et al., 2013). Schneider and McGrew (2018) emphasized the inclusion of an inductive reasoning test for an adequate Gf assessment. The second test should measure deductive reasoning, as it is unrelated to a content domain. In this sense, employing figural content to assess inductive and deductive reasoning could be the most effective approach for Gf measurement.

The exploration of whether inductive and deductive reasoning are fundamentally distinct cognitive processes remains an ongoing discussion within cognitive science. The behavioral studies (Hayes et al., 2018; Stephens et al., 2018) support a single-process account of reasoning. These studies argue that both inductive and deductive reasoning can be explained through a common cognitive mechanism, albeit with distinct decision thresholds for each reasoning type. They propose that reasoning, irrespective of being inductive or deductive, operates along a continuum of cognitive evaluation, where the core processes remain fundamentally interconnected yet are adaptable to context-specific demands.

In contrast, a neuroimaging study by Goel and Dolan (2004) provides compelling evidence for the neural dissociation between these reasoning types, where the left inferior frontal gyrus exhibited a predilection for deductive reasoning, while the left dorsolateral prefrontal cortex was more involved during inductive reasoning tasks. This neural specificity emphasizes a functional bifurcation, suggesting that each reasoning type may utilize different cognitive strategies or processes. Similarly, a meta-analysis of brain scans identified different patterns of cortical activation for both the process (induction vs. deduction) and content (verbal vs. visuospatial) factors of Gf (Santarnecchi et al., 2017).

Currently, a variety of commercial Gf tests are available, including Raven's Progressive Matrices (RPM; Raven et al., 1988), Cattel's Culture Fair Intelligence Test (CFIT; Cattell et al., 1973), Test of Nonverbal Intelligence (TONI-4; Brown et al., 2010), and Naglieri Nonverbal Ability Test (NNAT-3; Naglieri, 2016). These tests have gained extensive utilization in both practical applications and academic research. In clinical and industrial contexts, copyright-protected commercial tests present notable benefits. Nevertheless, commercial tests meet with resistance within the research community, primarily due to their expense and limited administration flexibility. Some tests have been developed for research purposes, such as the series tests (e.g., Kyllonen et al., 2019) and matrices tests (e.g., Chierchia et al., 2019; Heydasch et al., 2013; Koch et al., 2022), and several tests in The International Cognitive Ability Resource project (ICAR; Condon & Revelle, 2014). Nonetheless, the current tests possess certain limitations. Firstly, they measure exclusively only one narrow ability of Gf (i.e., inductive reasoning), resulting in an inadequate representation of Gf. Secondly, these tests are structured in a fixed-item format, which has a limited measurement range and is better suited for test-takers with moderate levels of ability.

Advancements in Gf measurement have been achieved through the development of computerized adaptive test (CAT) versions. CAT can address the limitations of fixed-item tests, which have a restricted measurement range. Gf tests that have been developed using CAT include the Adaptive Matrices Test (AMT; Hornke et al., 2000), Fluid Intelligence Multistage Test (FIMT; (Martín-Fernández et al., 2016), and the Scrambled Adaptive Matrices (SAM; Klein et al., 2018). These measures have proven to be practically useful for both practitioners and researchers. However, they assess only a single, narrow aspect of Gf.

1.2. Computerized Adaptive Testing (CAT)

1.2.1. Item response theory (IRT) as a foundation of CAT

CAT is a methodology designed to increase the measurement efficiency of the tests. Currently, most CATs are based on IRT models. IRT is a group of mathematical models describing individuals' ability to interact with test items. IRT posits that the estimated ability parameters are independent of the specific items administered to individuals (Emberton & Reise, 2000). This characteristic enables the comparison of individuals' abilities irrespective of the group of items they encounter. IRT employs a mathematical function to model the likelihood of a correct response to an item based on the examinee's ability level. An examinee with a higher ability level is more likely to answer an item correctly, irrespective of the item's difficulty. Conversely, an easier item is more likely to be answered correctly by any examinee regardless of their ability.

Differences between IRT and Classical Test Theory (CTT)

IRT differs from CTT in various aspects. Reise and Henson (2003) identified fundamental differences between the two approaches. Regarding the definition of ability, under CTT, an ability estimate is considered a true score, i.e., a score that an examinee would likely achieve if they repeatedly took parallel forms of a given test. In contrast, the ability estimate derived from IRT represents the examinee's position on an ability continuum, known as theta, which predicts their responses to individual items. In terms of scoring, CTT typically employs summed scores, with ability estimates ranging from the lowest to the highest possible test scores. Conversely, in IRT, theta estimation can be obtained through advanced statistical techniques like maximum likelihood estimation or Bayesian estimation. Under

CTT, test properties such as descriptive statistics and reliability are sample-dependent. In contrast, in IRT, both test and item properties are sample-independent and remain invariant across different samples. Regarding the calculation of the standard error of measurement (SE), in CTT, SE is inversely related to the reliability coefficient and is assumed to be constant for all examinees, regardless of their score or ability level. In IRT, SE varies depending on the theta estimate, making it unique for each examinee.

Table 1 shows the difference between IRT and CTT in many aspects summarized from Embretson and Reise (2000) and Hambleton et al. (1991).

Table 1
Comparative table outlining key aspects of CTT and IRT

Aspect	**Classical Test Theory (CTT)**	**Item Response Theory (IRT)**
Focus	Entire test	Individual items within a test
Key concept	Sum score	Item characteristic curves (ICCs)
Score interpretation	Based on raw scores or transformed scores (e.g., percentiles)	Based on a latent trait or ability, often depicted on a scale
Error estimation	Assumes constant error across all test scores	Error varies across the ability continuum
Item analysis	Limited to item difficulty, discrimination, and distractor analysis	Detailed analysis of item properties using ICCs
Assumptions	Assumes homogeneity of test items and unidimensionality	Allows for multidimensionality, local independence
Applicability	Well-suited for shorter tests with homogeneous items	Preferable for tests requiring detailed item analysis and adaptivity
Data needed	Requires less data, simpler models	Requires extensive data for accurate model estimation

Reise and Henson (2003) mentioned that IRT offers many advantages over CTT: (a) estimate both item and person parameters within the same model; (b) allow for person-free item parameter estimation and item-free trait level estimation; (c) enable optimal scaling of individual differences; and (d) facilitate significant applications such as Computerized Adaptive Testing (CAT), linking scales, and evaluating Differential Item Functioning (DIF).

In IRT, the information about the SE of estimated ability is determined by *items and tests information function*. Each item provides varying amounts of information about the examinees. Easy items provide more information about those with lower abilities on the continuum and less information about those at the higher end. Computing item information constitutes a fundamental aspect of CAT since the items chosen for each examinee must yield the greatest information about that individual's abilities. Under a 1PL (Rasch) model, the item information is defined as:

$$I_i(\theta)= P_i(\theta)Q_i(\theta) \qquad (1)$$

Where I_i is the item information function of item i, $P_i(\theta)$ is the probability of a correct answer to item i, and $Q_i(\theta)$ is 1- $P_i(\theta)$. From this formula, the maximum item information function can be achieved if the $P_i(\theta)$ is 0.5. $P_i(\theta) = 0.5$ can be achieved if the item difficulty is matched to theta (see equation 6). Therefore, the maximum item information function will be obtained when the item difficulty (b) administered to the examinee equals their ability (theta). Item information calculation is independent of each other, and the test information is calculated by summing all the item information together.

$$TI(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad (2)$$

Test information is related to the precision of measurement, or SE, in the following way:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \qquad (3)$$

which is the square root of the reciprocal of the test information.

Models in IRT

Although IRT models were initially created for test items scored dichotomously (correct or incorrect), they can now be used for any model (e.g., polytomous). However, since this dissertation focuses on measuring ability, which typically uses dichotomous scores, this section will focus on the dichotomous model. Dichotomous models are often categorized based on the number of item parameters included in the model: the one-parameter model (1PL or Rasch model, Rasch, 1960), the two-parameter model (2PL, Birnbaum, 1968), and the three-parameter model (3PL, Birnbaum, 1968) or by the number of dimensions involved (unidimensional or multidimensional). The 3PL model is named because it uses three parameters: discrimination (*a*), difficulty (*b*), and pseudo-guessing (*c*). The 2PL model assumes no guessing in the data, but items can vary in difficulty and discrimination. The 1PL model assumes guessing is part of the ability, and all items have equivalent discriminations.

There is theoretically a four-parameter model (4PL, Barton & Lord, 1981) with an upper asymptote, denoted by *d*. However, this model is rarely used in practice.

The ability level, denoted as theta (θ), is assumed to follow a normal distribution in the population, with a mean of zero and a standard deviation of one. The relationship between theta and the probability of a correct response is established for each test item, resulting in the item characteristic curve (ICC). *Item difficulty* corresponds to the theta value (ability level), at which half of the examinees answer the item correctly. *Item discrimination* is determined by the slope of the ICC at the inflexion point, indicating how effectively the item distinguishes individuals' ability levels. Items with steeper slopes discriminate better. *Pseudo-guessing* parameter estimates the probability of a person with a very low ability answering the item correctly.

For the 3PL model, the probability of a correct response for person *j* to item *i* is:

$$P_{ij}(\theta_j) = c_i + (1 - c_i)\left[\frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}\right] \qquad (4)$$

where θ_j is the ability level of person *j*, b_i is item difficulty, a_i is the item discrimination, and c_i is the pseudo-guessing parameter. The 2PL model is nested in the 3PL model with the pseudo-guessing parameter set to zero. The probability of a correct response for person *j* to item *i* for the 2PL model is :

$$P_{ij}(\theta_j) = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]} \qquad (5)$$

The 1PL model is nested in the 2PL model with the discrimination parameter set to one. The probability of a correct response for person *j* to item *i* for the 1PL model is :

$$P_{ij}(\theta_j) = \frac{\exp[(\theta_j - b_i)]}{1 + \exp[(\theta_j - b_i)]} \qquad (6)$$

Mathematically, the Rasch Model can be seen as a 1PL IRT model where the item discrimination parameter is held constant across all items, and the pseudo-guessing parameter is not used. However, proponents of Rasch consider it a distinct approach. IRT works in an exploratory manner, building a model to fit data. Instead of relying on pre-defined assumptions about item discrimination, the IRT model calculates item discrimination using examinee data and considers this information in estimating ability levels. In contrast, the Rasch model aims to establish a consistent measurement scale across examinees and subsequently test whether the data fit that model (Stemler & Naples, 2021). Misfitting responses under Rasch models require a diagnosis for exclusion from the dataset based on substantive reasons.

A simulation study by Stemler and Naples (2021) indicated that Rasch estimate is *test-free* and *person-free*, as is the goal of CAT, while the IRT estimate is not. Apart from the results of the simulation study from Stemler and Naples (2021), Rasch model will be used in this dissertation since it has good measurement properties such as specific objectivity and sufficient statistics. Specific objectivity ensures that the measurement of a person's ability or an item's difficulty is consistent, regardless of the specific items or persons involved in the test. On the other hand, sufficient statistics imply that the total score of correct responses can efficiently estimate a person's ability or an item's difficulty without losing any significant information.

Multidimensional IRT

Even though most CATs rely on unidimensional IRT models, there is a growing trend towards adopting multidimensional IRT (MIRT) models (Reckase, 2009). Van der Linden and Hambleton (1997) highlighted that unidimensional models may not always be suitable for real tests. When the dimensions are correlated, responses to items measuring one dimension can offer insights into the examinee's ability on another dimension in the battery. This additional information is overlooked by traditional CTT and IRT scoring methods but can be effectively addressed using MIRT models. MIRT models are used when several different abilities (i.e., more than one theta) contribute to generating the manifest responses for an item. In particular, the unidimensional IRT model is suitable when a single factor is derived from the test items. In contrast, MIRT models are utilized when more than one factor is found to be significant.

MIRT models are categorized into two types based on item level: within-item multidimensional IRT models and between-item multidimensional IRT models (Wang & Chen, 2004). In the within-item multidimensionality model, a single item may load on multiple dimensions. However, in cognitive abilities testing, frequently, a test comprises multiple subtests, each of which is designed to measure a single dimension (Makransky & Glas, 2013). This particular IRT model is suitable for the goal of my study. Hence, for this dissertation, I examine a model in which each item exclusively provides information directly about the specific dimension it aims to measure, while also indirectly contributing to other dimensions through their intercorrelation. In the literature, this model is s referred to as between-item MIRT, often called MIRT models with a simple structure (Kim et al., 2020), or multi-unidimensional models (Sheng & Wikle, 2007).

The multidimensional generalization of the Rasch model in literature is often modelled using multidimensional random coefficients multinomial logit model (MRCMLM; Adams et al., 1997). MRCMLM is a flexible model that can incorporate the most unidimensional and multidimensional Rasch-based models, including dichotomous Rasch model (Rasch, 1960), rating scale model (Andrich, 1978), and partial credit model (Masters, 1982). This model is for a test item with the highest score category for item *i* equal to K_i. For dichotomous items, $K_i = 1$, and there are two score categories, 0 and 1. The score category is represented by *k*. The random variable X_{ik} is an indicator variable that shows whether or not the observed response is equal to *k* on item *i*. If the score is *k*, the indicator variable is assigned a 1; otherwise, it is 0. For the dichotomous case, if $X_{i1} = 1$, the response to the item was correct, and a score of 1 was assigned. The MRCMLM model can be written as follows:

$$P(\mathrm{X}_{ik} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) = \frac{\exp\left(\mathbf{b}'_{ik}\boldsymbol{\theta}_j + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)}{\sum_{k=1}^{K_i} \exp\left(\mathbf{b}'_{ik}\boldsymbol{\theta}_j + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)} \qquad (7)$$

where $\boldsymbol{\theta}_j$ has been collected into a **Dx1** column vector with **D** corresponding to the number of dimensions for person *j*. **A** is a design matrix with vector elements $\mathbf{a}_{\mathbf{ik}}$ that select the appropriate item parameter for scoring the item; **B** is a scoring matrix with vector elements $\mathbf{b}_{\mathbf{ik}}$ that indicate the dimensions that are required to obtain the score of *k* on the item; $\boldsymbol{\xi}$ is a vector of item difficulty parameters.

The model described above is designed to handle a broad range of scenarios, encompassing both dichotomous and polytomous scored test items. When dealing with dichotomous items, the multidimensional Rasch model can be expressed using the same equation.

$$P\left(\mathrm{U}_{ij} = 1; \mathbf{a1}, \mathrm{d} \mid \boldsymbol{\theta}_j\right) = \frac{\exp\left(\mathbf{a}_i\boldsymbol{\theta}_j + \mathrm{d}_i\right)}{1 + \exp\left(\mathbf{a}_i\boldsymbol{\theta}_j + \mathrm{d}_i\right)} \qquad (8)$$

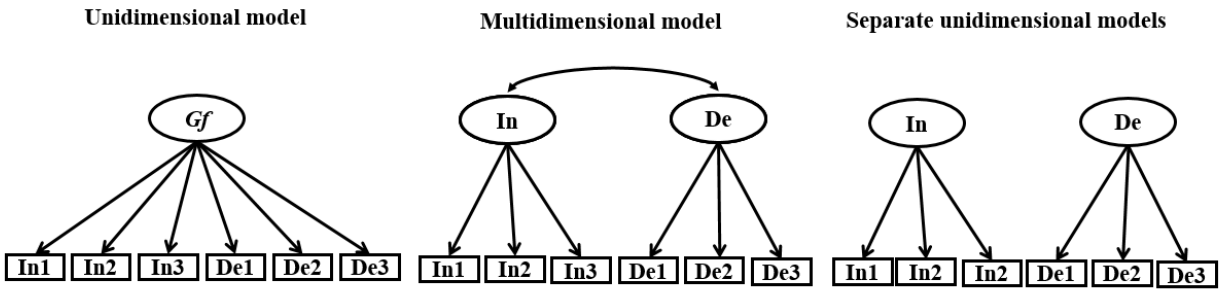
where $\mathbf{a_i}$ is a vector such that $\mathbf{a_i} = \mathbf{b_{ik}}$ and d_i is a scalar value equal to $\mathbf{a}'_{ik}\boldsymbol{\xi}$. Note that in equation 7, when k equals 0, the exponent of e becomes 0, so that term of the sum in the denominator is equal to 1.

In the context of between-item dimensionality, the $\mathbf{a_i}$-vector primarily comprises zero values except for a single element that indicates the dimension being measured by the item. That is, the test developer specifies the dimension that the item is designed to measure. To some extent, the test developer determines the values of the $\mathbf{a_i}$-vector elements instead of deriving them through usual statistical estimation procedures. In a two-dimensional scenario, $\mathbf{a_i}$-vectors with values of [1 0] or [0 1] indicate a between-item dimensionality.

Determining the model to be used is one of the essential steps in implementing adaptive testing (Magis & Barrada, 2017). The actual structure of ability dimensions is often unknown in several testing scenarios. Suppose the Gf measure consists of two subtests: induction and deduction. Three possible models could be employed: unidimensional, separate unidimensional, or multidimensional (Figure 2). Intuitively, the two subtests are related. One could assume that all items measure Gf and fit a unidimensional model. However, the estimation could be biased if the subtests do not precisely measure a single underlying construct. Alternatively, one may fit the unidimensional model separately for each subtest. However, the subscores obtained through this method are optimal only in the absence of correlations among the subtests. This is because separate unidimensional models do not consider the intercorrelation among distinct abilities. In such situations, opting for a multidimensional model becomes evident, as it takes into account the correlation between subtests during estimation, resulting in more efficient and accurate estimates (Sheng & Wikle, 2007).

Figure 2

The unidimensional and multidimensional model in Gf measurements



Multidimensional CAT (MCAT)

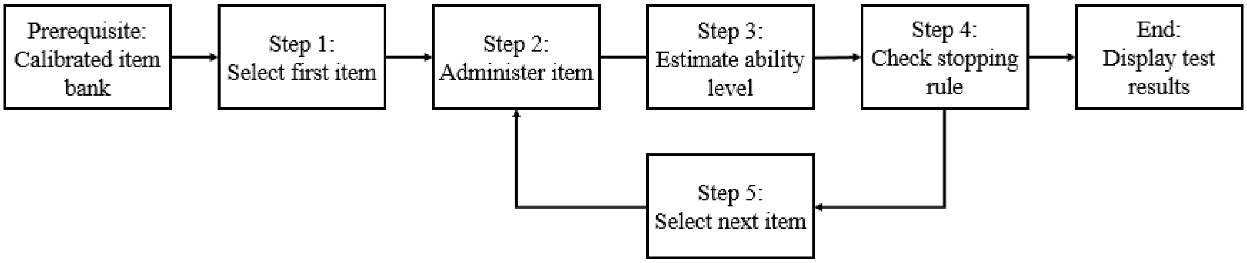
Multidimensional computerized adaptive testing (MCAT) integrates MIRT into CAT. MCAT offers several distinct and convincing benefits compared to unidimensional CAT (UCAT). *First,* MCAT yields greater information than UCAT. The abilities measured in MCAT are often correlated, and information provided by items of correlated dimensions leads to enhanced measurement efficiency. For example, Paap and colleagues (2019) reported that between-item MCAT was, on average, 20-38% shorter than UCAT when the correlation between the two measured dimensions was high (r = .80). Relatedly, MCAT provides substantially lower SE when the test length is equal in MCAT and UCAT (Segall, 1996). *Second*, MCAT can automatically ensure comprehensive content coverage through efficient item selection without relying heavily on the content balancing techniques often employed in UCAT. MCAT views targeted content domains as separate but highly intercorrelated dimensions, utilizing information from various sources across all dimensions simultaneously (Wang & Chen, 2004)

1.2.2. Components of CAT

CAT, in general, consist of four components (Reckase, 2009): (a) an item bank, (b) an item selection rule, (c) a scoring method, and (d) a termination criterion. The first item from the item bank is administered based on certain criteria (e.g., certain difficulty level, specific item, random). Scoring is performed in real-time. During a CAT session, the ability level is iteratively estimated. Items are presented based on the current trait estimate, which depends on the previous answers. If the examinee answers correctly, the next item will be harder, and vice versa. The process continues until the predetermined stopping rule has been met. Each component will be described in detail, but Figure 3 depicts the basic procedures of CAT (adapted from Oppl et al., 2017).

Figure 3

The procedure of computerized adaptive testing (CAT)


Prerequisite: Calibrated item bank
Step 1: Select first item
Step 2: Administer item
Step 3: Estimate ability level
Step 4: Check stopping rule
End: Display test results
Step 5: Select next item

Calibrated item bank

As a prerequisite for a CAT procedure, an item bank is developed, comprising items that have been previously administered and calibrated according to the chosen measurement model. Item banks for CAT can utilize 1-PL (Rasch), 2-PL, or 3-PL models and can be either unidimensional or multidimensional, depending on the fit of the item response data to the model. Reckase (2009) indicated that approximately 200 items in the pool are appropriate for participants from a standard normal distribution. However, the item difficulties in the bank must cover the entire range of the population's distribution of the trait being assessed.

Item selection methods

Item selection refers to choosing an item from the item bank to be administered to the examinee. Once an item is selected, it is marked so it cannot be reselected for the same examinee. The general idea for optimal item selection is to find the next item that minimizes the error and maximizes the information on the estimated theta. Several item selection methods are available for unidimensional or multidimensional CATs. The maximum Fisher-information (MI) method is the most commonly used, which selects the next items with the maximum information at the provisional ability level (Lord, 1980). However, the MI approach has a limitation since it assumes that the interim estimated theta is close to the true theta. In practice, this assumption is often violated, which may lead to local item selection problems, particularly in the early stage of the CAT, where the estimated theta is very different from the true theta. An extensively examined alternative strategy is the global information proposed by Chang and Ying (1996), known as Kullback-Leibler (KL) information. The KL method selects the items based on average global information when the estimators are not close to the true theta, employing the average global information procedure known as the KL index (KI).

Item selection in multidimensional models tends to be more technically complicated since items may measure more than one ability simultaneously, and redundant information may exist if the abilities are correlated. While many item selection methods exist for unidimensional models in the literature, a limited number of item selection methods are available for multidimensional models, such as D-rule, E-rule, T-rule, and KL. The D-rule focuses on maximizing the determinant of the information matrix, the T-rule selects items which increase the average unweighted information about the latent traits, while the A-rule attempts to reduce the marginal expected standard error for each θ by ignoring the covariation between traits (Chalmers, 2016). A different strategy for item selection is the KL (Chang and Ying 1996) which was then adapted by Veldkamp and van der Linden (2002) for MCAT. This method provides more stable, efficient, and precise ability estimates, especially when only a limited number of items have been administered, and is more flexible for shadow CAT design (i.e., CAT with several constraints) (Veldkamp & van der Linden, 2002). An attractive feature of KL information is that no matter how many dimensions there are, the KL information is always a scalar. Hence, KL is its immediate generalization from unidimensional to multidimensional adaptive testing (Wang et al., 2013).

Scoring method

There are two categories of ability estimation: maximum likelihood estimation method (MLE) and Bayesian methods. MLE aims to calculate the most likely ability score of the examinees given the responses to items in a test. The drawback of MLE lies in the absence of finite values for the maximum, which occurs when all responses are either correct or incorrect (van der Linden & Hambleton, 1997). Van der Linden (1999) highlighted that the MLE method is biased and inefficient, even when dealing with linear combinations of subtests in MCAT. Bayesian methods generally do not suffer from this limitation because they include additional information about the distribution of θ through a prior density function.

Several statistics, such as the expected a posteriori (EAP) and maximum a posteriori (MAP) methods, are often considered for theta estimation using the Bayesian approach. EAP calculates the expected value of the posterior distribution, considering the overall distribution of abilities, while MAP determines the most probable ability point based on the highest point of the posterior distribution. A prior distribution for the latent trait must be selected. Reckase (2009) recommended selecting it based on prior test data analyses or general knowledge about typical distributions in educational or psychological contexts. When information about the empirical theta distribution is scarce, the standard choice is often the multivariate normal distribution with an identity matrix for the variance-covariance matrix. In the context of MCAT, several simulation studies indicated that MAP outperformed EAP in terms of efficiency (Araci & Tan, 2022; ŞahiN & Gelbal, 2020). Seo and Weiss (2015) indicated that the MAP estimation method provided more accurate θ estimates than the EAP method under most conditions, and MAP showed lower observed standard errors than EAP under most conditions.

Termination criteria

Termination criteria determine when a CAT will stop. CAT could be terminated based on the number of administered items (*fixed-length*) or predetermined measurement precision (*variable-length*). The choice of stopping rule is often highly dependent on the test purpose and item bank characteristics. For example, if the test is conducted in standardized testing conditions (e.g., classical educational assessment), examinees will question the fairness of the testing if different test lengths are used. In this case, the fixed-test length could be used in the test design: the test will end after attaining a certain number of items. One of the consequences of fixed-length testing is that measurement accuracy may vary from one examiner to another. Therefore, a simulation study is needed to investigate the ideal number of items needed to administer.

In other conditions, equally precise scores among examinees are intended. It guarantees that decisions and interpretations based on test scores are equally precise for all individuals. In this case, the variable length could be used in the test design: the test will end when the standard error of the latest ability estimate is smaller than a specific criterion value. Equal measurement precision is beneficial as it aligns with CTT's *equal variance of measurement error* assumption. Additionally, this characteristic can be applied in other statistical tests that consider measurement error (Wainer, 2000). Finally, termination criteria can be based on other practical considerations, such as combining the test length and measurement precision or stopping the CAT after a specific amount of time has elapsed.

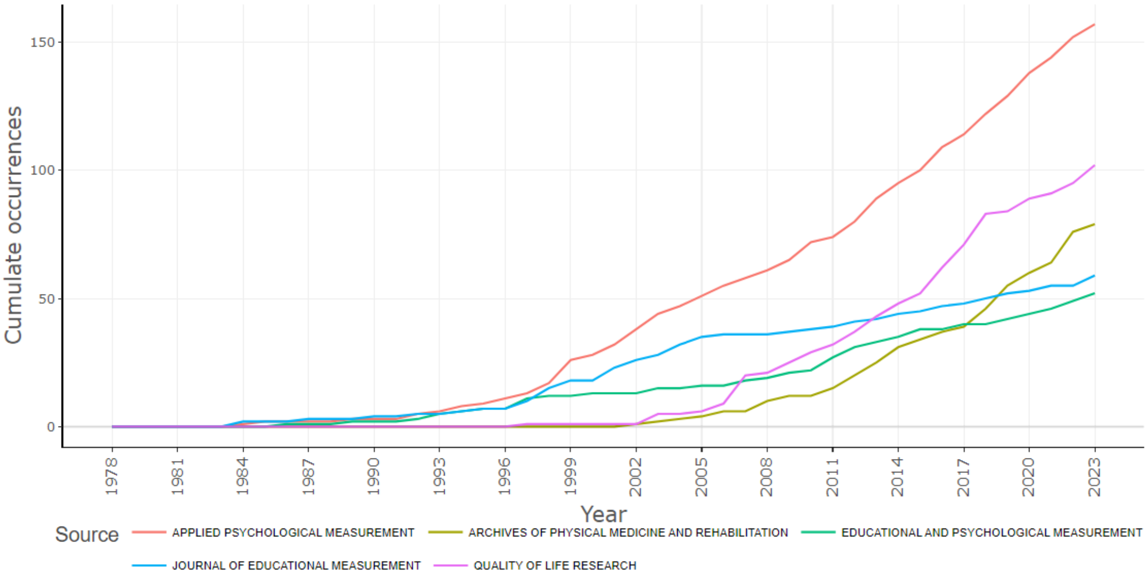
1.2.3. Trends in research on CAT[2]

Akhtar and Kovacs (2023a) performed a bibliometric analysis of research on CAT. Publications from 1978 to 2023 were collected from the Web of Science database using "*computer* adaptive test**" as the search term. There has been a significant increase in the volume of CAT research, as evidenced by the growing number of CAT-related papers published in journals over time. Early research output remained relatively stable but began to climb steadily in the early 2000s, with an accelerated increase in articles peaking just before 2023. Advancements in technology may contribute to this progress. Such growth is consistent with the bibliometric data suggesting an 11.78% annual growth rate in CAT publications. The trajectory of research volume in CAT not only expanded in quantity but also transitioned across disciplines. Initially, literature was predominantly rooted in psychology and education. However, after 2000, there was a notable influx of research from the health sciences, diversifying the application of CAT and contributing to the steep increase in publications (see Figure 4).

[2] This section contains exact copies of some segments of the following paper:
Akhtar, H. & Kovacs, K. (2023). Five Decades of Research on Computerized Adaptive Testing: A Bibliometric Analysis. [Manuscript submitted for publication]

Figure 4

Journals' production over time

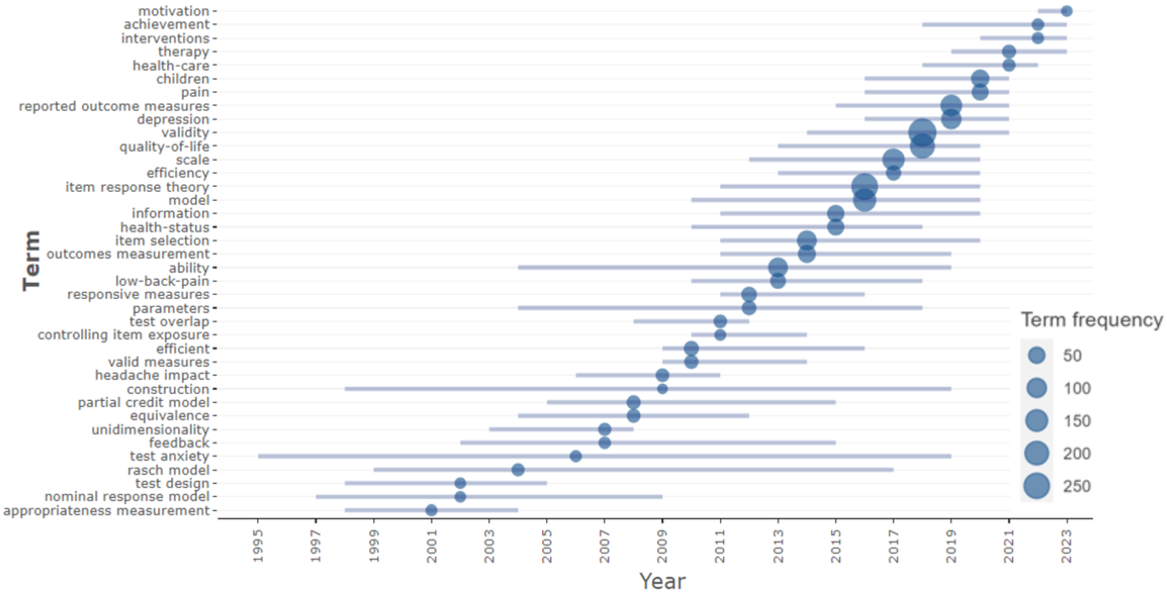


One of the main challenges in implementing CAT is the high development cost, particularly in creating CAT applications. Open-source online adaptive testing platforms (e.g., Concerto and mirtCAT) have played a significant role as they offer cost-effective solutions for developing and deploying CAT. The doubling of CAT-related articles in the last seven years, predominantly focusing on its applications, appears to align with the emergence of platforms like mirtCAT and Concerto. For example, mirtCAT was launched in 2015 (Chalmers, 2016), while Concerto was launched several years earlier (Scalise & Allen, 2015). As the availability of tutorials on these platforms increases, developing and implementing CAT is becoming more accessible and affordable. Consequently, a continued upward trend in the application of CAT across various areas is likely to be observed.

Figure 5 indicates the research trends in CAT. The bubbles correspond to research topics, with their size reflecting the frequency of each keyword's occurrence, and the grey bars denote the first and third quartiles of the occurrence distribution. The visual data indicates a pronounced increase in research emphasis on *validity* and *item response theory* within CAT literature. The increasing bubble sizes over the years indicate growing research interest and publication volume in these areas. As noted by the bar, test anxiety has maintained a steady presence in research discussions over the years. The topic of motivation also emerged recently, suggesting an interest in understanding the psychological aspects of test-taking.

Figure 5

Trend topic of research on CAT



Though periodically discussed in the literature over the years, the test-takers' perspectives on CAT (e.g., anxiety, motivation) are still considerably under-researched in the field. Yet, given the recent proliferation of research on the user experience aspects of testing, the psychological aspects of CAT are expected to become more relevant for test designers and to proliferate in the near future.

1.2.4. Benefits of CAT

CAT is widely applied in psychological, educational, and medical assessment. Unlike FITs (e.g., paper-based tests or computer-based tests where items are administered in sequence), CAT aims to choose optimal items based on selection criteria that capitalize on pre-calibrated item information and the test-takers' provisional trait estimates (Weiss, 1982). From a technical and psychometric perspective, CAT has many benefits over FIT.

First, CAT improves measurement efficiency. CAT selects the items that will be the most informative for the specific examinee. As increasing information on the examinee's ability is provided, the standard error of measurement decreases. This mechanism decreases the number of items administered without sacrificing precision (Lunz et al., 1994; Wainer, 2000). For example, in a validation study of CAT of Communicative Development Inventories, a 50-item adaptive version was comparable to the 680 items of the full scale (Kachergis et al., 2022). In general, administering fewer items also leads to quicker testing. Simms and Clark (2005) found that the adaptive version of the Schedule for Non-adaptive and Adaptive Personality took an average of 38% less time than the full computer-administered test. In research, administering fewer items is typically preferred as it reduces participant burden and fatigue (Gosling et al., 2003).

Second, CAT reduces floor and ceiling effects. Ceiling effects occur when test items are too easy so that large proportions of individuals obtain the best or maximum possible score, while floor effects occur when test items are too difficult so that large proportions of individuals obtain the minimum possible score. Ceiling and floor effects diminish the test's ability to provide precise measurements across a wide range of abilities, undermining its sensitivity and validity. By tailoring item selection to align with the examinee's ability level, it is possible to mitigate the impact of ceiling and floor effects. For example, by implementing CAT, Ware et al. (2005) demonstrated a remarkable reduction of ceiling and floor effects in a health rehabilitation measurement. In principle, CAT provides precise estimates over a wide range of abilities, while FIT is usually more precise for average examinees.

Third, CAT provides an easier way to measure growth, as measuring change over time is sometimes challenging. Using the same test multiple times carries the risk that examinees might recall their previous responses to identical items. Conversely, if distinct tests are employed on multiple occasions, assessing the extent of change becomes challenging because the tests are not on the same underlying scale. CAT offers an alternative approach to measure change or development that effectively tackles these challenges. CAT is well-implemented in measuring learning potential, which adopts the *test-train-retest* approach (de Beer, 2013). An examinee can take a CAT to acquire the baseline of ability. Subsequently, at a later time, the examinee can take another CAT to obtain a new estimate of ability. To prevent repetition of items, the CAT program can be designed to ensure that an item is not presented consecutively to the examinee. As ability or trait levels are consistently estimated on the same underlying scale through IRT scoring, change can be determined by assessing deviations from the baseline.

Fourth, CAT can enhance test security. In CAT, the test items presented to the examinee are tailored based on the examinee's previous answers. Therefore, the order and difficulty of items differ from one examinee to another. As a result, it is more difficult for individuals to share or obtain similar items, minimizing the risk of cheating. In addition, CAT eliminates the need for printing and distributing physical test booklets, which reduces the chances of unauthorized access or duplication.

Fifth, CAT provides a lot of flexibility: flexibility on item type as well as flexibility in test administration. The computer medium allows for many item types, including sound and video clips, animation, and other interactive media. For example, Harrison and Müllensiefen (2018) took benefit of this flexibility to test musical ability. In addition, CAT provides flexibility for test administrators to set the desired SE, depending on the stakes of testing (e.g., for high-stakes assessments, a very low SE is expected).

Despite the several benefits of CAT, there are several disadvantages or challenges to this testing method. *First*, to ensure that CAT precisely measures a wide range of abilities and to prevent items from becoming overexposed, CAT requires a large pool of items that can be rotated across test-takers. However, developing a large item bank can be time-consuming and costly since a larger sample size is needed (Burr et al., 2023). *Second*, CAT seldom permits test-takers to revisit and change their answers to items previously administered items because the scoring is performed in real-time, and CAT has adapted to the test-taker's performance level. This differs from FIT, where test-takers have the freedom to browse through all the items, skip some to be answered later, and review and possibly change answers. *Third*, early mistakes by high-ability test-takers can lead to considerable underestimation (Rulison & Loken, 2009). CAT can cause anxiety because test-takers cannot go back to change their early answers. If they get those wrong due to anxiety, the test does not adjust for those mistakes. *Finally*, because CAT is complex and unfamiliar for most test-takers, user needs to put more effort into public relations, explaining CAT and the reasons for using it (Thompson, 2011).

So far, the evidence focuses on the benefits of CAT from the test developer's and test administrator's perspectives. However, little is known about the benefits of CAT from the test-takers' perspective, and there are several concerns regarding the challenges of CAT. Some people might hesitate to use CAT because of their unfamiliarity with the system. For example, Goto et al. (2023) noted that Japanese education officials are questioning the use of CAT because it is deeply rooted in their academic system that all students simultaneously attempt to solve the same items. Therefore, when test items are individually tailored it may negatively impact students' acceptance. In the following section, I will provide literature reviews on the psychological aspects of test-taking, particularly when tested using CAT.

1.3. Psychological aspects of test-taking

1.3.1. Test-taking motivation in cognitive abilities testing[3]

It has been shown that variables other than ability (e.g., fatigue, anxiety, motivation, test format, test length) can impact performance in a cognitive abilities test (DeMars, 2010; Duckworth et al., 2011; Wolf & Smith, 1995). A lack of test-taking motivation becomes one of the main concerns in *low-stakes tests* (i.e., tests with no personal consequences for test-takers). Low test-taking motivation can manifest in low effort to complete the test, which will create construct-irrelevant variance in the test scores. Therefore, test scores may not reflect real ability. Numerous studies have demonstrated that motivated test takers perform better than unmotivated test takers (Duckworth et al., 2011; Eklöf et al., 2014; Wise & DeMars, 2005), even when the ability is accounted for (Cole et al., 2008; Silm et al., 2019; Thelk et al., 2009).

Expectancy-value theory

Expectancy-value theory is one theory that may explain the relationship between test-taking motivation and test performance. This theory is frequently used as a framework for test-taking motivation, a particular type of achievement motivation (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000a). According to this theory, achievement motivation for taking a test is a function of (1) *expectancy* (i.e., the expectation of success in solving the test items) and (2) *value* (i.e., the perceived values of the test). The expectancy component

[3] This section contains exact copies of some segments of the following paper:
Akhtar, H. & Firdiyanti, R. (2023). Test-taking motivation and performance: Do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? *Learning and Individual Differences*. 106, https://doi.org/10.1016/j.lindif.2023.102323

consists of ability *beliefs* (i.e., broad beliefs about competence in a given domain) and *expectancy* (i.e., expectancy for success on a specific task). The value component consists of four aspects: *importance*, *interest*, *utility*, and *cost*. Importance (or attainment value) refers to the personal importance of doing well on a task. Interest (or intrinsic value) refers to enjoyment from engaging in an activity. Utility refers to the perception that the task will be useful to meet future goals. *Cost* refers to the negative aspect of a task, including *loss of time* to engage in other desired activities, and the *effort* required to complete the task.

Measures of test-taking motivation

There are several methods to measure test-taking motivation, such as *self-reported measures* and *time-based measures*. Self-report instruments are the most common measures used for determining test-taking motivation. These instruments are administered to test-takers right after completing the test. Several self-report measures have been developed to measure test-taking motivation. For instance, *the current motivation questionnaire* (QCM; Rheinberg et al., 2001), *the student opinion scale* (SOS; Sundre & Moore, 2002), *the effort thermometer* (Baumert & Demmrich, 2001), and *the motivation instrument* (Knekta & Eklöf, 2015). The expectancy-value framework is often used in the measurement of test-taking motivation. However, most instruments that measure test-taking motivation do not include all the components of expectancy-value theory. Among these instruments, the motivation instrument (Knekta & Eklöf, 2015) is the most comprehensive expectancy-value-based questionnaire measuring five aspects of test-taking motivation (effort, expectancy, importance, interest, and test anxiety).

The other ways to measure test-taking motivation are time-based measures. The most extensively used time-based measure is Response Time Effort (RTE), proposed by Wise and Kong (2005). This measure attempts to quantify the proportion of rapid responses in the test based on the response times for each question. This measure is calculated based on the assumption that unmotivated test-takers will answer the question too quickly (i.e., before they have a chance to read and properly analyze the question). The benefit of RTE is that it is unobtrusive and does not disrupt test-takers. In addition, RTE is more beneficial than self-report when individuals are not interested in responding to questionnaires.

RTE is also considered a more objective measure of effort since its score is not influenced by response bias. Self-reports can reflect many things besides test-taking motivation (e.g., social desirability, perceived failure or lack of ability), making their interpretation less clear. However, RTE is a very specific, egregious form of non-effort. Sometimes, test takers do not answer rapidly but give less-than-full effort to items. Wise and Kuhfeld (2020) referred to these as partially engaged responses, and RTE may not represent those non-effortful responses. This idea is supported by findings showing that test-takers reduce their performance during the test even when their responses do not reflect rapid guessing behaviour (Nagy et al., 2022; Wise & Kuhfeld, 2021). They empirically found that non-rapid responses are not necessarily given with effort. These findings indicate that rapid guessing behaviour does not fully capture all aspects of disengaged response.

Although SRE and RTE are designed to measure test-taking effort, previous studies showed that the correlation between these measures is lower than expected. Wise & Kong (2005) found that the correlation between the two measures was *r* = .25, and even lower in more recent studies, *r* = .17 (Silm et al., 2019) and *r* = .18 (Hofverberg et al., 2022). A meta-analysis study found that the average correlation between SRE and test performance was *r* = .33, and the average correlation between RTE and performance was *r* = .72 (Silm et al., 2020). The difference between the two is noticeable, indicating that they may not reflect the same underlying mechanism of test-taking motivation (Silm et al., 2020). However, the reason behind this difference is not clear. RTE simply tallies rapid responses, which are typically more incorrect than correct, it logically could account for greater variance in test performance. In contrast, SRE is an overall measure of test-taking effort, potentially reflecting the extent of effort applied among other aspects.

1.3.2. Factors affecting test-taking motivation

Many factors could influence test-taking motivation. Some are related to the test context (e.g., high-stakes vs low-stakes), and some are related to the test design. Testing in the research setting is often considered low-stakes for test-takers because the test result has no personal consequences for them. Wise and DeMars (2005) identified interventions for improving test-taking efforts in low-stakes assessment: (a) increasing test relevance, (b) modifying test design, (c) promising feedback, and (d) providing external incentives. Rios

(2021) performed a meta-analysis to examine the effectiveness of those strategies. He found that the most significant improvements in the test-taking effort and performance were observed when external incentives were offered, followed by increasing test relevance. Moreover, negligible impact was detected for interventions that modified the assessment design or promised feedback. From the expectancy-value theory perspective, providing external incentives means increasing the utility value associated with the assessment, while increasing test relevance means increasing the importance value (Baumert & Demmrich, 2001; Wise & DeMars, 2005). It should be noted that Rios's meta-analysis was limited to educational assessment.

In a broader context, particular test characteristics are believed to influence test-taking motivation. When dealing with ability tests, examinees show a decrease in motivation if the items are too mentally taxing (Wise, 2006; Wise et al., 2009; Wolf et al., 1995). Therefore, test developers are recommended to reduce the number of response options and not use lengthy item stems to avoid the impression that the items are too mentally taxing (Wise, 2006; Wise et al., 2009). Certain item characteristics also influence test-taking motivation, such as item location (Akhtar, 2022b; Akhtar & Kovacs, 2023b; Pastor et al., 2019; Wise et al., 2009), item difficulty (Akhtar, 2022b; Asseburg & Frey, 2013), and item type (Sundre & Kitsantas, 2004). Items that appear at the end of the test and items that are too difficult for examinees tend to be answered carelessly.

Previous studies suggested that test-taking motivation can rise and fall during testing sessions (Barry et al., 2010; Barry & Finney, 2016; Penk & Richter, 2017). When an assessment involves two or more tests with different characteristics, the order of the test matters. When dealing with a cognitive ability test and mock exam, Wolgast et al. (2020) found students' efforts did not decrease when the cognitive ability test came first, but significantly decreased when the mock exam came first. Students had a higher level of accuracy on the cognitive ability test than on the mock exam. They found that presenting an easier test at the beginning of the testing session could be more motivating.

Akhtar & Kovacs (2023) also found the order effect when dealing with ability tests and non-ability tests. Taking non-ability tests first resulted in significantly higher effort for non-ability tests. However, the order of presentation did not matter for ability tests: neither effort nor performance varied as the function of the order of presentation in the case of ability tests. From the expectancy-value theory perspective, the mental taxation required to answer the ability test items correctly could lead to low expectancy, resulting in low effort. Motivation may also change across items during a single test. Previous studies have consistently found that effort either declines or follows a less systematic pattern as the test progresses (Akhtar, 2022b; Pastor et al., 2019; Penk & Richter, 2017; Wise et al., 2009).

1.3.3. Test-taking experience in adaptive testing

The term "test-taking experience" in this subsection describes the broader concept of test-taking motivation. This includes constructs in test-taking motivation (i.e., effort, expectancy, interest), affective reactions (e.g., anxiety during the test), and post-test reactions (e.g., feedback acceptance). In the previous section, I clearly explained the benefits of CAT from a psychometric and technical perspective. However, the test-taking experience in CAT is often neglected despite its special characteristics.

There are unique characteristics of CAT that set it apart from traditional fixed-item testing. *First*, the test items are tailored to the examinees' levels of ability, ensuring they are neither too easy nor too difficult. This differs from FIT, where items usually have a wide range of difficulty, from easy to difficult. Some argue that this aspect could boost motivation since test-takers consistently face items that offer an adequate challenge (Wise, 2014).

Second, in CAT using the Rasch or 2PL model, when item difficulty matches the estimated ability (theta), the probability of answering correctly is 50%, regardless of the examinees' abilities. This is in contrast to FIT, where the likelihood of correct answers varies with the examinees' abilities; those with higher abilities tend to answer more items correctly. However, this success rate might be perceived as too low, especially by high-ability examinees, affecting their test experience negatively. In particular, test-takers might become discouraged if they only answer half of the items correctly, which could diminish their motivation (Bergstrom et al., 1992). *Third*, several features common in FIT, such as the inability to skip or review — and possibly change— previously answered items, are not well-implemented in CAT. This characteristic leads to dissatisfaction among examinees (Vispoel et al., 2000).

Fourth, in CAT, the number of correct answers does not solely determine the final scores. Even if two examinees have the same number of correct responses, their final scores might differ significantly depending on the items they answered correctly. In contrast, in FIT, the number of correct answers largely determines the final score. For instance, some institutions question the use of CAT because it challenges their academic system where all examinees simultaneously attempt to solve the same items; thus, adopting CAT might affect examinees' acceptance of the test (Goto et al., 2023). This negative impact might be mitigated by informing examinees about how CAT works. Ortner and Caspers (2011) suggested that informing examinees about the mechanisms and procedures used in adaptive testing led to better outcomes than when only standard instructions were provided, as it helped to alleviate negative psychological effects. This procedure could mitigate the negative impacts of divided attention on irrelevant thoughts and emotions, thereby enhancing their cognitive and emotional resources for improved task performance (Ortner & Caspers, 2011). In broader context, Orive and Gerard (1987) suggested that familiar stimulus can serve as an anxiety reducer, especially when simple stimuli are involved under stress conditions. This effect was attributed to a cognitive-evaluative process where familiar stimuli are likely perceived as less threatening or more controllable, thereby diminishing the anxiety response. For these reasons, the test-taking experience in CAT might differ from that in FIT. However, the impact is non-directional, as both better and worse experiences with CAT compared to FIT are possible.

It has been frequently claimed that CAT provided a better experience than FIT (Deville, 1993; Wainer, 2000; Weiss, 1982). This claim is then often emphasized by commercial test developers and providers (e.g., Thompson, 2011). The argument for the claim is that the examinees in CAT are not administered items that are not too easy or too difficult. Therefore, low-ability examinees will not be frustrated dealing with the items that are too difficult for them, and high-ability examinees will not get bored with items that are too easy for them. This claim was partly supported by Betz and Weiss (1976): students reported higher motivation levels in CAT than in FIT. However, they also found that students reported higher anxiety levels in CAT than in FIT. In fact, some typical features of adaptive tests, such as the inability to skip the items, adversely impact the examinee's reactions (Tonidandel & Quiñones, 2000). Recent literature on the psychological effect of CAT shows mixed results. Some studies indeed found that CAT provided better experience (e.g., Fritts

& Marszalek, 2010), while others found the opposite (e.g., Ortner et al., 2014), and some did not find clear evidence favouring either position (e.g., Ling et al., 2017).

One aspect that is often neglected when evaluating CAT is feedback acceptance. Feedback acceptance is important in testing as it correlates with perceived fairness (Tonidandel et al., 2002), which, in turn, indirectly affects the face validity of the test. Tonidandel et al. (2002) found that participants were more likely to accept feedback if their perceived performance was consistent with their actual performance. In FIT, actual performance is typically closely related to perceived performance (Macan et al., 1994) because the final test score depends on the number of correct answers. This is not expected to be the case in CAT, where the relationship between actual and perceived performance is noticeably weaker (Powell, 1994). To date, no studies directly compare test acceptance in CAT and in FIT.

1.4. Motivation for this dissertation

This dissertation will address two problems. The first problem pertains to the availability of Gf measures for non-commercial use. Although there are numerous tests available for researchers to measure Gf, as discussed earlier, these existing tests come with several limitations: (a) they are developed in fixed item formats, limiting their measurement precision for both low- and high-ability examinees, (b) they only measure one narrow ability of Gf, and (c) not all tests are freely accessible to researchers. As a measure of Gf, it is important to measure at least two narrow abilities to be adequately represented (Flanagan et al., 2013; Schneider & McGrew, 2018). There is a need for flexible, accessible, efficient, and comprehensive tests measuring Gf. My dissertation addresses this need by developing a new figural multidimensional CAT that assesses two narrow abilities of Gf. Incorporating multiple figural tasks helps to prevent dependence on a single test format, often figural matrices, which might not comprehensively represent the broad domain. Moreover, using figural items can minimize cultural and linguistic biases in assessments, a crucial aspect in multicultural settings (Akhtar, 2022a).

The second problem concerns the evidence of the advantages of CAT over FIT. As presented above, the psychometric and technical benefits of CAT over FIT are undeniable. However, the evidence regarding the effect of CAT from the test-takers' perspective is limited, despite many claims that CAT provides a better experience. The psychological impact of CAT on test-takers is an under-researched area. Although multiple studies have been conducted to test this claim, it remains unclear whether CAT results in a better experience. Such an understanding would require a synthesis of previous research. However, at the time of this writing, no syntheses have been conducted related to the psychological effect of CAT over FIT. Therefore, this dissertation addresses this limitation by systematically synthesizing literature as well as verifying it in an empirical study.

1.5. Research questions and overview of the studies

This dissertation aims twofold. *The first* is to develop a new multidimensional CAT measuring two narrow abilities of Gf: inductive and deductive reasoning. *The second* is to compare the psychometric and psychological aspects between CAT and FIT. My empirical work aims to address these two general research questions:

1. **Is measurement precision different under adaptive testing and non-adaptive testing?** Specifically, is MCAT more efficient compared to separate-unidimensional CAT or FIT? How many items must be administered to reach a desirable level of measurement precision in MCAT? Is the estimated ability from the MCAT equivalent to that from FIT?
2. **Is test-taking experience different under adaptive testing and non-adaptive testing?** Specifically, are reactions to an adaptive test more favorable than to a FIT? Is feedback acceptance different under CAT compared to FIT?

The next three chapters explain four separate but related studies, each aiming at specific goals. In brief, the first study aims to compare psychological aspects of CAT and FIT. The second and third studies aim to develop a new multidimensional CAT for measuring Gf and evaluate its psychometric aspect. Meanwhile, the last study aims to compare the psychometric and psychological aspects of CAT and FIT in real testing. Research Question #1 is addressed through Chapters 3 and 4, while Research Question #2 is addressed through Chapters 2 and 4. The following is a brief overview of the studies:

Chapter 2 presents a systematic review and meta-analysis to synthesize previous research on the psychological impact of CAT over FIT. The purpose of Chapter 2 is to gain a comprehensive understanding of the effects of CAT on motivation and anxiety in comparison to traditional FIT. We examined major databases to search for articles that employed empirical studies directly comparing CAT and FIT. The main issue we wanted to address in this study was whether CAT was more motivating and induced less anxiety than FIT.

Chapter 3 aimed to develop a multidimensional CAT for measuring Gf and evaluated its psychometric aspects through a simulation study. We created 530 items divided into two subtests, all of which were administered to a large sample from Indonesia. We also performed

Monte Carlo simulations to evaluate the potential performance of the MCAT in comparison with separate-unidimensional CAT or FIT. The main issues we wanted to address in Chapter 3 were: (a) whether the Gf construct fits better in a unidimensional, separate-unidimensional, or multidimensional model; (b) whether we could generate a Gf test that has a wide range of item difficulties; (c) whether the measures are valid indicators of Gf, as shown by correlations with external measures; (d) whether the MCAT was more efficient compared to the UCAT or FIT; and (e) how many items need to be administered in high-stakes and low-stakes testing.

Chapter 4 presents an empirical study to investigate the psychometric and psychological impacts of CAT in the context of a multidimensional fluid reasoning test. Participants were randomly assigned to one of two conditions, varying in test types (MCAT vs. FIT). The main issues we wanted to address in Chapter 4 were: (a) whether measurement precision differs under MCAT and FIT; (b) whether the test-taking experience differs under MCAT and FIT.

Chapter 2: The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis[4]

2.1. Background and aims

Although many studies have been carried out on the psychometric aspects of Computerized Adaptive Testing (CAT), its psychological aspects are less researched. It has been frequently claimed that because in CAT, the presented items are matched to test-takers' ability, CAT can be more motivating and less anxiety-inducing than traditional FIT (Wainer, 2000; Weiss, 1982). The reasoning behind this claim is that test-takers with lower ability do not become anxious by items that are too difficult for them, while test-takers with higher ability are not bored by items that are too easy for them. However, the literature on CAT's psychological effects shows mixed results.

To our knowledge, currently, there is no systematic review and meta-analysis of the psychological impact of CAT compared to FIT. The purpose of this chapter is to gain a comprehensive understanding of the effects of CAT on motivation and anxiety. We aimed to synthesize the literature regarding the evidence of the effect of CAT on test-takers' motivation and anxiety compared to FIT. Motivation refers to test-taking motivation from the expectancy-value theory (Wigfield & Eccles, 2000a), which has two main components: expectancy for success and the perceived value of a task (importance, enjoyment, usefulness of the task, and effort). Anxiety refers to state anxiety, defined as a temporary emotional condition elicited by a specific situation (Spielberger, 1972). In this context, state anxiety is the anxiety in response to certain testing conditions.

[4] Chapter 2 is based upon the following paper:
Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis. *Assessment*, 30(5), 1379–1390. https://doi.org/10.1177/10731911221100995

2.2. Methods

2.2.1. Eligibility criteria

To be included in this review, studies had to meet the following criteria: 1) Original research, 2) written in English, 3) contained a comparison of state anxiety and/or state motivation (i.e., anxiety and motivation as a reaction of certain testing conditions) between CAT and FIT. The following studies were excluded: 1) oral/poster presentations, 2) studies that did not report original findings, 3) studies that did not directly compare the effect of CAT versus FIT on state anxiety and motivation. Sample characteristics and test categories were not among the inclusion and exclusion criteria.

2.2.2. Information sources and search strategy

We performed a search on seven databases where we could potentially identify peer-reviewed journal articles as well as grey literature: PsycINFO, PubMed/Medline, Scopus, Google Scholar, Proquest, EbscoHost Open Dissertation, and Web of Science, for articles published between January 1st, 1990 and December 1st, 2021, for the following keywords: “computer* adaptive test*”, “motivation”, “anxiety”, with Boolean operators AND and OR - "computer* adaptive test*" AND ("motivation" OR "anxiety"), in the title, abstract, or keywords. For the Google Scholar search result, we only extracted the 500 most relevant articles from 3190 results. The papers referenced in key articles were also reviewed to ensure no relevant studies were excluded. Duplicate results were removed.

2.2.3. Selection process

Two reviewers surveyed the title and abstract of each article in order to select articles that match the inclusion criteria. The shortlisted papers were evaluated for eligibility by the same two reviewers. Any duplicates were deleted from the final pool of papers. When necessary, the authors of the included articles were contacted for supplementary data.

2.2.4. Data extraction and data items

Two reviewers analyzed the studies, using the following classifications: 1) the psychological aspect investigated in the study (motivation, anxiety, or both), 2) characteristics of participants, 3) the construct measured by the tests, 4) the testing method compared with CAT, 5) the outcome measure, 6) document type, and 7) mean and standard deviation of each group. For the outcome measure, we only measure state anxiety and/or motivation, i.e., motivation and anxiety as a reaction to certain testing conditions. Additionally, the specific study design and the nature of the test were also considered in each study. Any disagreements between the reviewers were resolved by consensus. Articles were included only if they featured an independent variable related to the type of testing (i.e., CAT and FIT) and a direct comparison of its effect on state anxiety and/or motivation.

2.2.5. Quality assessment

Additionally to the aspects listed above, studies were also assessed with the Mixed Methods Appraisal Tool (MMAT)-2018 (Hong et al., 2018). Every included study was evaluated first on the basis of 1) the clarity of the research questions and 2) whether the collected data were adequate to address the research questions. If the answer was affirmative in both cases, then the included studies were assessed based on study design. Each of the questions was answered with “No,” “Yes,” or “Cannot tell”.

2.2.6. Meta-Analytical Procedures

The 3.3 version of the Comprehensive Meta-Analysis (CMA) software was used to compute the individual effect sizes and conduct the analyses (Borenstein et al., 2015). The dependent variable in the present meta-analysis was the standardized mean difference between the CAT and FIT groups on the outcome measures of anxiety and motivation. In consideration of the great variability of sample sizes and different outcome measures in the primary studies, the Hedges' *g* estimate was calculated by using the pooled standard deviations (Hedges, 1983). When more than one appropriate outcome measure was reported in a primary study, the average of these effect sizes was computed. The average effect size and the corresponding 95% confidence interval were calculated using the random-effects model, which incorporate heterogeneity across the included studies (Borenstein et al., 2011). Studies were weighted with the reverse of their variance based on sample size to account for differences (Borenstein et al., 2011). Before calculating the average effect size, individual studies were screened for outlying effect size values, with a standardized residual exceeding ± 3.29 considered as an outlier (Tabachnick & Fidell, 2013). A positive effect size indicated less anxiety or more motivation in the CAT condition compared to the FIT. Instead of Cohen's (1988) classical benchmarks of effect sizes (small = 0.2, medium = 0.5, large = 0.8), benchmarks from social sciences were used (small = 0.05, medium = 0.15, large = 0.20) as suggested by Bakker et al. (2019) and Kraft (2020). These benchmarks were further supported by a previous meta-analysis in which a significant positive effect of self-adaptive testing was compared to computerized-adaptive testing with 0.19 Cohen's d effect size (Pitkin & Vispoel, 2001).

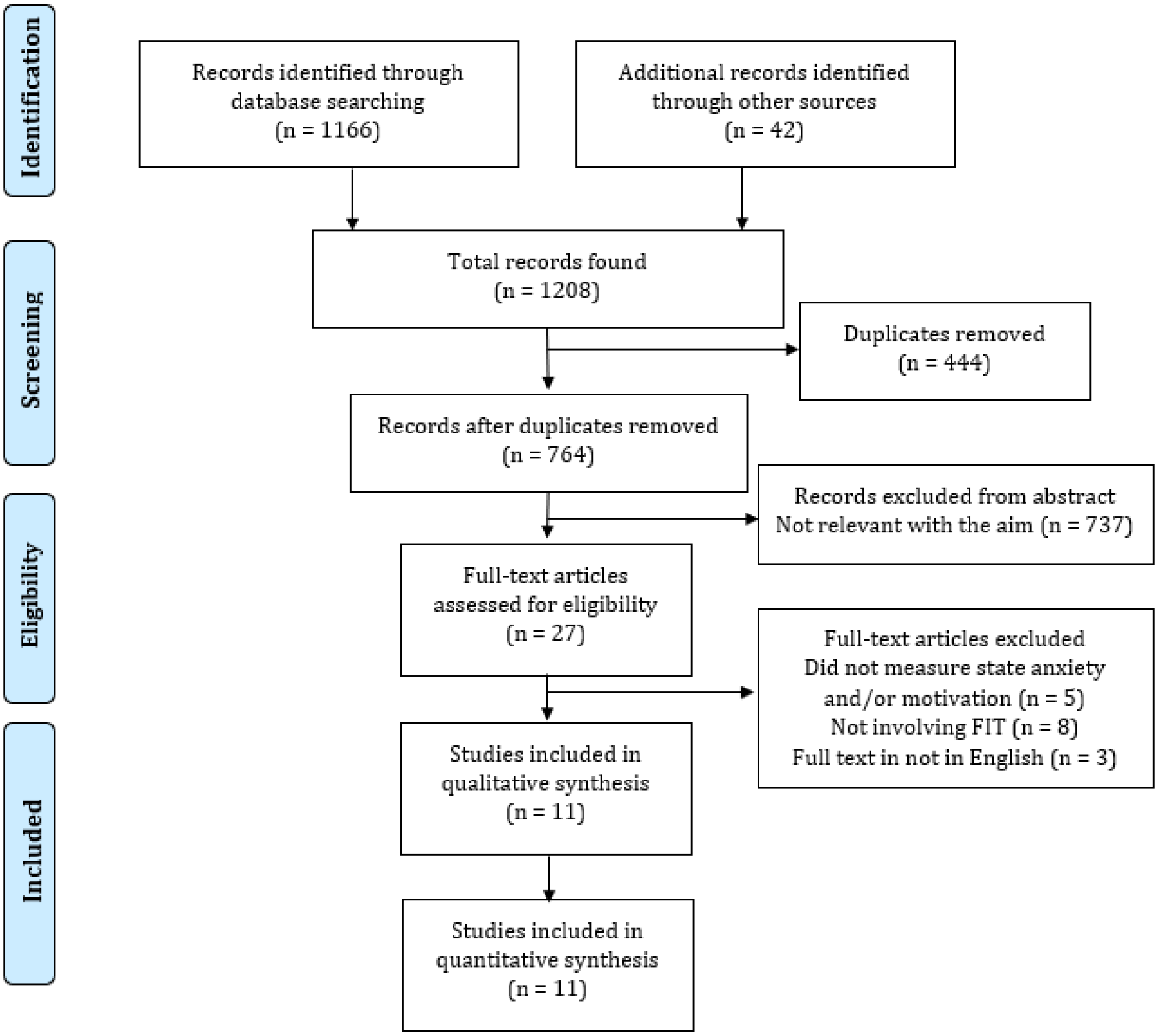
The heterogeneity of the effect sizes was estimated with the *Q*-statistic and the I^2 estimate, indicating between-study variance caused by systematic differences across primary studies beyond sampling error (Higgins et al., 2021). I^2 values above 75% suggest a substantial relative heterogeneity between primary studies in relation to total variability, which might be explained by factors on the study-level (Higgins et al., 2021). As I^2 informs about the relative percentage of between-study heterogeneity, but not the size of true variance, the absolute random variance was observed as well, referred to as Tau^2 or T^2 (Borenstein et al., 2017).

To address publication bias, grey literature was also included (i.e., theses, conference papers), and the symmetry of Begg's funnel plot and Egger's regression test was examined (Egger et al., 1997). As Sterne et al. (2011) suggested, publication bias was tested only for the overall effect, as under 10 studies, this test of asymmetry is underpowered. Subgroup analyses were performed to assess different types of CAT tests efficacy compared to FIT, in those cases where there were at least two studies to be included.

2.3. Results

The initial search produced 1208 potential articles which decreased to 764 after duplicates were removed. The titles and abstracts of the remaining articles were surveyed according to the inclusion criteria, which were met by 27 articles. Finally, after reading the full text of the articles, 11 were included in the study. Thirteen articles were removed because they did not mention any comparison of state motivation and/or state anxiety between CAT and FIT. Three papers were removed because the full-text article was not in English, only the abstract. Figure 6 illustrates the phases of article selection in accordance with PRISMA guidelines.

Figure 6
PRISMA Flowchart of the current study


Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1166)
Additional records identified through other sources (n = 42)
Total records found (n = 1208)
Duplicates removed (n = 444)
Records after duplicates removed (n = 764)
Records excluded from abstract Not relevant with the aim (n = 737)
Full-text articles assessed for eligibility (n = 27)
Full-text articles excluded Did not measure state anxiety and/or motivation (n = 5) Not involving FIT (n = 8) Full text in not in English (n = 3)
Studies included in qualitative synthesis (n = 11)
Studies included in quantitative synthesis (n = 11)

2.3.1. Characteristics of included studies

The characteristics of the included studies are summarized in Table 2. Most studies were conducted in western countries: five in the USA (Arvey et al., 1990; Fritts & Marszalek, 2010; Kiskis, 1991; Ling et al., 2017; Powers, 2001), two in Spain (Olea et al., 2000; Revuelta et al., 2003), one in Germany (Ortner et al., 2014), and one in Australia (Martin & Lazendic, 2018). Only two studies were conducted in non-Western countries: Malaysia (Mohd Ali et al., 2019) and Korea (J. Kim & McLean, 1995).

The sample size varied considerably in the included studies, ranging from 127 (Kiskis, 1991) to 12736 (Martin & Lazendic, 2018) participants. All of the studies were conducted in educational settings, except for the study by Arvey (1990) and Kiskis (1991), who conducted their study in organizational setting. All tests measured maximum performance. Four studies compared CAT with Paper-and-Pencil Fixed Item Testing (PPFIT) (Arvey, 1990; Kim & McLean, 1995; Fritts & Marszalek, 2010; Powers, 2001), five studies compared CAT with Computer-Based Fixed Item Testing (CBFIT) (Ling et al., 2017; Martin & Lazendic, 2018; Olea et al., 2000; Ortner et al., 2014; Revuelta et al., 2003), and two study compared CAT with both PPFIT and CBFIT (Mohd Ali et al., 2019, Kiskis, 1991)

Table 2

Summary of selected studies characteristics

Author(s)	Document type	Country	Psychological aspect	Participants	Construct measured by the test	Testing method to compare	Outcome measure
Kiskis, 1991	Thesis	USA	Anxiety	Applicants at personnel agency (n=127)	Clerical aptitude	PPFIT, CFIT	STAI, TAI
Kim & McLean, 1995	Conference paper	Korea	Anxiety	College students (n=208)	Math (algebra)	PPFIT	TAI
Olea et al., 2000	Journal article	Spain	Anxiety	Undergraduate students (n = 184)	English vocabulary	CFIT	SAS
Powers, 2001	Journal article	USA	Anxiety	GRE Test-takers (n = 1100)	Verbal reasoning, quantitative reasoning, analytical writing	PPFIT	TAI
Revuelta et al., 2003	Journal article	Spain	Anxiety	University students (n = 557)	English vocabulary	ECAT, CFIT	SAS
Fritts & Marszalek, 2010	Journal article	USA	Anxiety	Junior high school student (n = 132)	Math and reading ability	PPFIT	STAIC
Mohd Ali et al., 2019	Journal article	Malaysia	Anxiety	University students (n = 300)	Math (algebra)	CFIT, PPFIT	FTA (SV, CI, & PET)
Arvey, 1990	Journal article	USA	Anxiety, Motivation	Army (n=535)	Vocational aptitude	PPFIT	TAS (M&S)
Ling et al., 2017	Journal article	USA	Motivation, Anxiety	Middle school students (n = 789)	Mathematics problem-solving	ECAT, CFIT	QCM (C&I), AQ
Ortner et al., 2014	Journal article	Germany	Motivation	Secondary school students (n = 174)	Figural reasoning	CFIT	QCM (FF & PS)
Martin & Lazendic, 2018	Journal article	Australia	Motivation	Elementary and secondary school students (n = 12,736)	Numeracy skills	CFIT	MES (PME & NME)

PPFIT = Paper-and-Pencil Fixed-Item Test, CFIT = Computerized Fixed-Item Test, ECAT = Easier Computerized Adaptive Testing, TAS = Test Attitude Survey, M&S = subscale of motivation and comparative anxiety, SAS = State-Anxiety Scale, TAI = Test Anxiety Inventory, STAIC = State-Trait Anxiety Inventory for Children, STAI = State-Trait Anxiety Inventory, FTA = Friedben Test Anxiety Scale, SV, CI, & PET = subscale of Social Views, Cognitive Impairment, and Physical and Emotional Tension, QCM = Questionnaire on Current Motivation, AQ = Anxiety Questionnaire, C&I = Subscale of Challenge and Interest, FF & PS = subscale of Fear of Failure and Probability of Success, MES = Motivation and Engagement Scale, PME & NME = subscale of Positive Motivation and Engagement and Negative Motivation and Engagement

2.3.2. Quality assessment of Included Studies

None of the eleven included studies had major problems that endanger their quality. All studies had clearly formulated research questions and reported appropriate data collection. A few studies did not meet one of the methodological criteria. For example, in Powers' study (Powers, 2001), examinees were not randomly assigned to modes of exposure (CAT vs. FIT) but were allowed to self-select themselves into one of the two conditions. In addition, Powers did not control testing mode (computer-based vs. paper-based) and score-reporting (immediately vs. several weeks later) as possible confounders that could affect the result of the study. Another study that did not meet one of the criteria is the only one by Fritts and Marszalek (Fritts & Marszalek, 2010), who compared two groups from two different school districts. The two districts' testing conditions or test-taker characteristics could be different enough to confound the difference in anxiety. The summary of the quality assessment is presented in Table 3.

Table 3

Risk of bias assessment of the studies on the effect of CAT on motivation and anxiety

Author(s)	Screening questions		Methodological quality criteria			
	Are the research questions clear?	Do the collected data allow for addressing the research questions?	Are the measurements appropriate?	Are there complete outcome data?	Are the confounders accounted for in the design and analysis?	During the study period, did the exposure occur as intended?
Kiskis, 1991	Yes	Yes	Yes	Yes	No	Cannot tell
Kim & McLean, 1995	Yes	Yes	Yes	Yes	Yes	Yes
Olea et al., 2000	Yes	Yes	Yes	Yes	Cannot tell	Yes
Powers, 2001	Yes	Yes	Yes	Yes	No	No
Revuelta et al., 2003	Yes	Yes	Yes	Yes	Cannot tell	Yes
Fritts & Marszalek, 2010	Yes	Yes	Yes	Yes	No	Yes
Ortner et al., 2014	Yes	Yes	Yes	Yes	Yes	Yes
Arvey, 1990	Yes	Yes	Yes	Yes	Cannot tell	Yes
Ling et al., 2017	Yes	Yes	Yes	Yes	Yes	Yes
Martin & Lazendic, 2018	Yes	Yes	Yes	Yes	No	Yes
Mohd Ali et al., 2019	Yes	Yes	Yes	Yes	Cannot tell	Yes

2.3.3. Instruments

The included studies used different instruments to measure anxiety and motivation. Anxiety was measured by the following scales: State-Anxiety Scale (SAS) (Olea et al., 2000; Revuelta et al., 2003), Test Anxiety Inventory (TAI) (Powers, 2001), State-Trait Anxiety Inventory (STAI) (Ling et al., 2017), State-Trait Anxiety Inventory for Children (STAIC) (Fritts & Marszalek, 2010), The Friedben Test Anxiety Scale (FTA) (Mohd Ali et al., 2019), and Comparative Anxiety subscale of Test Attitude Survey (TAS) (Arvey, 1990). Motivation was measured by the Questionnaire on Current Motivation (QCM) (Ling et al., 2017; Ortner et al., 2014), the Short Motivation and Engagement Scale (Short MES) (Martin & Lazendic, 2018), and Motivation subscale of TAS (Arvey, 1990).

Some of the studies also reported subscales scores (Arvey, 1990; Ling et al., 2017; Martin & Lazendic, 2018; Mohd Ali et al., 2019; Ortner et al., 2014), and some of them reported multiple outcome measures (Ling et al., 2017; Kiskis, 1991). Although two studies (Ling et al., 2017; Ortner et al., 2014) used the QCM as a measure of motivation, they measured different factors; Ling and colleagues measured the ‘Challenge’ and interest’ factors and modified the scale to adjust the context of their research, while Ortner and colleagues measured the ‘Probability of success’ and ‘Fear of failure’ factors. TAI was also administered on one occasion (Fritts and Marszalek, 2010), but we excluded this study in our review since TAI measures trait anxiety (with items such as “I feel very panicky when I take an important test”), and it was administered before the achievement tests. In comparison, Powers (2001), Kim & McLean (1995), and Kiskis (1991) have modified the questionnaire TAI to measure test anxiety after taking a test. Some of the studies measured additional constructs, too. For example, Power (2001), Kiskis (1991), and Fritts and Marszalek (2010) measured computer anxiety. In our review, we only included measures of state anxiety and/or motivation.

2.3.4. Overall Effect of Test Type on Anxiety and Motivation: Meta-Analytical Results

As there were no outlier studies based on the standardized residuals, all 11 studies were included in the meta-analysis of the overall effect of test type on anxiety and motivation. A meta-regression analysis revealed that the year of publication among the included studies had no effect on the overall effect size (coefficient = -0.002, p = .78). The funnel plot showed a symmetrical distribution, which suggested no publication bias (see Figure A1). Similarly, Egger's regression test showed no signs of publication bias ($t = 0.51$, $p = .63$). Figure 7 shows the forest plot with a non-significant small effect of test type on overall anxiety and motivation. The overall effect was significantly heterogeneous, with a high proportion of observed variance (84%) reflecting real differences in effect size (see Figure 7).

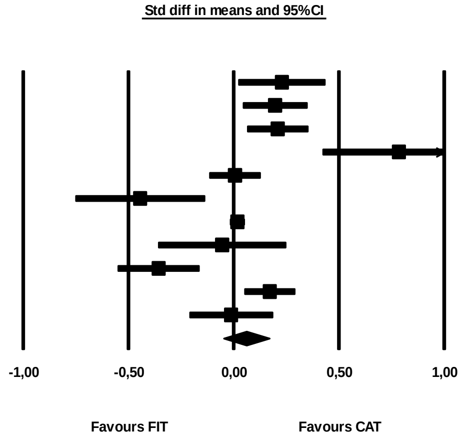
As one of the included studies (Martin & Lazendic, 2018) had a sample size of over 12.000 participants, its relative weight in the overall analysis was twice that of the weight of the smallest sample. For this reason, a sensitivity analysis was performed with the exclusion of the Martin and Lazendic (2018) study ($k = 10$, $g+ = .07$, SE = .08, 95% CI [−0.08, 0.21], $p = .37$), but still indicating a non-significant small sized effect.

Subgroup analyses of different comparisons of CAT, PPFIT, and CFIT were non-significant, except for ECAT's overall effect on motivation and anxiety in contrast to PPFIT and CFIT, showing a large positive effect (see Table 4).

Figure 7

Forest Plot of the Overall Effect of Test Type on Anxiety and Motivation

Study name	Subgroups within study	Outcome	Statistics for each study						
			Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Olea et al., 2000	Combined	SAS anxiety	0,229	0,105	0,011	0,023	0,435	2,175	0,030
Powers et al., 2001	PPFIT - CAT	TAI anxiety	0,197	0,078	0,006	0,044	0,350	2,531	0,011
Revuelta et al., 2003	Combined	SAS anxiety	0,210	0,074	0,005	0,064	0,355	2,828	0,005
Fritts & Marszalek, 2010	PPFIT - CAT	STAIC anxiety	0,786	0,185	0,034	0,423	1,149	4,243	0,000
Ling et al., 2017	Combined	Combined	0,006	0,062	0,004	-0,115	0,128	0,102	0,919
Ortner et al., 2014	CFIT - CAT	Combined	-0,444	0,157	0,025	-0,751	-0,137	-2,834	0,005
Martin & Lazendic, 2018	CFIT - CAT	Combined	0,018	0,018	0,000	-0,017	0,053	0,980	0,327
Kiskis, 1991	Combined	Combined	-0,055	0,155	0,024	-0,358	0,249	-0,352	0,725
Kim & McLean, 1995	PPFIT - CAT	TAI anxiety	-0,357	0,099	0,010	-0,550	-0,163	-3,607	0,000
Arvey, 1990	PPFIT - CAT	Combined	0,172	0,061	0,004	0,052	0,292	2,804	0,005
Mohd-Ali et al., 2019	Combined	Combined	-0,012	0,101	0,010	-0,210	0,186	-0,118	0,906
			0,060	0,056	0,003	-0,049	0,169	1,084	0,278



Note. This figure demonstrates a forest plot with the individual study effect sizes and the total effect size (Hedges' g) of test type on anxiety and motivation combined. Negative effect size favours the FIT groups (PPFIT and CFIT), and positive effect size favours the CAT groups (CAT and ECAT). The total effect is demonstrated in the last row.

Table 4

Effects of Testing Type on Anxiety and Motivation

	Effects based on standardized mean differences and heterogeneity								
	k	Mean effect size (g^+)	95% CI	*p*	SE	*Q* value	*p*	I^2	T^2
Overall	11	0.06	[-0.05; 0.17]	.28	0.06	61.46	.001	84%	0.02
Effect favours CAT to PPFIT & CFIT	11	0.04	[-0.09; 0.16]	.56	0.06	67.37	.001	85%	0.03
Effect favours CAT to PPFIT	6	0.11	[-0.14; 0.35]	.39	0.12	39.26	.001	87%	0.07
Effect favours CAT to CFIT	7	-0.02	[-0.16; 0.12]	.79	0.07	24.25	.001	75%	0.02
Effect favours ECAT to PPFIT & CFIT	2	0.22	[0.09; 0.36]	.001	0.07	0.08	.77	0%	0.01
Anxiety	9	0.09	[-0.06; 0.23]	.23	0.07	46.37	.001	83%	0.04
Effect favours CAT to PPFIT & CFIT	9	0.06	[−0.11; 0.23]	.49	0.09	57.43	.001	86%	0.06
Effect favours CAT to PPFIT	6	0.08	[−0.16; 0.31]	.52	0.12	36.82	.001	86%	0.07
Effect favours CAT to CFIT	5	0.02	[−0.22; 0.26]	.86	0.12	20.37	.001	80%	0.06
Effect favours ECAT to PPFIT & CFIT	2	0.22	[0.09; 0.35]	.001	0.07	0.05	.82	0%	0.01
Motivation	4	0.03	[-0.15; 0.21]	.75	0.09	31.67	.001	91%	0.03
Effect favours CAT to PPFIT & CFIT	4	-0.03	[-0.25; 0.19]	.78	0.11	36.48	.001	92%	0.04
Effect favours CAT to CFIT	3	-0.15	[-0.38; 0.07]	.18	0.12	12.28	.002	84%	0.03

Note. k = number of included studies; g^+ = Hedges' *g* effect size; 95% CI = 95% confidence interval; *p* = significance value; *Q* = Cohrane's Q value to test heterogeneity; I^2 = percentage of relative variance across studies due to heterogeneity; T^2 = absolute between-study variance; CAT = computerized adaptive testing; ECAT = Easier Computerized Adaptive Testing; PPFIT = Paper-and-Pencil Fixed Item Testing; CFIT = Computerized Fixed Item Testing

2.3.5. Effect of CAT on Anxiety

Four of the nine articles that discuss anxiety found significantly lower levels of reported anxiety when taking a CAT. Fritts and Marszalek (Fritts & Marszalek, 2010) compared state anxiety of junior high school students after taking a standardized achievement test. The result of the analysis showed that examinees who took a traditional test had a higher mean state anxiety score than examinees who took the CAT, after controlling for computer anxiety and test anxiety. Powers (Powers, 2001) also compared examinees' anxiety after they took the Graduate Record Examination (GRE) Test – albeit several days after actually taking the test – and found that the PBT sample reported higher anxiety levels than the CAT sample. The same effect of CAT on anxiety was also found by Ling and colleagues (Ling et al., 2017). However, they used two types of CAT: Easier CAT (ECAT) and regular CAT. The ECAT was a version of CAT in which items were chosen at a lower difficulty level than the examinee's estimated ability, thus increasing the probability of arriving at a correct answer from the 50% that is regularly applied in a CAT. They compared middle school students' state anxiety after taking mathematics problem-solving tests and found that ECAT resulted in lower anxiety than either regular CAT or CFIT.

Five of the nine studies did not find a statistically significant effect of test conditions on anxiety. The goal of the study by Olea and colleagues (2000) was to examine the effect of being able to review and change previous answers on computerized tests, both fixed and adaptive; they also compared participants' state anxiety before and after taking an English vocabulary test. A similar study was conducted by Revuelta and colleagues (Revuelta et al., 2003). Their main goal was to investigate the effect of item selection and the ability to review previous items on computerized testing. However, they also compared participants' state anxiety among three types of tests: CAT, ECAT, and CFIT. Arvey (1990) compared the anxiety of Armies after taking the CAT and FIT versions of The Armed Service Vocational Aptitude Battery (ASVAB), while Kiskis (1991) compared the anxiety of applicants at a personnel agency after taking the CAT and FIT version of clerical aptitude test.

2.3.6. Meta-Analytical Results: Anxiety

As shown in Table 4, a non-significant small effect of testing type on anxiety was found. The effect was heterogeneous, with 83% of the observed variance reflecting differences in effect size. Subgroup analyses of different comparisons of CAT, PPFIT, and CFIT were non-significant, except for ECAT's overall effect on anxiety in contrast to PPFIT and CFIT, indicating a large positive effect (see Table 4)

2.3.7. Effect of CAT on Motivation

Two of the four articles reported a positive effect of CAT on motivation (Ling et al., 2017). Arvey (1990) reported that the CAT version of the Armed Services Vocational Aptitude Battely (ASVAB) had significantly higher scores on the Motivation factors compared to the paper-and-pencil version of the ASVAB. In addition, Ling and colleagues (2017) compared three types of tests: ECAT, regular CAT, and CFIT and found that ECAT resulted in higher motivation than regular CAT or CFIT. However, they did not find any significant difference of motivation between regular CAT and CFIT.

Another study compared test-relevant motivation and engagement in elementary and secondary school students who completed a numeracy test and reported the lack of a statistically significant effect of test condition on motivation (Martin & Lazendic, 2018). Finally, one of the papers even reported a negative effect of CAT on motivation in secondary school students (Ortner et al., 2014). During a break in the testing session, state motivation was measured, and 'Fear of failure' was higher in the CAT condition than in the CFIT condition. Moreover, the 'probability of success' in the CAT condition was lower than in the CFIT condition. These results might explain why students found CAT more motivating than CFIT.

2.3.8. Meta-Analytical Results: Motivation

As shown in Table 4, there was a non-significant small effect of testing type on motivation. The effect was heterogeneous, with 91% of the observed variance reflecting differences in effect size. Subgroup analyses about different comparisons of CAT, PPFIT, and CFIT were non-significant (see Table 4), although ECAT type of testing was not compared to FIT types of tests as there was only one study measuring this.

As the Martin and Lazendic (2018) study, with a sample size of over 12.000 participants and a relative weight twice of the weight of the smallest study in the subgroup analysis, a sensitivity analysis was performed with the exclusion of this study (*k* = 3, *g*+ = .005, SE = .17, 95% CI [−0.32, 0.33], *p* = .98), but still indicating a non-significant small sized effect.

2.4. Discussion

This review examined the effect of CAT on motivation and anxiety in comparison to traditional FIT, based on eleven studies. The general result of our review and meta-analysis suggested no significant effect of test type on anxiety and motivation when comparing CAT with FIT. This is in contrast with the claims articulated in early work on CAT (Betz & Weiss, 1976; Wainer, 2000).

Only two studies on motivation and four studies on anxiety in our review supported the benefits of CAT, while one of them showed the opposite result: a decrease in motivation under CAT. It should be also noted that the single study which demonstrated a positive effect of CAT on motivation and anxiety (Ling et al., 2017) compared two types of CAT, easier CAT (ECAT) and regular CAT. They found that only ECAT, but not traditional CAT, resulted in higher motivation and lower anxiety than regular FIT.

It is possible that there are methodological reasons for the null findings. For example, in the study of Ortner and colleagues (Ortner et al., 2014), test-takers were not given specific information about how CAT works. That such information might be relevant is highlighted in an earlier study (Ortner & Caspers, 2011) that informing examinees about the mechanisms of adaptive testing led to higher scores than presenting standard instructions. Another possibility is that participants are uncomfortable with certain features in CAT, such as the inability to review or skip items (Tonidandel et al., 2002; Tonidandel & Quiñones, 2000). The difference between low-stakes vs. high-stakes testing situations could also affect motivation and anxiety. For example, Revuelta and colleagues noted that a lack of an effect of test type on anxiety may be due to the floor effect caused by the low-stakes nature of the test (Revuelta et al., 2003). On the other hand, in high-stakes testing (e.g., in the GRE test), Powers found that those who took PBT reported more anxiety than those who took CAT

(Powers, 2001). Uncontrolled confounders were also found in few studies, such as different school districts (Fritts & Marszalek, 2010) or other pre-existing differences (Powers, 2001). Some studies controlled some potential confounders (e.g., trait anxiety, computer anxiety, ability), while others did not.

In addition, several of the reviewed studies also discussed the different conditions of CAT that could affect motivation and anxiety. In our analysis, using ECAT had significant large effect on anxiety in comparison with FIT. It was in line with previous studies (Häusler & Sommer, 2008; Tonidandel et al., 2002) that found respondents' reactions to be more favorable under easier computerized adaptive tests. A possible explanation for this finding from the expectancy-value theory perspective is that using easier items can result in higher expectancy, as examinees are consistently given items below their ability level. This could lead to increased motivation and reduced anxiety overall. This characteristic is not present in regular CAT, where examinees typically start with medium difficulty items. This approach could lead to a low perceived success probability (Frey et al., 2009; Ortner et al., 2014), which is analogous with low expectancy. However, using easier items is not optimal from the perspective of measurement efficiency (B. A. Bergstrom et al., 1992; Häusler & Sommer, 2008). For example, it takes 100 items to reach a SEM of .20 if the probability of a correct response is 50%, 104 items if 60%, and 119 items if 70% (Bergstrom et al, 1992). However, the increase in test length did not lead to an increase in test duration (Hausler & Sommer, 2008).

Another condition that could lower examinees' level of state-anxiety is allowing them to review previously administered items and change their responses (Olea et al., 2000; Revuelta et al., 2003). However, from the perspective of test developers permitting item review is difficult, since the test algorithm has to be more complicated and testing time typically increases by 37%-61% (Vispoel et al., 2000).

Further, the specific procedures employed by the reviewed studies also provide valuable information about the psychological aspects of using CAT. For example, Olea and colleagues (Olea et al., 2000) suggested that providing detailed, item-level feedback on performance after the exam leads to decreased state anxiety and an increased ability estimate level. Future investigation in this topic is needed.

The ability level of the examinees might also mediate results. In our review, only three studies investigated the relationship between performance, testing mode, and psychological effects (Ling et al, 2017; Ortner et al., 2014; Powers, 2001). Ling and colleagues (2017) reported that examinees with higher abilities tended to report less anxiety and less engagement for each mode of testing (CAT, ECAT, and FIT). However, under the ECAT condition, lower-ability examinees reported less anxiety and more engagement than in regular CAT and FIT conditions. A similar result was found by Powers (2001): the relationship between performance and anxiety was similar for each mode of testing (CAT and FIT). Yet a different result was reported by Ortner and colleagues (2014): motivation was equal for high- and low-performance examinees in the CAT condition, but in the FIT condition, high-performance examinees experienced a higher motivation. Evidence for the interaction between ability and mode of testing is still inconclusive, and thus, future research in this area is required.

Specifically, in the study investigating constructs related to fluid reasoning, two studies were analyzed with contradictory findings. The first study, conducted by Powers (2001), examined verbal, quantitative, and analytical reasoning. The results were positive: examinees who took the CAT reported less anxiety than those who took the FIT. In contrast, the second study by Ortner et al. (2014), which focused on figural reasoning, found a negative effect: examinees who took the CAT reported a lower probability of success and greater fear of failure than those who took the FIT. These differing outcomes may be attributed to the distinct procedural contexts of the studies. Examinees in the first study were informed about the test format they were engaging with, in contrast to those in the second study, who were not provided with such information. Furthermore, the first study was set in a high-stakes environment where the test outcomes had substantial implications for the participants, unlike the second study. The inconsistent results of these two studies merit further exploration in the future.

Several studies could be relevant to this research, but their full texts are unavailable in English. For instance, Frey et al (2009) investigated the impact of adaptive testing using the Frankfurt Adaptive Concentration Test and found that test-taking motivation was significantly lower in the adaptive condition compared to the non-adaptive condition, largely due to perceived success probability. Meanwhile, Elbarbary (2020) demonstrated that variable-length adaptive tests were more effective in reducing test anxiety and enhancing positive attitudes towards online exams than both fixed-length adaptive tests and traditional linear computer tests, the latter two showing no significant difference.

This review has several limitations. First, it only considered studies that contained a comparison of motivation and/or anxiety between CAT and FIT, but not a comparison within CAT conditions, such as the ones carried out by Hausler and Sommer (Häusler & Sommer, 2008) as well as Toninandel and colleagues (Tonidandel et al., 2002), who compared different item selection methods and their impact on examinee's motivation.

Second, the number of studies included in the meta-analysis was small. However, as Davey and colleagues (2011) reported, the average meta-analysis in some fields includes a median of three studies. Third, our review only included English-language studies. Fourth, several of the reviewed studies did not control for possible confounder variables such as trait anxiety, computer anxiety, test-taker's ability, and testing context (low- and high-stakes).

Chapter 3: Development and Evaluation of Multidimensional Computerized Adaptive Test for Measuring Fluid Reasoning[5]

3.1. Background and aims

Although many Gf tests have been developed, there is a lack of figural tests measuring two narrow factors simultaneously. From a CHC perspective, to adequately represent a measure of Gf, it is essential to assess at least two narrow abilities (Flanagan et al., 2013; Schneider & McGrew, 2018). In addition, there is a need for flexible, accessible, efficient, and comprehensive tests measuring Gf for research purposes. For the reasons mentioned above, multidimensional CAT can be the solution. MCAT has been considered more efficient and beneficial than separately administered unidimensional CAT or fixed-item tests (Chien & Wang, 2017). Although Gf could be approached from different perspectives, we focus on the CHC model for this study. This model is highly influential in contemporary psychometric testing and is familiar to most users of such tests (Flanagan & Dixon, 2014). Using the CHC model as a guiding framework allows us to position our tests in an accepted and well-known taxonomy of cognitive abilities and thus facilitates the interpretation of test results.

The current study aimed to develop and evaluate a Multidimensional Induction-Deduction Computerized Adaptive Test (MID-CAT), a test that measures two process factors of Gf: induction and deduction. Induction, or rule inference, is the ability to observe a phenomenon and discover the underlying principles. Deduction, or rule application, pertains to logical reasoning based on established rules and premises. Furthermore, figural content was used to reduce bias due to fluency in a language.

The tests consisted of two tasks – the odd-one-out and sudoku-like tasks – measuring two narrow abilities of Gf: induction and deduction. To solve the-odd-one-out task, test-takers need to identify the similarities between the figures in order to find the odd one. That is, examinees need to *infer a new rule* to arrive at a solution. This means they engage in inductive reasoning when solving these items. The primary rationale for choosing the odd-one-out task over other forms like matrices, series, or analogies is to avoid redundancy, given

[5] Chapter 3 is based upon the following paper:
Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis. Assessment, 30(5), 1379–1390. https://doi.org/10.1177/10731911221100995

the established tests for these types in academic research (e.g., Condon & Revelle, 2014; Koch et al., 2022; Kyllonen et al., 2019). Despite all these tasks assessing inductive reasoning, the odd-one-out, being less common yet potentially complementary, offers a unique approach that could further elucidate the distinct cognitive strategies employed in inductive reasoning. Additionally, the sudoku-like task is a relatively new invention as a figural psychometric test of deductive reasoning, contrasting with the majority of deductive reasoning tasks, which are typically verbal. To solve sudoku-like task, test-takers are already provided with the rule that is required to solve the task, and examinees need to *apply this rule* to arrive at a solution. This means they engage in deductive reasoning when solving these items.

This chapter is divided into two parts. The first one discusses the development of the item bank using the multidimensional Rasch model. The second part discusses the results of a simulation study that evaluates the potential performance of MCAT compared with separate-unidimensional CAT (UCAT) or FIT.

3.2. Study 1: Development of the item bank

The purpose of Study 1 was to create fluid reasoning items and investigate the psychometric properties of the item pool. The main issues we wanted to address in study 1 were (a) whether the Gf construct fits better in a unidimensional, separate-unidimensional, or multidimensional model; (b) whether we could generate a Gf test that has a wide item difficulty range; and (c) whether the measures are valid indicators of Gf as shown by correlations with external measure. Study 1 was divided into two stages (study 1a and 1b). Study 1a was a pilot study aimed at creating an initial item pool and checking the appropriateness of the use of a multidimensional model as well as the item difficulties. Study 1b aimed to exclude items with poor psychometrics characteristics, introduce a set of new items, and validate the tests.

3.3. Method

3.3.1. Participants

The total number of participants in Study 1 was 2247. Data were collected in two waves (study 1a and 1b). For study 1a, 206 participants (148 females) completed the tests.

Participants were undergraduate students in the Faculty of Psychology, University of Muhammadiyah Malang (M_{age} = 19.87, SD_{age} = 0.74, range = 18 - 22). Most participants live in urban areas (53%), followed by those in rural (31%) and suburban (16%) areas.

For study 1b, the participants were 2041 Indonesians (1258 female) with M_{age} = 23.99, SD_{age} = 7.49, range = 14 - 59. Most participants either hold or are pursuing a bachelor's degree (64%), followed by those with a senior high school diploma or who are currently attending high school (25%). Those either holding or pursuing master's (9%) and doctoral degrees (2%) make up the remainder. Most participants live in urban areas (50%), followed by those in rural (30%) and suburban (20%) areas. Participants were recruited in May-July 2022 using various strategies, including advertisements on social media (Instagram, Twitter, Facebook, and WhatsApp groups), as well as invitations extended to teachers, lecturers and their students. No monetary incentives were offered for participating in this study. All participants received the result at the end of the test but were also notified that the test was still under development.

3.3.2. Measures

For study 1a, we wrote an initial 50 items for each test, varying in expected difficulty. We administered all items after item reviews by the authors as well as an expert in cognitive and cross-cultural psychology, and through cognitive interviews with research assistants (i.e., we showed items to research assistants and asked them to think aloud to ensure the proper understanding of the task). All the tasks in Study 1 were programmed for the PsyToolkit platform (Stoet, 2010, 2017), along with demographic questions.

The odd-one-out task

The odd-one-out tasks (hereinafter called “induction test”) were developed to measure inductive reasoning. Induction test items were created by varying (a) the type of stimulus that appeared in the picture (shape, colour, position, number, size), (b) the number of stimulus types that appeared in the picture, and (c) the principal relationship among stimuli. More difficult items had more variation in stimulus and a more complex principal relationship (see Figure 8). There are six pictures in an item, and one out of the six pictures has the most different characteristics based on a certain principle. Examinees were asked to find the picture most different from the others. To solve this task, the examinees need to identify the similarities between the figures in order to find the odd one. That is, examinees need to infer a new rule in order to arrive at a solution. This means they engage in inductive reasoning when solving these items.

Sudoku-like task

Sudoku-like task (hf. deduction test) was developed to measure general sequential (deductive) reasoning. The deduction test was a modified version of the classic 6x6 Sudoku puzzle by changing the stimulus from number to shape. Items were created by varying the relational complexity among stimuli, i.e., the number of constraints on which they depend (for further detail, see Lee et al., 2008). Examinees were asked to replace the question mark, so there was only one of each shape in any column, row, or mini-grid. The more difficult items had more relational complexity (see Figure 8). To solve this task, examinees are required to identify the missing shape using the general rule that each figure appears only once in each column, row, and mini-grid. That is, examinees are already provided with the rule that is required to solve the task, and examinees need to apply this rule in order to arrive at a solution. This means they engage in deductive reasoning when solving these items

Figure 8

Sample items of the induction (A) and deduction (B) test.

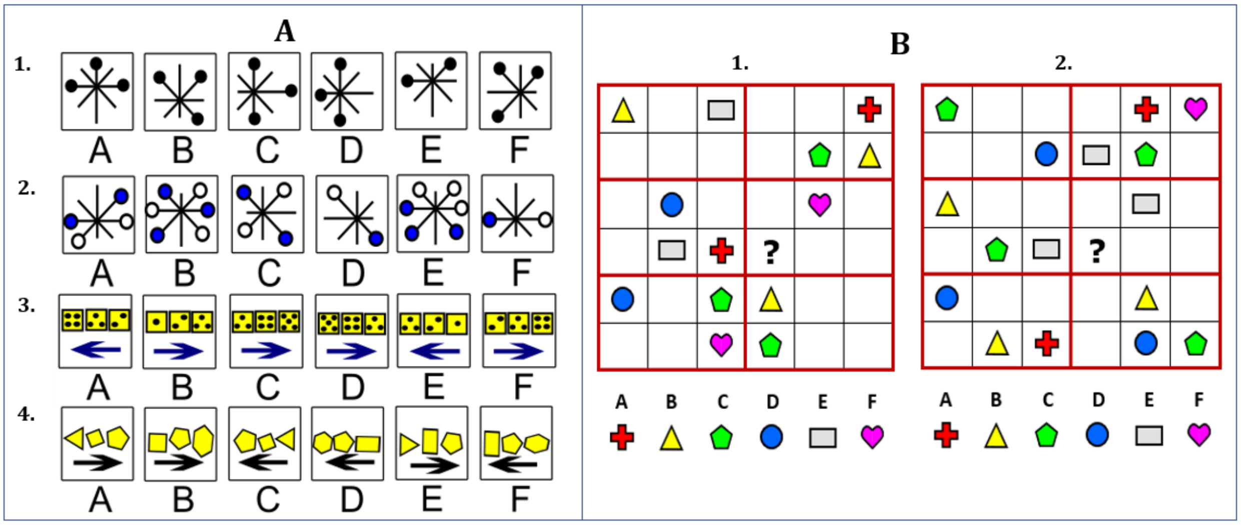


Figure 8 depicts sample items of the tests. Item A1 is easier than A2 because it has fewer stimuli. A1 only varies in the number and position of the dots, while item A2 also varies in colour. Item B1 is easier than B2 because it provides more information based on the premises. For item B1, after excluding the shapes in the same column, row, and mini-grid as the question mark, the only possible answer is D. However, for item B2, after excluding the shapes in the same column, row, and mini-grid as the question mark, there are still four possible solutions. Therefore, test-takers need to use the inclusion-exclusion strategy and solve other cells first until they can determine the answer.

Following Study 1a, for Study 1b, new items were written. Based on the evaluation of study 1a, we wrote additional items, creating an item pool of 530 items (265 for both induction and deduction tests). An additional measure was administered to investigate the validity of the tests: Hagen Matrices Test – Short form (HMT-S, Heydasch et al., 2013). HMT-S is a six-item matrix test intended to measure fluid reasoning. The previous investigation found that the reliability of HMT-S was 0.60 (Heydasch et al., 2013). Matrices tests, such as HMT-S, require individuals to make accurate generalizations based on observed patterns, which is a key component of inductive reasoning. Therefore, the high correlation with HMT-S indicates that the test measures Gf.

3.3.3. Procedure

Data collection was divided into two waves (study 1a and 1b). At the beginning of all studies, participants were informed about the goal of the research and about technical details to complete the test. For study 1a, participants who were willing to participate in the study completed the demographic questions and the 100 items of the tests (50 items of each task). The test administration was self-paced. The testing was carried out on the Psytoolkit platform (Stoet, 2010, 2017). Participants were allowed to take breaks between each subtest but not between items of either subtest. Participants used their own devices (PC or laptop) to complete the tests. Respondents were given instructions for completing the test, two exercise items, and an explanation of the solution of the exercise items. It took participants about 20-40 minutes to complete each subtest.

For study 1b, we developed 13 forms of the combined induction-deduction test, with a balanced level of item difficulty in each form. Each form contained 50 items, divided into two subtests with 25 items for induction and 25 for deduction, and included five anchor items in each subtest (i.e., items presented in all 13 forms). Anchor items were chosen based on study 1a to be proportionally representative in content as well as item difficulties. Anchor items provided the statistical adjustment needed to include multiple test forms on a common metric scale.

In both studies tests were administered in an unproctored online environment; participants used their own devices (PCs or laptops) to complete the test. All participants received a report of their scores at the end of the test. The report contained a warning that the test was under development. Response times were recorded. Research Ethics Committee of Eotvos Lorand University, Hungary, approved the research protocol for both studies (license numbers for study 1a and 1b are 2021/54 and 2022/291, respectively).

The data from non-effortful test-takers were excluded as they might have negatively impacted estimates of item parameters (Rios & Soland, 2021; Wise & DeMars, 2006). We excluded participants with a Response Time Effort (RTE; see Wise & Kong, 2005) of less than 0.8, as recommended by Rios and colleagues (2017). We used the 10% Normative Threshold (NT10) approach to determine the threshold of rapid guessing response. This approach proposes using 10% of the average response time for the threshold. If a participant responds slower than the threshold, their response is considered appropriate solution behaviour (SB). RTE was calculated by summing the SB index values across all items and dividing by the number of items in the test (see Wise & Ma, 2012 for detailed procedures to calculate RTE).

3.3.4. Analysis

All analyses were performed in R software (R Core Team, 2012). Data were analyzed using the dichotomous Rasch model (Rasch, 1960). The multidimensional random coefficients multinomial logit model (MRCMLM; Adams et al., 1997) was used to estimate both unidimensional and multidimensional Rasch models. MRCMLM is part of the Rasch family, which has good measurement properties, such as specific objectivity and sufficient statistics.

The first analysis compared three models: unidimensional, separate-unidimensional, and multidimensional. The Bayesian information criterion (BIC; Schwarz, 1978), The

Akaike information criterion (AIC; Akaike, 1974), and the *p*-value of the likelihood ratio (LR) test were used to evaluate model-data fit. The model with the lowest BIC and AIC values was considered the best.

To analyze item parameters, we calculated item difficulty (*p*) and item-total correlations (r_{it}) in the sense of classical test theory for each subtest separately. Items with negative r_{it} were removed. Parameters and item fit were estimated using MML estimation in the 'TAM' packages (Robitzsch et al., 2022). Data visualization was prepared using ‘WrightMap’ (Irribarra & Freund, 2014) and ‘mirt’ packages (Chalmers, 2012). A fitted ‘TAM’ object was converted into a ‘mirt’ object using ‘sirt’ package (Robitzsch, 2023). For MRCMLM, item fit was assessed using the residual-based approach (i.e., infit and outfit mean square) suggested by Adams and Wu (2007). Misfit items (i.e., infit or outfit < 0.5, or infit or outfit > 1.5; Wright & Linacre, 1994) were removed from the item bank. We estimated the item difficulty (*b*) for all items fitting the Multidimensional Rasch model. Items in all forms were calibrated using concurrent calibration. The final theta of each dimension was then correlated with theta scores of HMT-S to investigate the convergent validity of the test.

3.4. Results

3.4.1. Initial response screening

The data collection platform did not allow missing item responses, except for timing out due to item time limits (120 seconds). Fewer than 1% of responses were in this category and were treated as incorrect (score 0). Initial response screening was investigated with RTE. RTE values near 1 indicate a strong examinee effort for the test. In our analysis, we filtered out participants with an RTE value of less than 0.8. For studies 1a and 1b, only 193 and 1757 participants were used to estimate item parameters, respectively. For Rasch-type models, a sample size of 150 is usually sufficient (Sahin & Anil, 2017), and no extra sample sizes are required for the multidimensional approach (Wang et al., 2004).

3.4.2. Model comparison

Model comparison was conducted to investigate which model fits the data better: unidimensional, separate-unidimensional, or multidimensional. Model comparison in Table 5 shows that the multidimensional model has the lowest AIC and BIC values, indicating that this model fits the data better than the unidimensional and separate-unidimensional models. Similarly, the LR test also showed that the multidimensional model is a significantly better fit than both the unidimensional and separate-unidimensional models. Given the advantages of the multidimensional model, analyses were based on the multidimensional model.

Table 5

Comparison of unidimensional, two-unidimensional, and multidimensional model.

Model	AIC	BIC	logLik	Npars	LR Test
Unidimensional (1)	20714.14	21043.67	-10256.07	101	(1 vs 2), p < 0.001
Separate-unidimensional (2)	20596.20	209228.99	-10196.10	102	(1 vs 3), p < 0.001
Multidimensional (3)	**20538.01**	**20874.06**	-10166.00	103	(2 vs 3), p < 0.001

Note: AIC = Akaike information criterion, BIC = Bayesian information criterion, logLik = Log-likelihood, Npars = number of parameters, LR Test = Likelihood ratio test

3.4.3. Multidimensional Rasch analysis (study 1a)

Prior to Rasch analyses, we calculated item difficulty and r_{it} using classical test theory for each test. Out of the 50 items per test, participants answered on average 26.13 items (*SD* = 5.87) correctly for the induction test, and 20.76 items (*SD* = 8.99) for the deduction test. Overall, the developed items were of medium difficulty for the induction test (*M* = 0.52, SD = 0.29) and deduction test (*M* = 0.42, *SD* = 0.16). The average r_{it} for the induction test was *M* = .30, *SD* = 0.13, and the deduction test was *M* = .40, *SD* = 0.12. Four items with negative r_{it} were removed for the following analyses.

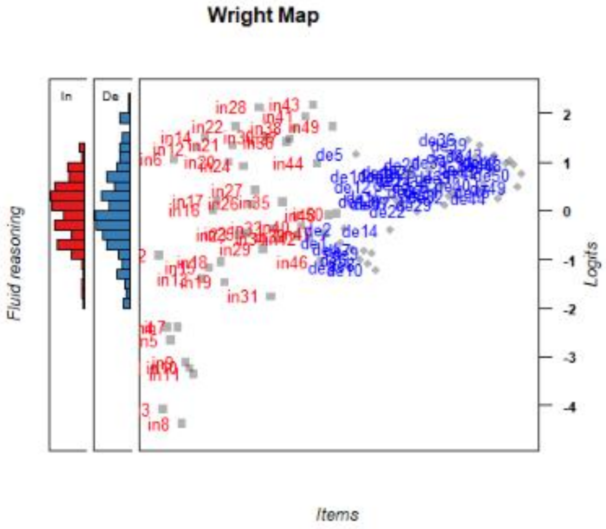
Multidimensional Rasch analysis was conducted to investigate item fit (infit and outfit), item difficulty (*b*), person ability (theta), and reliability. One item did not fit the Rasch model (i.e., outfit > 1.5) and was removed from the pool. Data were re-analyzed, and all items fit the Rasch model. The mean of infit was 0.98 (*SD* = 0.07), and the mean of the outfit was 0.97 (*SD* = 0.19). The mean of *b* for the induction test was -0.37 (*SD* = 1.76), and for the deduction test was 0.34 (*SD* = 0.70). The empirical reliability[6] for the induction test was 0.80, and the deduction test was 0.88.

[6] Empirical reliability is defined as $1 - s/(s + v) = v/(s + v)$, where $\boldsymbol{v}$ denote the variance of theta estimates, and $\boldsymbol{s}$ denotes the average of the squared standard error

The Wright map of initial items is shown in Figure 9. The histogram on the left side of the Wright map represents the distribution of person measure, while the item label on the right side of the Wright map represents the distribution of item difficulty. Person measures and item difficulty levels were plotted on the same scale. A positive value in the Wright map indicates more difficult items. Figure 9 shows that items intended to measure deductive reasoning were only appropriate for average-ability test-takers, while high- and low-ability test-takers were not measured with sufficient accuracy with the existing items. Therefore, more items were added.

Figure 9

Wright map of initial items of inductive and deductive reasoning tests



Note: In = Induction, De = Deduction

3.4.4. Multidimensional Rasch analysis (study 1b)

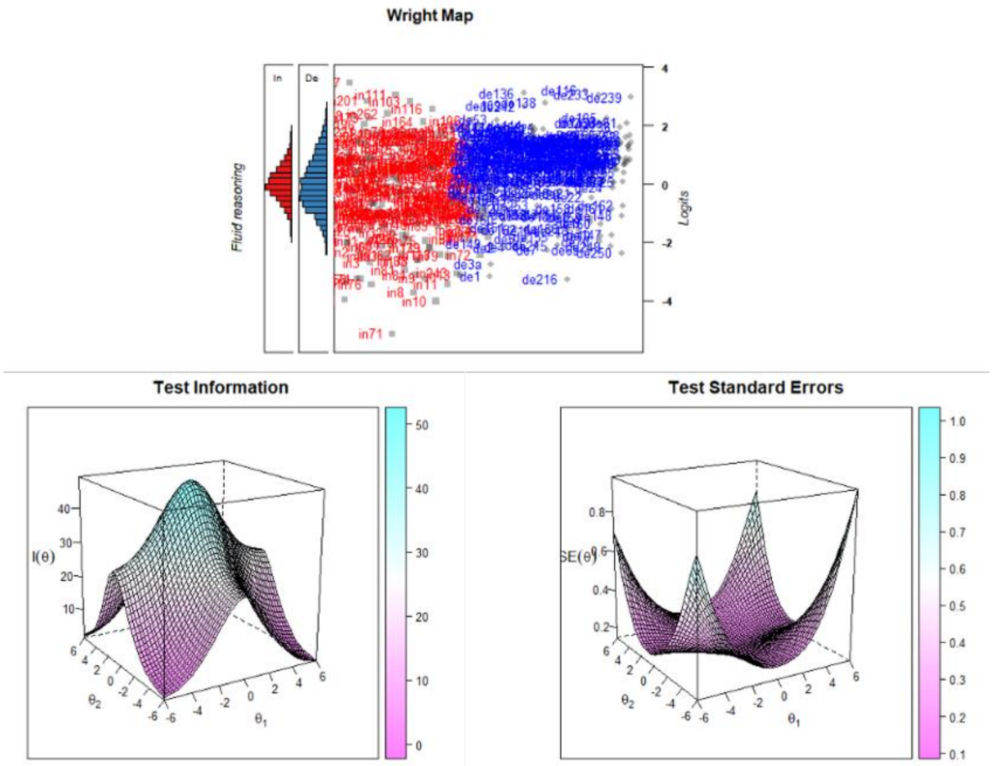
Following the recommendation from Study 1a, additional items were created for Study 1b. Prior to Rasch analyses, we calculated item difficulty and r_{it} in the sense of classical test theory for each test form. On average, the items in all forms were answered correctly by 47% of the participants. Out of the 25 items per test form, participants answered correctly on average 13.43 items (*SD* = 3.62) for the induction test and 9.95 items (*SD* = 4.47) for the deduction test. Overall, items difficulty (*p*) in the pool were medium for the induction test (*M* = .52, *SD* = 0.26) and deduction test (*M* = .39, *SD* = 0.20). Three items with negative r_{it}

were removed for the following analyses. The average r_{it} for the induction test was M = .34, SD = 0.13, and the deduction test was M = .39, SD = 0.13.

Multidimensional Rasch analysis showed that 11 items did not fit the Rasch model and were excluded. The mean of infit was 1.00 (SD = 0.09), and the mean of outfit was 1.02 (SD = 0.19). The mean of b for the induction test was -0.21 (SD = 1.47), and the deduction test was 0.56 (SD = 1.16). The empirical reliability for the induction test was 0.73, and the deduction test was 0.81. The Wright map, test information function, and standard errors of the final items are shown in Figure 10. The Wright map shows that the final items of the two tests have a wide range of difficulty that makes it possible to precisely measure participants with a wide range of abilities. Similarly, the test information and standard errors align with the Wright Map, indicating that all items in the bank could precisely measure a wide range of ability, particularly for examinees with average ability. Even for examinees with extreme ability (e.g., θ = -2.0 or θ = 2.0), the standard error remains below 0.3.

Figure 10

Wright map, test information function, and test standard errors of the final items



Note: In = Induction, De = Deduction, θ1=induction, θ2=deduction

3.4.5. Relation with external measure

We computed Pearson correlations to examine the correlation among the induction test, deduction test, and HMT-S. The correlation between the induction and deduction scores with HMT-S was *r* = .51 and *r* = .46, respectively. All tests correlated moderately with the HMT-S, indicating convergent validity and supporting the tests developed here as measures of Gf. However, these correlations were lower than expected. The few items and low reliability of HMT-S ($r_{xx'}$ = .53) possibly caused the correlation to be lower. After correcting for unreliability of measurement, the corrected correlation between the induction and deduction tests with HMT-S was *r* = .81 and *r* = .70, respectively. The factor correlation between the induction and deduction tests was *r* = .72. The corrected correlation is the raw correlation between x and y divided by the square root of the product of the reliability of x and the reliability of y.

3.5. Discussion

The main goal of Study 1 was to develop an item bank of fluid reasoning tests based on the multidimensional Rasch model. The final item bank consisted of 516 items (261 items measuring induction, 255 items measuring deduction). Overall, the proportion of correct answers for the deduction test was lower than for the induction test, indicating that the deduction test was slightly more difficult. This finding was further corroborated by the average *b* parameter -0.21 (induction) and 0.55 (deduction). The distribution of item difficulty and person theta in the Wright map (Figure 10) indicates that the items of the induction and deduction tests cover a wide range of difficulty. The items are appropriate to precisely measure the wide range of test-takers' abilities. The tests also showed satisfactory convergent validity with HTM-S.

3.6. Study 2: A simulation study of MCAT

The purpose of study 2 was to conduct Monte-Carlo simulations to evaluate the potential performance of MCAT in comparison with separate-unidimensional CAT or non-CAT. The main issue we want to address in Study 2 was to determine (a) whether MCAT was more efficient compared to UCAT or FIT, and (b) the number of items needed to be administered in high-stakes and low-stakes testing.

3.7. Method

The final item bank used in the simulation study contained 516 items measuring two latent traits (261 items measuring induction, 255 items measuring deduction). All steps in the simulation study were performed using the mirtCAT package (Chalmers, 2016) in R. Item parameters were based on the calibration results in study 1. Person theta scores were generated using the mirtCAT package. The theta parameters were drawn from a standard multivariate normal distribution (*M*=0, *SD*=1) with an inter-factor correlation of *r* = 0.72, and the sample size was fixed to 1000. Since we have two latent traits, items were divided into two blocks: induction and deduction. First, items were administered from the 'induction block'. Items from the 'deduction block' were only presented after the stopping criteria for the first block had been met. This is often called a multi-unidimensional model (Sheng & Wikle, 2014), where unidimensional blocks are clustered together for smoother presentation. In this condition, item selection was constrained to avoid intermixing items from different dimensions. Items selected for the first block were constrained to items that measure induction. Similarly, items for the second block were constrained to items that measure deduction. The MCAT simulation design differed in terms of test type and stopping rule.

3.7.1. Test type

There are two conditions of test type: CAT and FIT. Each test type has two conditions of the model: separate-unidimensional and multidimensional. MCAT refers to multidimensional CAT, UCAT refers to separate-unidimensional CAT, MFIT refers to multidimensional FIT, and UFIT refers to separate-unidimensional FIT. For the CAT, Kullback-Leibler Information Criteria (KL) was used for the item selection method. KL was introduced by Chang and Ying (1996), and Veldkamp and van der Linden (2002) adapted KL information for item selection in multidimensional adaptive testing. This method offers a potential advantage compared to Fisher Information (FI) methods because it can account for uncertainty associated with the $\hat{\theta}$ values when only a small number of items have been administered (Chang & Ying, 1996). The KL is appealing because it can be used for both unidimensional and multidimensional adaptive testing (Wang et al., 2013). In the case of shadow CAT design (i.e., CAT with several constraints), it was shown that it is feasible if item selection is based on the KL rather than the FI measure (Veldkamp & van der Linden, 2002).

The FIT version of the test was developed specifically for this study as a benchmark. The test was assembled using Automated Test Assembly performed using ‘xxIRT’ package (Luo, 2016). The induction and deduction test consisted of 20 items each. Twenty items are typically sufficient for low-stakes testing to reach a reliability of at least 0.80 for most test-takers (see Bergstrom et al., 1992). In order to maximize the reliability of most test-takers, the absolute objective of the test assembled was to have mean $b = 0$ and $SD = 1$, and the relative objective was to select items with higher r_{it}. All test-takers were administered the same items in the same sequence: the easiest to the hardest.

3.7.2. Stopping rule

As a stopping rule, the precision-based termination rules were utilized with three conditions: SE < 0.32 (equivalent of a reliability of 0.90[7]), SE < 0.45 (equivalent of a reliability of 0.80), and SE < 0.54 (equivalent of a reliability of 0.70). The fixed number of items was also simulated under four conditions: k = 40, k = 30, k = 20, and k = 10. Only a fixed number of items (i.e., k = 20) were performed for FIT. For ability estimation, Bayesian Maximum A Posteriori (MAP) was used.

Finally, all 16 conditions were tested to evaluate the performance of the MCAT in comparison to UCAT or FITs. Five criteria were used to evaluate the MCAT: test length, reliability, bias, root means square error (RMSE), and correlation between estimated and true theta (rxt).

1. *The test length* was simply the number of items the MCAT required to terminate. It was important for precision-based stopping rule conditions and was a measure of the efficiency of the MCAT.

[7] In classical test theory, the standard error of measurement (SE) is approximated with the equation SE = SD (1− rxx)½, where SD is the standard deviation of the observed scores, and rxx is the reliability. Assuming that the SD of theta is 1, specifying a reliability of .90 for rxx gives a SE of .32.

2. *Reliability* is equal to the mean reliability under each participant's stopping rule (Wainer, 2000). Reliability is defined as:

$$Reliability \; = \; 1 - SE^2$$

3. *Bias* measured the signed difference between the estimated and true theta. It was calculated by

$$Bias = \frac{\sum_{j=1}^{N}(\widehat{\theta_J} - \theta_j)}{N}$$

4. *RMSE* measured the absolute difference between the test-estimated and true theta. It was calculated by

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\widehat{\theta_J} - \theta_j)}{N}}$$

5. *The correlation between estimated and true theta* (r_{xt}) is the Pearson correlation between the test-estimated theta values and the true theta values.

3.8. Results

The complete findings of the simulation study are shown in Table 6. As shown in Table 6, MCAT outperformed both UCAT and FIT in all criteria. However, the efficiency of the MCAT varied depending on the stopping rule. When the precision-based stopping rule was applied, the test length of the MCAT was shorter than UCAT. Based on the average total items used, MCAT was 5-14% shorter than UCAT. The benefits of MCAT over FIT was also varied for different test-takers with different ability levels. For example, as shown in Figure 11, when the test length was fixed at 20 items, MCAT resulted in lower SEs than MFIT, especially for test-takers with very high or very low theta scores (i.e., theta < -2 or > 2).

Table 6

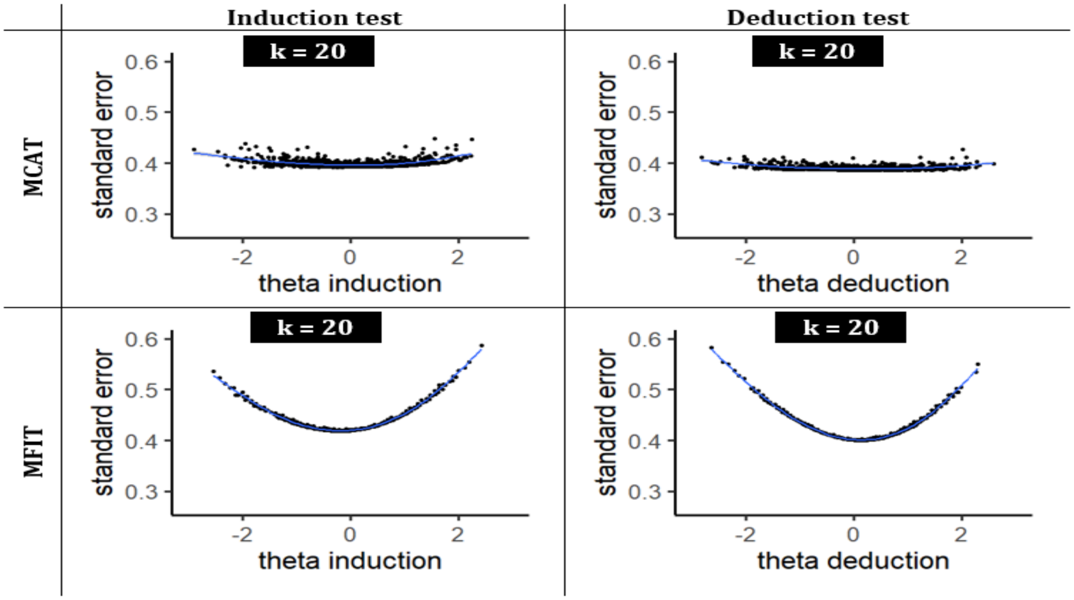
Results of the simulation study

Test type	Stopping rule	Test length		Reliability		Bias		RMSE		r_{xt}	
		In	De	In	De	In	De	In	De	In	De
MCAT	SE < 0.32	37.9	33.3	0.91	0.9	0.01	0.01	0.32	0.31	0.95	0.95
	SE < 0.45	17.93	14	0.83	0.8	-0.02	0.01	0.44	0.45	0.9	0.89
	SE < 0.54	11.56	8.08	0.75	0.72	-0.03	0.01	0.52	0.54	0.86	0.83
	k = 40	40	40	0.91	0.91	-0.01	0.01	0.31	0.3	0.95	0.95
	k = 30	30	30	0.89	0.89	-0.01	0.01	0.35	0.33	0.94	0.94
	k = 20	20	20	0.84	0.85	-0.01	0.01	0.41	0.39	0.91	0.92
	k = 10	10	10	0.73	0.75	-0.02	-0.02	0.53	0.49	0.85	0.86
UCAT	SE < 0.32	37.97	37.42	0.9	0.9	-0.01	-0.01	0.32	0.31	0.95	0.95
	SE < 0.45	17.98	17.11	0.8	0.8	-0.04	0.03	0.45	0.44	0.9	0.89
	SE < 0.54	11.6	10.72	0.72	0.72	0.01	0.01	0.55	0.53	0.84	0.84
	k = 40	40	40	0.9	0.91	-0.01	-0.01	0.31	0.3	0.95	0.95
	k = 30	30	30	0.87	0.88	-0.01	0.01	0.36	0.34	0.94	0.94
	k = 20	20	20	0.82	0.83	-0.01	0.01	0.43	0.41	0.91	0.91
	k = 10	10	10	0.68	0.7	-0.01	-0.01	0.58	0.54	0.82	0.83
MFIT	k = 20	20	20	0.81	0.82	0.01	0.02	0.44	0.41	0.9	0.91
UFIT	k = 20	20	20	0.78	0.8	-0.02	0.01	0.49	0.47	0.88	0.88

Note: KL = Kullback-Leibler Information Criteria, SE = standard error of estimate, k = number of items per test, F1 = induction, F2 = deduction, RMSE = root means square error, r_{xt} = correlation between estimated and true theta.

Figure 11

The number of items answered and standard error as a function test-takers theta

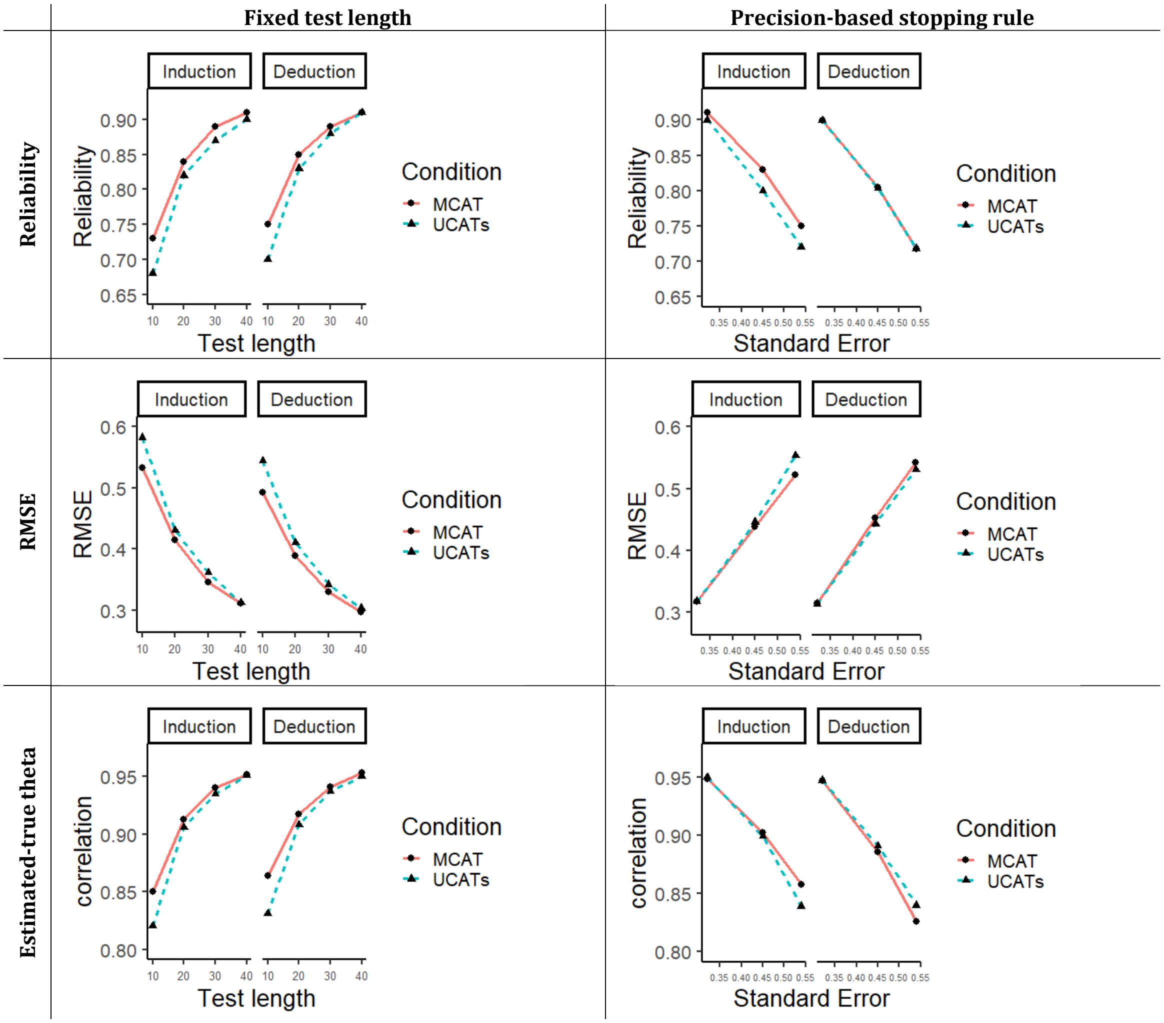


The efficiency of the MCAT in this study was more salient in the second task (deduction). The test in our study was designed based on a between-item model. Therefore, when the precision-based stopping rule was applied, the efficiency (i.e., shorter test length) of using the MIRT model was found only after completing the first block. However, in the separate-unidimensional model, the number of items used in the two subtests was relatively equal. At the same time, the induction test outperformed the deductive one in other aspects (i.e., reliability, RMSE, and r_{xt}) when the multidimensional model was applied. As shown in Figure 12, the induction test has higher reliability and r_{xt}, while the RMSE was lower than in the case of the deduction test. The difference increased as precision decreased (i.e., as SE increased). This is not the case in the unidimensional model, where reliability, r_{xt}, and RMSE values are relatively equal.

Finally, when test length was fixed – i.e. termination did not depend on accuracy – the benefits of MCAT could be demonstrated in both the induction and deduction tests. Across all conditions in the simulation, MCAT has lower SE and RMSE, while reliability and rxt were higher compared to two-unidimensional CAT or FIT. In all conditions, absolute bias was very low (less than 0.05), indicating no systematic error in ability estimation. A one-way ANOVA was conducted to examine the effect of different test types on estimated theta. The analysis revealed that test type was not significantly associated with estimated theta, $F(15, 15984) = 0.07$, $p = 1.00$, $\eta^2 < 0.001$ for induction, and $F(15, 15984) = 0.079$, $p = 1.00$, $\eta^2 < 0.001$ for deduction (see appendices for illustration, Figure A2).

Figure 12

Simulation results with different stopping rules



3.9. Discussion

The purpose of study 2 was to compare the performance of MCAT with unidimensional CAT or FIT. Our findings show that MCAT outperformed UCAT and FIT, but this is also the function of whether the test is fixed length or a precision-based stopping rule is applied. When the precision-based stopping rule is used, test efficiency (i.e., shorter test length) is only applied for the deduction test. The induction test has higher reliability and r_{xt} and a lower RMSE, but, in consequence, it is substantially longer than the deduction test. When fixed test length is used, all criteria are relatively equal between the induction and deduction tests. It should be noted that our finding is relevant only for the test that does not allow intermixing items between dimensions. The first test equals unidimensional CAT, but for the subsequent test, the estimation of the provisional ability benefits from multidimensional IRT. Without such constraint, item order is solely determined by item selection criteria, resulting in intermixed administrations of items from various dimensions, which could lead to different measurement efficiency (see Kroehne et al., 2014). In general, the multidimensional model outperforms the separate-unidimensional model, regardless of the item selection procedure (adaptive or not).

When compared to FIT, the main advantage of CAT is that it consistently produce relatively stable SE for all levels of test-takers' abilities. In contrast, FIT is only optimal for measuring test-takers with average abilities. Please note that FIT in this study was designed to have a mean M = 0 and SD = 1. Figure 11 illustrates that for test-takers with average abilities (i.e., -1 < theta < 1), the precision levels of MCAT and MFIT are not significantly different. However, MCAT demonstrates higher precision than MFIT for test-takers with low- or high- abilities (theta < -1 or theta > 1).

What are the consequences of the simulation for real-time testing? As the required accuracy of the test result is the function of the stakes of the testing session, there is no universal recommendation. Instead, the stopping rule criterion of SE < 0.32 (corresponds to a reliability of 0.90) is recommended for high-stakes assessments, while SE < 0.45 (corresponds to a reliability of 0.80) can be used for lower-stakes assessments. However, using a precision-based stopping rule resulted in a different number of administered items in the two tests. Alternatively, specifications can be provided in terms of the number of items: for high-stakes assessment, ideally, 40 items per subtest are administered, while for lower-stakes assessment, administering 20 items per subtest is sufficient.

3.10. General discussion of Chapter 3

The main purpose of this study was to develop a computerized adaptive test, MID-CAT, that measures two process factors of Gf: induction and deduction. The test was designed to be a flexible and efficient instrument that is entirely free for non-commercial use.

The output of study 1 was an item bank consisting of 516 items with a wide range of difficulty calibrated using the Rasch model. The validity of the test as a measure of Gf was demonstrated by its high correlation with HTM-S. In study 2, we employed simulations to compare MCAT with UCAT and non-CAT, and found that MCAT provides greater measurement efficiency: greater precision for fixed test lengths or shorter test lengths for precision-based stopping rule.

One notable aspect of the MID-CAT is its unique task format to measure fluid reasoning. The odd-one-out task has been widely used previously as a measure of inductive reasoning, as it requires individuals to identify the picture that does not fit with the others based on a set of rules or patterns (Ruiz, 2009). On the other hand, the Sudoku-like task is a relatively new invention as a psychometric test of deductive reasoning. Sudoku is a popular logic puzzle where individuals must infer the missing digits in a 9 × 9 array according to the general rule. Lee et al. (2008) suggested that individuals rely solely on pure deductions to solve Sudoku. Unlike traditional Sudoku puzzles, in our task test-takers only need to find the missing value in the missing cell. This kind of task makes it possible to manipulate the relational complexity and determine expected item difficulties.

From a CHC perspective, the test adequately represents fluid reasoning since it measures two different narrow abilities (Flanagan et al., 2013). An additional advantage of the test is that both kinds of items are non-verbal; given the growing interest in cross-cultural comparisons of cognitive abilities, the MID-CAT provides a more culturally fair measure because it is not influenced by language barriers.

Finally, the output of this study is a calibrated item bank of 516 items that provides a valuable resource for researchers. The item bank has a wide range of difficulties, allowing for the creation of tests tailored to specific populations or research questions. Even though the MID-CAT is particularly designed for research purposes, it can be administered in a higher-stakes context, too. This test is considered flexible, accessible and efficient for any research design. For instance, a more moderate SE threshold (e.g., SE < 0.45) will suffice if the test is used for screening. However, if, for some reason, a fixed-item stopping rule is required then 20 items per task is sufficient.

Even though this current article aims to provide a strong foundation for the MID-CAT, there is certainly room for further development. All items in MID-CAT were purely written by humans despite several methods to generate figural items automatically (e.g., Blum & Holling, 2018). It enabled us to create varied items with entirely different stimuli, sometimes using irregular shapes. However, we found some unusual empirical item difficulty levels, as some items were harder or easier than expected. This is particularly the case for the induction test. For example, we expected items that A4 and A3 (Figure 8) would have similar difficulty since they are based on a similar rule (i.e., the number of sides and dots). But apparently, the empirical item difficulties differ by 0.9 logits, showing that A3 is easier than A4. The complexity perceived by the item writer might not match the complexity perceived by test-takers. In a previous study with a similar task (the-odd-one-out), this was not the case when items were written more simply (i.e., varied only in the number of simple symbols) and systematically (Ruiz, 2009). Therefore, future research could adopt a more experimental strategy by manipulating item properties to identify the factors that influence item difficulty. Specifically, an investigation could identify whether a specific stimulus is associated with item difficulty.

We found that the multidimensional model was superior to both a single unidimensional model and separate unidimensional models, suggesting that inductive and deductive reasoning are independent yet related abilities. However, the factors in the multidimensional model might reflect methodological differences (sudoku-like vs. the odd-one task) in addition to different cognitive processes. In fact, the correlation between these two types of reasoning is high ($r = 0.72$). Given the ongoing debate regarding whether these two types of reasoning are fundamentally distinct, MID-CAT can be instrumental in addressing this issue. A notable feature of MID-CAT is its use of figural content to measure both inductive and deductive reasoning, which is believed to correlate closely with *g* (Carroll, 1993; Wilhelm, 2005). This aspect is particularly significant, as previous behavioral studies predominantly utilized verbal tasks to examine the dimensionality of reasoning (Hayes et al., 2018; Stephens et al., 2018).

Finally, some limitations should be noted. First, the performance of MID-CAT was achieved using a simulated dataset but has not yet been replicated with real data simulations or real-time CAT conditions. Second, the current study relied on social media advertising to recruit participants, which could have introduced selection bias. Third, the test was administered in an unproctored online study, therefore the generalization of the results to proctored testing is questionable. Fourth, the validity of the test was demonstrated through its correlation with HTM-S scores. As validation is an ongoing process, additional validity evidence is needed.

To conclude, the results of this study indicate that MID-CAT is feasible for measuring process factors of fluid reasoning in a precise and efficient way. Moreover, the test can be adaptively adjusted to accommodate the particular context in which fluid reasoning is needed to be measured.

Chapter 4: Psychometric and Psychological Evaluation of Multidimensional Computerized Adaptive Testing[8]

4.1. Background and aims

Research on computerized adaptive testing (CAT) has provided evidence of the psychometric advantages of CAT over traditional fixed-item tests (FIT). Although it has often been claimed that CAT provides a better user experience than FIT (e.g., Thompson, 2011), the evidence supporting this claim is not unequivocal. Several studies have compared FIT with unidimensional CAT with mixed results. In addition, to our knowledge, no studies have examined the psychological aspects of MCAT.

As mentioned earlier in the literature review, CAT possesses unique characteristics that differentiate it from FIT, such as adaptive item selection based on test-taker ability, which is often claimed to increase motivation (Wise, 2014). However, this also implies that CAT maintains a 50% success rate, which is considered too low to sustain motivation (B. A. Bergstrom et al., 1992). Additionally, CAT employs a complex scoring system that might be unfamiliar to most users, potentially negatively affecting their acceptance (Goto et al., 2023). However, informing examinees about adaptivity before testing could mitigate the negative psychological impact associated with unfamiliar testing formats (Ortner & Caspers, 2011). These characteristics might influence the test-taking experience in CAT compared to FIT. However, the impact is non-directional, as both better and worse experiences with CAT compared to FIT are possible.

As for MCAT, this method also has different features than unidimensional CAT. *First*, in multi-unidimensional CAT, the presentation of the item in the second test highly depends on the result of the first test. When test-takers perform well in the first test, they start the second test with a relatively difficult item, unlike separate unidimensional tests that often begin with easy items as a warm-up for test-takers (B. A. Bergstrom & Lunz, 1999). *Second*, MCAT often used scores on one test as collateral information to indicate ability on another test. Although using collateral information improves measurement precision, it is difficult to

[8] Chapter 4 is based upon the following study:
Akhtar, H., & Kovacs, K. (in press). Measurement Precision and User Experience with Adaptive versus Non-Adaptive Psychometric Tests. *Personality and Individual Differences*.

explain to lay people why a person's score on the first test depends partly on their performance on the second test, or vice versa (Wang et al., 2004). These issues may negatively affect public acceptance of the test, although it might be mitigated by informing examinees about how CAT works (Ortner & Caspers, 2011).

This study examined the effects of MCAT on both measurement precision and test-taking experience, focusing on test-taking motivation and feedback acceptance. Expectancy-value theory (Wigfield & Eccles, 2000b) is often used to explain why individuals are motivated to take tests. According to this theory, individuals are more motivated when they believe they can succeed (high *expectancy*) and *value* the test. In expectancy-value theory, expectancy reflects the test-taker's perception of how they will perform. However, the same dimensions that drive expected performance are believed to be relevant too for evaluating past performance. The *value* components examined in this study were *interest* and *cost* (i.e., anticipated *anxiety* and *effort* required to complete the task). Test taking-motivation can vary across different test items, influenced by the difficulty level of each administered item (S. L. Wise & Smith, 2011). CAT and FIT differ in the adaptivity of items administered; thus, different testing types could lead to different expectancy, interest, anxiety, and effort. Therefore, this study aims to examine these variables—expectancy, interest, anxiety, and effort—as components of test-taking motivation, to understand how they differ across testing types.

Another potential aspect differentiating MCAT from FIT could be test-takers' *feedback acceptance*, which is linked to perceived fairness (Tonidandel et al., 2002) and indirectly influences the face validity. Tonidandel et al. (2002) found that participants were more likely to accept feedback if their perceived performance was consistent with their actual performance. In FIT, actual performance is typically closely related to perceived performance (Macan et al., 1994) because the final test score depends on the number of correct answers. This is not expected to be the case in CAT because, in an ideal CAT scenario, all test-takers would answer around 50% of the items correctly (Bergstrom et al., 1992). Therefore, how individuals estimate their own performance (i.e., *self-estimated performance*) is a central aspect when comparing test-takers' experience in adaptive and non-adaptive tests.

This study also interested in the test duration difference between CAT and FIT. It is often claimed that CAT results in shorter test to reach certain level of SE, and thus shorter test duration (Wainer, 1993). However, when CAT uses fixed-test length, its efficiency effect due to shorter test is in question. It has been noted that test-takers generally allocate more time to questions they get wrong compared to those they answer correctly (Bergstrom et al., 1994; Chae et al., 2019; Hornke, 2000; Preckel & Freund, 2005; Yang et al., 2002). When the number of questions is constant, a reduced test duration suggests good test economy. According to the distance–difficulty hypothesis (Ferrando & Lorenzo-Seva, 2007), the time taken to respond to a question decreases as the gap between an individual's ability level and the task's difficulty increases. People tend to spend more time on tasks that match their abilities and less time on tasks that are either too simple or too hard. In CAT, the test-taker encounters questions that are suited to their ability level, which typically leads to a longer test duration.

Several studies have compared FIT with unidimensional CAT. However, to our knowledge, no studies have examined the psychological aspects of MCAT. Moreover, the psychometric and psychological aspects are frequently researched separately—the current study aimed to investigate both the psychometric and psychological impacts of CAT compared to FIT. The psychological aspects investigated in this study, later termed "test-taking experience," comprise effort, expectancy, interest, anxiety, self-estimated performance, and feedback acceptance. The tests used in the study consist of two subtests that measure two narrow abilities under Fluid reasoning: Induction and Deduction. Three questions were examined:

1. Is measurement precision different under MCAT and FIT?
2. Is test duration different under MCAT and FIT?
3. Is test-taking experience different under MCAT and FIT?
4. Are there different patterns of rapid guessing behavior between MCAT and FIT?

Measurement precision is operationalized as the standard error of the ability estimate (SE) after completing 20 subtest items. Test duration is operationalized as the time spent completing 20 items of each subtest. Test-taking experience is operationalized as test-takers' effort, expectancy, interest, anxiety, self-estimated performance, and feedback acceptance. Test-taking experience is operationalized as test-takers' effort, expectancy, interest, anxiety, self-estimated performance, and feedback acceptance. Two measures of effort were used: self-reported effort (SRE) and response time effort (RTE, Wise & Kong, 2005). SRE provides a global indicator of test-taking effort based on participants' self-ratings right after completing the tests. RTE is a time-based measure based on the assumption that unmotivated participants will answer items too quickly (i.e., before they have adequate time to read and fully consider the correct answer) (Wise & Kong, 2005). RTE makes it possible to investigate changes in participants' effort during a test session because response time data is available for each item.

Studies have consistently found that CAT outperforms FIT in terms of precision; therefore, we hypothesized that the SE in MCAT would be lower than in FIT. Regarding test duration, based on the distance–difficulty hypothesis, we hypothesized that the test duration in MCAT would be longer than in FIT. As for the test-taking experience, we posited non-directional hypotheses because both better and worse experiences with CAT compared to FIT are plausible, given the mixed findings in previous research.

4.2. Method

4.2.1. Participants

A total of 286 Indonesian adults aged 18-40 (M = 25.5, SD = 5.79) participated in this study. Participants were recruited through social media advertising (e.g., Instagram, Facebook, WhatsApp). No monetary incentives were provided to participants. A total of 140 participants completed the MCAT (97 females), and 146 participants completed the FIT (101 females). Participants mainly hold High School diplomas (37.76%) or Bachelor's degrees (37.76%). Residence distribution is nearly even, with 48.25% from rural and 51.75% from urban areas. Only 212 participants, 106 in each group, completed the questionnaires evaluating their test-taking experiences after the test.

4.2.2. Measures

Multidimensional Induction-Deduction Computerized Adaptive Test (MID-CAT)

MID-CAT (Akhtar & Kovacs, 2024) is a multidimensional computerized adaptive test measuring two process factors of fluid reasoning: induction and deduction. The test used in this chapter was the same as that in the previous chapter (see Figure 8 for sample items). The test consists of two subtests: one consisting of odd-one-out items (hf. Induction test) and one consisting of sudoku-like items (hf. Deduction test). Items in the Induction test contained six pictures. Test-takers need to identify the one that is different from the others based on a particular principle. Items in the deduction test are modified versions of the classic 6x6 Sudoku puzzle. Test-takers have to find which shape can replace the question mark so that each shape occurs only once in any column, row, or mini-grid.

The item bank for this study consists of 516 items with a wide range of difficulty. All items fit with the Rasch model. A fixed length stopping rule (20 items for each subtest) was applied. The test was developed using the mirtCAT package (Chalmers, 2016) in R. Details on the items and parameters are available in the online repository at https://osf.io/h74wd.

Multidimensional Induction-Deduction Fixed-Item Test (MID-FIT)

MID-FIT is a non-adaptive version of MID-CAT, consisting of 20 items for each subtest. Items from the item bank were selected using automated test assembly (ATA) methods using the xxIRT package in R (Luo, 2016). MID-FIT was assembled from items with average difficulty (M = 0, SD = 1) with the highest item-total correlation. The reliability of the tests was estimated using the data from the calibration of the item bank. Items were presented in increasing difficulty. The empirical reliability of the induction and the deduction test were 0.76 and 0.8, respectively. The Cronbach's Alpha reliabilities for this study's sample data were 0.88 for the induction test and 0.90 for the deduction test.

Test-taking motivation instrument

The Test-taking motivation questionnaire (Knekta & Eklöf, 2015) was adapted by Akhtar and Firdiyanti (2023). The original instrument was developed for a school context. The original instrument was developed for a school context. They translated and modified the instrument into a more general context. For instance, the original item of "*Compared with other students, I think I did well on this test*" was modified to "*Compared with other test-takers, I think I did well on this test*". We used the three relevant subscales: effort (hf. SRE for Self-Reported Effort, to differentiate from RTE, Response Time Effort), expectancy, and interest. SRE refers to participants' self-evaluation of the effort invested in the test. SRE consists of four items (i.e., "*I did my best on this test*", "*I worked with all items in the test without giving up, even when an item was difficult*", "*I felt motivated to do my best on this test*", "*I spent more effort on this test than I do on other tests*"). Expectancy refers to participants' perceptions of how well they performed. The expectancy subscale consists of two items (i.e., "*I did well on this test*", “*Compared with other test-takers, I think I did well on this test*”). Interest refers to how much participants enjoyed taking the test. The interest subscale consists of four items (i.e.,” *I am very curious about the result I received on this test*”, "*It was fun to do this test*", “*I looked forward to doing this test*”, “*I learned something new by doing this test*”). Participants rated their agreement on all items on a 4-point Likert-type scale ranging from 1 (strongly disagree) to 4 (strongly agree). The alpha reliability for SRE, interest, and expectancy was .74, .74, and .65, respectively.

State Anxiety questionnaire

The post-test state anxiety questionnaire was developed in Indonesian. The questionnaire items were similar to those used by Attali and Powers (2010). The instructions read as follows: "*How well do the following adjectives describe your feelings during the test that you just completed*?" The questionnaire consists of 12 adjectives (*Calm, Tense, Worried, Secure, Frightened, Anxious, At Ease, Nervous, Content, Jumpy, Pleasant, Confused*). Participants rated the given adjective's relevance on a 4-point Likert-type scale (*not at all, a little, moderately, very much*). The anxiety total score was defined as the sum of all item scores after reversing the scores of positive adjectives (Calm, Secure, At ease, Content, Pleasant). The Cronbach's alpha reliability was 0.90.

Self-estimated performance

Self-estimated performance was measured with a single-item: *"out of 40 items, how many items do you think you answered correctly?"*. To make it directly comparable to actual scores, we transformed the score to a percentage.

Feedback acceptance

Feedback acceptance refers to participants' belief that the feedback accurately reflects their performance. Test acceptance was assessed using a three-item scale from Nease et al. (1999). The item were "*The feedback I received is an accurate evaluation of my performance*", “*I agree with the feedback provided*”, “*It is hard to take feedback seriously*”.

Participants rated their agreement on a 5-point Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The Cronbach's Alpha reliability was 0.82.

RTE

RTE is a time-based measure of effort. The RTE index is based on the notion that answers provided below a particular time threshold indicate rapid guessing, as opposed to solution behaviour. The threshold for this study was set to be 5 seconds for all items, as used in the PISA tests (Buchholz et al., 2022). Therefore, if a participant responded slower than 5 seconds, their response was considered appropriate solution behaviour. In contrast, if a participant responded quicker than 5 seconds, their response was considered rapid guessing behaviour. The RTE index was calculated by summing the number of items reflecting solution behaviour and dividing by the number of items in the test. The RTE index was calculated for a specific subtest and the whole testing session.

4.2.3. Design

A between-subject design was used in the study; the independent variable was test type (MCAT vs. FIT). Participants in the MCAT group were informed about the adaptivity of item selection, whereas no such information was provided in the FIT group. Participants were randomly assigned to one of the two conditions. The dependent variables were measurement precision (i.e., SE), test duration, and test-taking experiences (i.e., interest, anxiety, expectancy, self-reported effort, response time effort, self-estimated performance, and feedback acceptance).

4.2.4. Procedures

Participants who were willing to participate in the study registered online. After reading the research description, participants consented and answered demographic questions. The link to complete the test was sent by email. The test was completed in an unproctored online environment. After completing the tests, participants were directed to questionnaires measuring test-taking experience. Participants who completed the questionnaires received feedback on their test scores. The feedback provided includes a description of the construct measured by the test, test results presented to participants at the end (i.e., theta and percentile), and a caution regarding the use of test results (for educational or entertainment purposes only, not clinical). After receiving the result, participants were asked to complete the feedback acceptance scale. The study protocol was approved by the research ethics committee of the Eotvos Lorand University (number 2022/529).

4.2.5. Analyses

All data analyses were performed in R (R Core Team, 2012). First, we reported the descriptive statistics for each group (Table 7). Then, we examined the correlations among the variables studied.

The first research question (*Is measurement precision different under MCAT and FIT?*) was examined with 2x2 mixed ANOVA. Test type was used as a between-subject factor, while the SE of the first and the second subtest (SE1 and SE2) were used as repeated measure factors. It should be noted that CAT is expected to exhibit equiprecision, meaning that SEs in CAT are not as varied as in FIT. This characteristic violates the assumptions of homogeneity of variance. Therefore, as expected, Levene's test indicated a lack of equality of variances and scatterplots are advised to be interpreted along with the statistical results.

The second research question (*Is test duration different under MCAT and FIT?*) was examined with 2x2 mixed ANOVA. Test type was used as a between-subject factor, while the test duration of the first and the second subtest (time1 and time2) were used as repeated measure factors.

The third research question (*Is test-taking experience different under MCAT and FIT?*) was examined with ANCOVAs. The p-values were adjusted using Benjamini & Hochberg (BH; 1995) method to account for the multiple comparisons issue. For RTE, since data were available for each subtest, the analysis was performed using 2x2 mixed ANCOVA. Test type was used as a between-subject factor, while the RTE index of the first and the second subtest (T1 and T2) were used as repeated measure factors. Composite theta (ability) was used as a covariate as it might affect test experience (Akhtar & Firdiyanti, 2023). For all analyses using ANCOVA, we utilized partial eta-squared (η_p^2) to measure effect sizes. Following guidelines by Ferguson (2009) for social sciences, a practically significant effect size is identified at 0.4. Additionally, effect sizes of 0.25 are classified as moderate, and sizes of 0.64 are deemed strong.

The fourth research question (*Are there different patterns of RGB between MCAT and FIT?)* was examined using Spearman's rank-order correlation, along with a visual inspection of the frequency of RGB as the function of item position. The Spearman correlation was used to determine the correlation between item position and RGB. In addition, the proportion of RGB for each item was calculated and plotted as a function of item position, with separate lines representing each testing mode.

4.3. Results

4.3.1. Descriptive statistics

Table 7 shows the descriptive statistics for FIT and MCAT. As expected, the percentage of correct scores was close to 50% for both FIT and MCAT. The standard deviation of the proportions of correct answers was twice as large in the FIT than in the MCAT tests, indicating that the raw scores were more varied in FIT than in MCAT. Most participants overestimated their scores: self-estimated performance was higher than the actual percentage of correct answers. Test performance was not different across CAT vs. FIT.

Table 7

Means (and Standard Deviations) of variables studied

Variables	MCAT	FIT	*d*
Percentage of correct	0.49 (0.11)	0.50 (0.22)	0.07
Theta induction	-0.04 (0.96)	0.10 (0.96)	0.15
Theta deduction	-0.49 (1.11)	-0.28 (1.10)	0.19
SE induction	0.39 (0.01)	0.43 (0.03)	1.91***
SE deduction	0.39 (0.01)	0.44 (0.03)	1.72***
SRE	3.47 (0.54)	3.51 (0.54)	0.07
Interest	3.43 (0.52)	3.46 (0.52)	0.04
Expectancy	3.06 (0.66)	3.13 (0.66)	0.07
Anxiety	2.16 (0.58)	2.00 (0.66)	0.25
Self-estimated performance	0.60 (0.20)	0.66 (0.19)	0.35*
Acceptance	3.86 (0.74)	3.86 (0.75)	0.01
RTE	0.97 (0.10)	0.94 (0.14)	0.28*
Test duration	24.94 (11.40)	21.99 (9.79)	0.28*

Note: MCAT=Multidimensional Computerized Adaptive Test, FIT = Fixed Item Tests, SRE = Self-Reported Effort, RTE = Response Time Effort, test duration = total test duration (in minutes), *d* = Cohen's *d*, * p < 0.05, *** p < 0.001.

Table 8 shows that test-taking experience scores were moderately correlated with test performance. However, the correlation coefficients were lower in the MCAT group than in the FIT group. For instance, the correlation between reported effort and test performance was significantly higher in the FIT condition (*r* = 0.44), whereas there was no correlation between reported effort and test performance in the MCAT condition (*r* = 0.04). Participants with more expectancy, effort, interest, and less anxiety tended to perform better. The two measures of test-taking effort, SRE and RTE, correlated weakly (*r* = 0.18).

Table 8

Correlations between variables studied

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Composite theta	-						
2. SRE	0.24***	-					
3. RTE	0.45***	0.18**	-				
4. Expectancy	0.25***	0.64***	0.08				
5. Interest	0.22**	0.55***	0.05	0.51***	-		
6. Anxiety	-0.33***	-0.20**	-0.06	-0.35***	-0.18**	-	
7. Self-estimated performance	0.48***	0.37***	0.17*	0.47***	0.29***	-0.42***	-
8. Acceptance	0.30***	0.44***	0.08	0.38***	0.42***	-0.15*	0.15*
MCAT condition							
1. Composite theta	-						
2. SRE	**0.04**	-					
3. RTE	0.51***	0.11	-				
4. Expectancy	**0.08**	0.60***	0.01	-			
5. Interest	0.14	0.60***	0.01	0.56***	-		
6. Anxiety	-0.29*	**-0.04**	0.02	-0.25**	-0.11	-	
7. Self-estimated performance	0.41***	0.27**	0.15	0.39***	0.24*	-0.32**	-
8. Acceptance	0.20*	0.43***	0.04	0.43***	0.47***	-0.06	0.20*
FIT condition							
1. Composite theta	-						
2. SRE	**0.44*** **	-					
3. RTE	0.44***	0.28**	-				
4. Expectancy	**0.40*** **	0.68***	0.18	-			
5. Interest	0.29**	0.50***	0.11	0.45***	-		
6. Anxiety	-0.35***	**-0.34*** **	-0.16	-0.44***	-0.25*	-	
7. Self-estimated performance	0.54***	0.48***	0.20*	0.55***	0.35***	-0.48***	-
8. Acceptance	0.40***	0.45***	0.25*	0.33**	0.36***	-0.23*	0.11

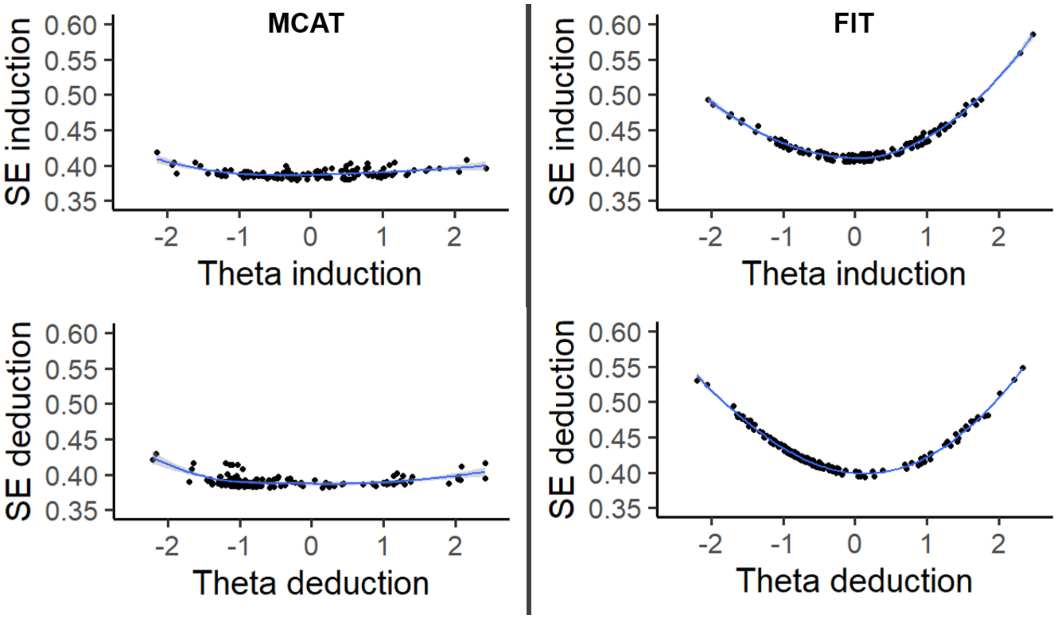
Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, SRE = Self-Reported Effort, RTE = Response Time Effort. Fisher's *r*-to-*z* transformation was used to assess the significance of the difference between the two correlation coefficients. A significant difference between the two correlation coefficients is printed in bold.

4.3.2. Measurement precision under MCAT and FIT

The analysis of SE using mixed Anova revealed a significant main effect of test type ($F(1, 284) = 272.21$, $p < 0.001$, $\eta_p^2 = 0.489$): the MCAT had lower SE than the FIT. In addition, the main effect of time (SE1 vs SE2) was also significant ($F(1, 284) = 9.49$, $p = 0.02$, $\eta_p^2 = 0.032$), showing that participants' SE increased in the second subtest. The relationship between participants' abilities (thetas) and SE is shown in Figure 13.

Figure 13

Relationship between theta and standard error (SE) in MCAT and FIT group

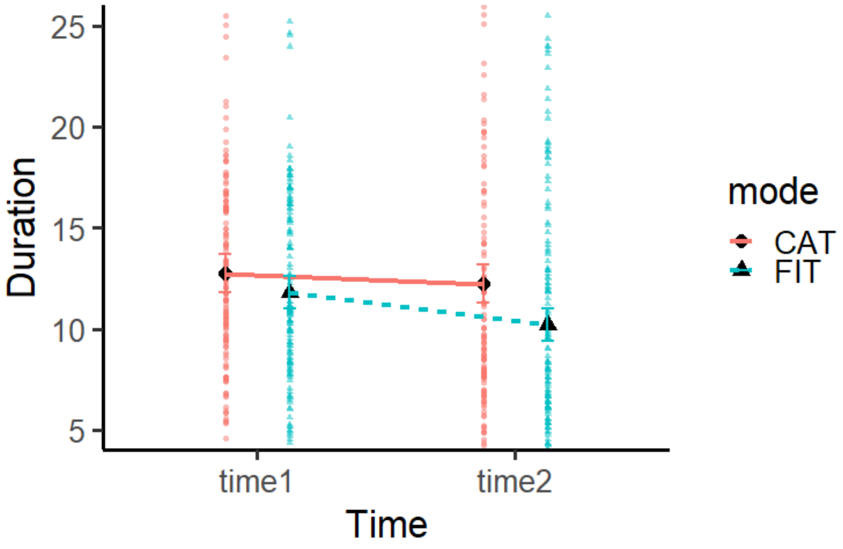


4.3.3. Test duration under MCAT and FIT

The analysis of test duration using mixed Anova revealed a significant main effect of test type ($F(1, 284) = 5.61$, $p = 0.019$, $\eta_p^2 = 0.019$): the MCAT had longer test duration than the FIT. In addition, the main effect of time (time1 vs time2) was also significant ($F(1, 284) = 11.50$, $p < 0.001$, $\eta_p^2 = 0.039$), showing that participants' test duration decreased in the second subtest. Figure 14 illustrates the effect of test position and test type on test duration.

Figure 14

Test duration on the first subtest (time1) and second subtest (time2) based on the test type



4.3.4. Test-taking experience under MCAT and FIT

The effects of test type and ability on six dependent variables (expectancy, SRE, interest, anxiety, self-estimated performance, and acceptance) were examined using ANCOVA. The summary of the results is presented in Table 9. Among all comparisons, a significant effect was only found for self-estimated performance: participants in the FIT condition reported a higher number of items answered correctly than those in the MCAT condition. For ability, significant effects were observed on all dependent variables. High-ability participants[9] reported higher effort, interest, expectancy, self-estimated performance, and acceptance, and demonstrated lower anxiety compared to low-ability participants. Specifically, the analyses revealed a medium effect size for expectancy ($d = 0.46$, $p < 0.001$), interest ($d = 0.356$, $p = 0.011$), acceptance ($d = 0.517$, $p < 0.001$), and anxiety ($d = 0.497$, $p < 0.001$). Additionally, a large effect size was observed for self-estimated performance ($d = 0.82$, $p < 0.001$).

[9] Ability was classified into two groups: high-ability (theta $\geq$ 0) and low-ability (theta < 0)

Table 9

ANCOVA results for dependent variables by test type

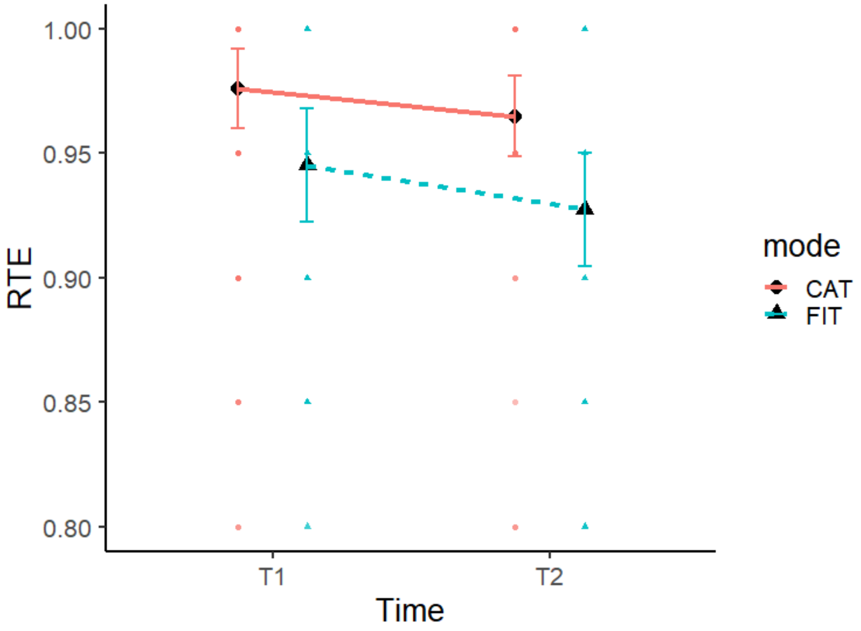
Dependent variable	Effect	*df*	*F*	*p*	Corr-p	η_p^2
Self-reported effort	Test type	1, 209	0.272	0.602	0.903	0.001
	Ability	1, 209	12.838	< 0.001		0.058
Interest	Test type	1, 209	0.100	0.752	0.903	0.001
	Ability	1, 209	10.241	0.001		0.047
Expectancy	Test type	1, 209	0.487	0.489	0.903	0.002
	Ability	1, 209	13.154	< 0.001		0.059
Anxiety	Test type	1, 209	3.732	0.055	0.164	0.018
	Ability	1, 209	24.278	< 0.001		0.104
Self-estimated performance	Test type	1, 209	8.381	0.004	0.025	0.039
	Ability	1, 209	59.793	< 0.001		0.222
Acceptance	Test type	1, 209	0.001	0.974	0.974	0.001
	Ability	1, 209	21.188	< 0.001		0.090

Note: Corr-p = corrected p-value for multiple comparisons

The analysis of RTE using 2X2 mixed Anova revealed a significant main effect of test type ($F(1, 283) = 12.10$, $p < 0.001$, $\eta_p^2 = 0.041$) and ability ($F(1, 283) = 76.35$, $p < 0.001$, $\eta_p^2 = 0.212$): taking the MCAT resulted in higher effort than taking the FIT. In addition, the main effect of time (T1 vs T2) was also significant ($F(1, 283) = 3.88$, $p = 0.049$, $\eta_p^2 = 0.013$), showing that participants' effort decreased in the second subtest. Figure 15 illustrates the effect of test position and test type on RTE.

Figure 15

Response Time Effort (RTE) on the first subtest (T1) and second subtest (T2) based on the test type

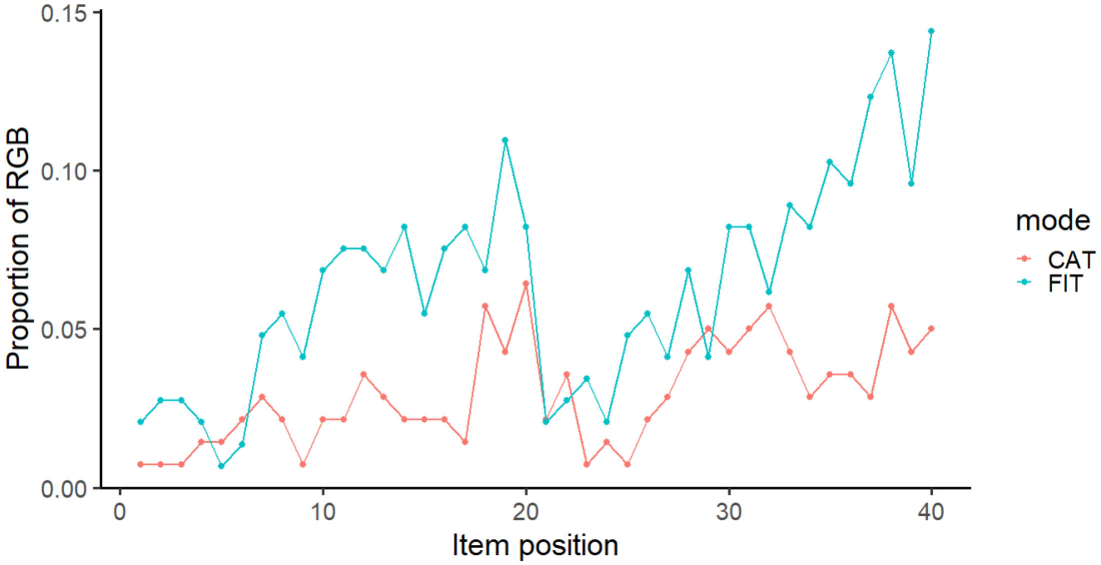


4.3.5. The pattern of rapid guessing behaviour in MCAT and FIT

The proportion of rapid guessing behaviour for each item was calculated and plotted. Figure 16 depicts the pattern of rapid guessing behaviour in CAT and FIT conditions. As shown in Figure 16, the proportion of rapid guessing behaviour in both MCAT and FIT conditions increased as the test progressed. However, the increasing rapid guessing response appeared more pronounced in the FIT group, especially in the final items, where the difficulty level was higher. In the MCAT group, the Spearman's correlation between item position and the proportion of rapid guessing behavior was $r = 0.72$ with $p < 0.001$ for the first test, and $r = 0.58$ with $p < 0.001$ for the second test. In the FIT group, it was $r = 0.86$ with $p < 0.001$ for the first test, and $r = 0.92$ with $p < 0.001$ for the second test.

Figure 16

Rapid guessing behaviour (RGB) across item position



4.4. Discussion

The main goal of this study was to examine whether test type affects measurement precision and participants' test-taking experience. Our findings demonstrate that MCAT is more precise than FIT (i.e., it has a lower SE with the same number of items). FIT, in general, is most appropriate for test-takers with medium ability levels, but when participants' abilities are below or above average (i.e., -1 < theta < +1), SE increases substantially. This is not the case in MCAT where SE is independent of ability (see Figure 13). This empirical result is in agreement with previous findings from simulation studies (Paap et al., 2019). These results are also identical to those found in the previous chapter based on a simulation study (see Figure 11).

As expected, completing 40 items in CAT requires more time than completing the same number of items in FIT, consistent with the distance–difficulty hypothesis (Ferrando & Lorenzo-Seva, 2007). This result is particularly important when aiming to claim the efficiency of CAT. Even though CAT generally requires a fewer number of items to achieve a certain level of precision, the number of items does not always correlate directly with test duration.

Although the MCAT leads to increased precision, it does not impact the overall test-taking experience. At the same time, participants' self-estimated performance was higher in the FIT condition: they reported a higher number of items answered correctly than in the MCAT condition. The higher self-estimated performance in FIT may be due to its structure, where questions start off very easy and become progressively harder. This contrasts with the MCAT, where the test begins with a question of medium difficulty, and the following questions are chosen based on adaptive criteria. Therefore, those who score high in MCAT do not encounter a single item of difficulty below average, but they solve several such items in FIT. This finding is consistent with the explanation of the Primacy bias (Anderson & Barrios, 1961), which suggests that earlier items have a larger relative effect on one's overall self-evaluation of performance. Several studies support this explanation, confirming that test-takers believed they answered more items correctly when the items were sorted from easiest to hardest, but not when ordered randomly (Bard & Weinstein, 2017; Jackson & Greene, 2014).

Although we did not find a difference in expectancy between the MCAT and FIT groups, the correlation between expectancy and test performance in the MCAT group is significantly lower than in the FIT group (see Table 8). This finding further indicates that participants in the CAT condition did not have a proper view of their own performance on the test, probably thanks to evaluating their own performance based on the number of items they believed to have answered correctly. Since in CAT, the actual test performance is not a direct function of the number of correct answers, as in FIT, participants might misjudge their overall performance.

Interestingly, we found different results when comparing FIT and MCAT based on self-report and time-based measures of effort: we found differences in RTE, but not in SRE. Researchers have suggested that SRE and RTE might not reflect on the same underlying mechanism of test-taking motivation (Akhtar & Firdiyanti, 2023; Silm et al., 2020). SRE is a subjective measure of effort that can reflect things other than test-taking motivation. Akhtar and Firdiyanti (2023) suggested that SRE is best predicted by individuals' perceptions of their performance. The relationship between perceived and actual performance is weaker on an adaptive test (Powell, 1994), as confirmed in our findings. Test takers may be biased toward claiming that they put in less effort than they actually did to justify their perceived failure.

RTE, on the other hand, reflects test-taking efforts based on RGB. Our results indicate that participants in MCAT spent more time with individual items. A distinct feature of CAT is that it provides items that are neither too difficult nor too easy for test-takers. Providing sufficient challenge for test-takers could retain their engagement on the test items. In FIT, items were sorted from the easiest to the hardest, and unmotivated participants exhibited RGB mostly at the end of the test (i.e., on the most difficult item). Our findings indicate that a match between test difficulty and the test-taker's ability is still desirable from a motivational point of view.

Our study has several limitations. First, we used uniform 5-second threshold applied to all items to determine RTE. There are other methods for setting item-specific thresholds, such as normative thresholds (see Soland et al., 2021). However, those methods require large samples for each item, and in our study, the sample size was limited. Second, the proportion of dropout participants between the two conditions is unequal and may potentially indicate bias. Third, the alpha reliability for expectancy was low (α = .65), so conclusions related to this measure should be interpreted with caution. Finally, since the test-taking experience is highly dependent on the stakes of the test, our conclusion should be limited to the context of unproctored online low-stakes testing.

Chapter 5: General Discussion

It is remarkable that after five decades of research on CAT, the use of this testing method in still underresearched. Despite the rapid advancements in research on the technical aspects of CAT, this type of test has not been fully implemented, especially in the field of cognitive ability research. Moreover, the evidence of whether CAT results in better psychological experience is not clear-cut (Wise, 2014). However, the rise of open-source adaptive testing platforms such as Concerto and mirtCAT has made creating and implementing CAT for research more cost-effective. Furthermore, the increasing availability of tutorials for these platforms makes the development and deployment of CAT more accessible and affordable. Therefore, using adaptive testing as a routine method in cognitive ability research is certainly feasible. This dissertation aimed twofold: 1) to develop a new multidimensional CAT measuring induction and deduction, and 2) to compare the psychometric and psychological aspects between CAT and FIT.

This dissertation contributes to the literature by providing a measure of Gf that is flexible, efficient, and entirely free for non-commercial use as well as pioneering empirical studies on the psychometric and psychological differences between CAT and FIT. This chapter begins with an overview of the main research questions and empirical findings. Subsequently, it delves into a discussion about how the output of this dissertation can be used for future research and the practical implications of these findings. Finally, it concludes by offering recommendations for further research.

Research Question 1: Is measurement precision different under adaptive and non-adaptive testing?

Measurement precision is substantially different in adaptive and non-adaptive tests. Specifically, in adaptive tests, measurement precision is relatively stable for all ability ranges. In contrast, in FIT, measurement precision depends on the items included in the test. For most cases, test items in FIT are designed to focus on assessing individuals of average ability, aiming for greater overall measurement precision. Consequently, this approach has led to less measurement precision for those with high and low abilities. While many studies found the same conclusion from simulation studies only (e.g., Ozdemir & Gelbal, 2022; Paap et al., 2019; Yao et al., 2014), the conclusion from this dissertation is based on the results of simulation (Chapter 3) and real-time testing (Chapter 4).

Moreover, this dissertation shows the psychometric benefits of MCAT over separate UCAT or FIT. The simulation study in Chapter 3 indicates that MCAT provides greater measurement efficiency: greater precision for fixed test lengths or shorter test lengths for precision-based stopping rule. When the precision-based stopping criterion is employed, test efficiency (i.e., shorter test duration) only appears for the second test. Although possessing greater reliability and r_{xt} while having a lower RMSE, the first test turns out to be considerably lengthier than the second. When fixed test length is employed, all the criteria exhibit a relatively similar balance between the first and second tests. The efficiency of MCAT over separate UCAT is predictable since the correlation between two narrow abilities is high. Paap et al. (2019) noted that the MCAT is substantially more efficient than separate UCAT, and the efficiency was more prominent when the correlation between dimensions was higher.

In all conditions in simulation studies, there is no difference in estimated ability between MCAT, UCAT, and FIT. Similarly, in real testing, the estimated ability was not significantly different when samples were randomly assigned to FIT and MCAT groups. Therefore, test type does not affect ability estimation. Even though examinees are administered different items, the estimated ability of MCAT is equivalent to non-CAT.

Research Question 2: Is the test-taking experience different under adaptive and non-adaptive testing?

The question regarding test-taking experience in adaptive and fixed-item tests is answered through systematic review and meta-analysis (Chapter 2) as well as empirical investigation using our test (Chapter 4). The systematic review and meta-analysis examined the effect of CAT on motivation and anxiety compared to traditional FIT based on 11 studies. The results demonstrated no overall effect of test type on anxiety and motivation. It should be noted that the study's results vary depending on the context and the outcome measures used. Modifying traditional item selection in CAT by selecting easier items (i.e., ECAT) could result in a better experience. When comparing ECAT with FIT, samples tested with ECAT showed less anxiety.

I also examined the psychological effects of using CAT compared to FIT. I compared MID-CAT with MID-FIT and found that CAT had a minimal impact on the test-taking experience. The CAT and FIT groups had no discernible differences in expectancy, interest, and anxiety. Moreover, participants' acceptance of CAT and FIT does not differ significantly. However, a different result was obtained based on response times analysis; participants exerted greater effort when dealing with CAT. Individuals in the FIT group displayed a more optimistic outlook regarding their performance, even though their actual performance closely resembled that of the CAT group. Notably, there was a weak correlation between participants' self-assessed performance and their actual performance in the MCAT condition, indicating that test-takers might misjudge their true performance.

5.1. Significance of this Dissertation

This dissertation has two outputs: 1) a multidimensional Gf test that is flexible, efficient, and accessible, and 2) the evidence on the advantage of CAT over FIT in terms of psychometric properties and test-taking experience. The first output is the calibrated item bank of Gf tasks consisting of 516 items measuring two narrow abilities of Gf: induction and deduction. The final item pool has a wide range of difficulties that could precisely measure individuals with a wide range of ability scores. To my best knowledge, no multidimensional Gf test has been developed in an adaptive version specifically for non-commercial purposes. MID-CAT can be a valuable resource for future research on cognitive abilities. All resources regarding this test, including data, test specification, and item properties, are available in the online repository (https://osf.io/h74wd/). Researchers are invited to adopt, modify, or further analyze the tests and data for their own purposes to gain more insight.

There are several conditions in which MID-CAT could be helpful for researchers. *First*, several research designs often include pretest and posttest administration. Parallel tests are often used to ensure that their resulting scores are placed on a common scale without item-learning effects (e.g., Kyllonen et al., 2019). For the same purpose, MID-CAT is preferred over parallel tests because it provides a more efficient assessment. *Second*, one challenge in cognitive ability research is maintaining participants’ test-taking motivation while working on a mentally taxing task. Our previous findings found that items at the end of the test may be answered carelessly by the participants, which could bias the results

(Akhtar, 2022b; Akhtar & Kovacs, 2023b). Collecting data using shorter tests like MID-CAT can maintain participants' motivation without sacrificing measurement precision. *Third*, given the ongoing debate among cognitive psychologists about whether inductive and deductive reasoning are fundamentally distinct, MID-CAT can play a crucial role in addressing this issue. Using figural content to assess both types of reasoning could reduce the cultural and linguistic biases inherent in assessments (Otero, 2017) while minimizing confounds between process and content factors. *Finally*, the flexibility in administering MID-CAT (e.g., number of items, precision level) benefits researchers because it can be adapted to their research goals and circumstances. The test can be administered online so that it can reach a large number of participants.

Prospective users might question how many items must be administered to reach a desirable level of measurement precision in MID-CAT. As the required measurement precision of the test result is the function of the stakes of the testing session, there is no single recommendation on how many items should be administered. The rule of thumb minimum reliability for high-stakes assessment is 0.9 (corresponds to SE of 0.32), while for low-stakes assessment is 0.8 (corresponds to SE of 0.45) (Downing, 2004). Test administrators could simply set the stopping rule to achieve this level of precision so that all examinees will have equal measurement precision. However, in the case of MID-CAT, where intermixing items between dimensions is not allowed, it will result in an imbalance in the number of items between two subtests. This is because the first test equals unidimensional CAT, while for the second test, the estimation of the provisional ability benefits from multidimensional IRT, resulting in a shorter test. Alternatively, the test administrator can set a fixed number of items. Based on the simulation study, ideally, 40 items per subtest are administered for high-stakes assessment, while administering 20 items per subtest is sufficient for lower-stakes assessment. For tests focusing on group scores, not individual scores, as in the research context, 10 items are sufficient.

One of the primary goals of my dissertation is to develop an efficient, flexible, and entirely free Gf test for research purposes. The administration of the tasks can be adjusted according to the research goal and situation. In situations where using an adaptive test platform is not feasible, the item bank can be easily configured in a fixed format. Even though fixed-item tests typically come with the cost of decreased precision, this type of testing is more practical in terms of administration (i.e., no special software needed). The items used in the fixed-item tests for this dissertation (MID-FIT) are accessible and can be freely utilized for non-commercial purposes. Comprehensive information for implementing MID-FIT into research projects is available on OSF (https://osf.io/h74wd/).

This dissertation also contributes to providing evidence on the comparability of CAT and FIT in terms of test-taking experience. Several commercial test developers claim that CAT leads to better experience and increased motivation because each examinee will be provided with an appropriate challenge (e.g., Thompson, 2011). Other researchers express concerns about the fairness of CAT, citing potential negative psychological reactions among examinees (e.g., Ortner et al., 2014; Tonidandel & Quiñones, 2000). Based on meta-analysis and empirical study, no substantial differences in psychological experiences between CAT and FIT users were found. Nonetheless, the use of ECAT (a CAT targeted at higher success rate) was associated with a better experience. It should be noted that this conclusion is most directly relevant to conventional FIT that is designed with a variety of item difficulties, similar to what was employed in this study. In principle, FIT could be designed with no variability in difficulty (e.g., only easy-difficulty items used), even though it is not practical from a psychometric perspective. It would lead to differences in the test experience between CAT and FIT for a larger portion of test-takers.

5.2. Implications

The studies in this dissertation imply that although it provides clear psychometric benefits, CAT may not result in a substantially different test experience for most examinees. As also noted by Ling et al. (2017), if examinees perceive a CAT as no different from a FIT, then the appeal of adaptive testing as a more efficient alternative to traditional fixed-item testing could potentially increase. Some might hesitate to use CAT over FIT in real testing because they do not understand how it works. In addition, researchers also found that certain features of adaptive tests lead to negative reactions (Ortner & Caspers, 2011; Tonidandel & Quiñones, 2000). In my study, examinees were informed about adaptivity before testing. This information is essential, as Ortner and Caspers (2011) found that informing examinees about the mechanisms and procedures employed in adaptive testing led to better results than presenting standard instructions. Providing additional information before testing could prevent the adverse psychological impact of using unfamiliar testing types. Therefore, based on the results of this dissertation, there are no arguments to advise against the use of adaptive testing.

The findings of this dissertation also indicate that MCAT outperformed UCAT or FIT in measurement efficiency. In brief, when dealing with multiple subtests measuring different abilities, transitioning from a unidimensional approach to a multidimensional one can result in significantly increased reliability and a more accurate measurement of the relationship between these abilities. When two subtests are highly correlated, item responses to another subtest are treated as collateral information, which can be used to increase the measurement precision (Wang et al., 2004). The multidimensional model offers greater flexibility and performs effectively, even where the actual test is unidimensional (Sheng & Wikle, 2007).

When using multidimensional CAT, one might doubt how a person's score on a test may depend partly on their performance on other tests. In addition, the correlation between the number of correct answers and the final scores (theta) is lower in CAT than in FIT. This can lead to distrust in the test results, which could indirectly damage the face validity of the test. However, the findings indicated no difference in feedback acceptance between the CAT and FIT groups. Such an observation is crucial for dispelling doubts regarding the adoption and utilization of CAT in various testing environments. As mentioned earlier, providing information about the scoring procedure in CAT before the testing session could prevent misunderstandings about CAT, which could minimize the adverse impact of using CAT.

This study also provides additional evidence that, for a small subgroup of examinees, test-taking effort declines as the test progresses. This is notably observed in FIT, where the items presented become increasingly difficult. This finding aligns with previous research (Akhtar, 2022b; Pastor et al., 2019; S. L. Wise et al., 2009). Based on these findings, researchers or practitioners might avoid using overly mentally taxing items in low-stakes assessments. Additionally, they might want to consider using shorter tests or implementing motivational interventions towards the end of the tests. However, shorter tests are typically less reliable. Using adaptive tests could be beneficial in this context, as they can maintain reliability while being relatively shorter compared to fixed-item tests.

Finally, modifying the item selection algorithm in CAT could be applied to enhance the test-taking experience. In the meta-analysis study, when comparing ECAT and FIT, we found that ECAT resulted in higher motivation than regular CAT or FIT. Due to only a few studies examining this topic, it is too premature to suggest moving from traditional CAT to select easier items. However, the empirical study in Chapter 4 also partially supports this suggestion. Starting the test with easier items, such as in the FIT group, results in a more optimistic performance evaluation. These results might be a promising foundation for implementing ECAT in real testing. However, selecting easier items means selecting suboptimal items, which could damage measurement efficiency. Therefore, practical considerations should be made to maximize the trade-off between test-taking experience and measurement efficiency.

Chapter 6: Conclusions and Future Directions

6.1. Conclusions

In summary, the adoption of CAT over FIT is underpinned by its unequivocal psychometric advantages. CAT's dynamic approach tailors the assessment to an individual's ability level, resulting in more reliable measurements. Remarkably, CAT achieves this psychometric excellence without compromising the test-taking experience. Test-takers assessed with CAT do not experience differences from those tested with FIT. Given the significantly improved efficiency of CAT while maintaining a similar test-taking experience, the appeal of using CAT is likely to increase.

Besides providing evidence of the psychometric and psychological impact of CAT, this dissertation also produces a multidimensional adaptive test for measuring two narrow abilities of fluid reasoning. The test is flexible, efficient, and entirely free for non-commercial use. All data regarding this test can be accessed and utilized by other researchers for their own purposes. Therefore, this dissertation contributes not only to the field of testing methodology but also to instruments that can be employed in cognitive ability research.

6.2. Future directions

Based on the research findings in this dissertation, there are several research agendas that merit further exploration. First, selecting easier items than the provisional estimated ability results in a better experience. Thus, CAT developers may consider modifying the CAT algorithm to optimize the experience from a psychological perspective. Investigating the ideal difficulty level for the items administered would be appealing. Future research could consider experimenting with different item selection algorithms. Item selection could be varied in expected success probabilities for administered items (e.g., 50%-90%) and evaluate how this affects both the test-takers' experience and the measurement efficiency.

Second, test-taking experience is a function of the stakes of the test. My studies were conducted in low-stakes assessments, with no personal consequences for participants. The psychological impact would have been different in high-stakes assessments where the personal consequences exist. Therefore, future research should extend the generalization of the findings by performing the research in the context of high-stakes assessment.

Third, it has been suggested that test-taking motivation affects test performance. In order to increase fairness in CAT, future research is expected to explore features of testing that could increase test-taking motivation. Test features such as the number of items, order of items, test review, and feedback presence could be manipulated in future research to see their effect on test-taking motivation.

Finally, it is important to consider the timeframe of the study (conducted between 2022 and 2023) when interpreting the findings. Given that CAT is not currently prevalent in Indonesia, almost all participants were experiencing a CAT for the first time. Participants' reactions to CAT could represent a novelty effect, and it is conceivable that their attitudes toward CAT may evolve as CAT becomes more commonplace. These attitudes may either become more positive or negative. Therefore, it is recommended that future research continues to investigate people’s attitudes toward CAT, recognizing that the testing environment and experience may influence these attitudes.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Adams, R. J., & Wu, M. L. (2007). The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (pp. 57–75). Springer. https://doi.org/10.1007/978-0-387-49839-3_4

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEICE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1093/ietfec/e90-a.12.2762

Akhtar, H. (2022a). Measuring Fluid Reasoning and Its Cultural Issues: A Review in the Indonesian Context. *Buletin Psikologi*, *30*(2), Article 2. https://doi.org/10.22146/buletinpsikologi.74475

Akhtar, H. (2022b). The pattern of test-taking effort across items in cognitive ability test: A latent class analysis. *Proceeedings of the 19th International Conference on Cognition and Exploratory Learning in the Digital Age (CELDA 2022).* http://dx.doi.org/10.33965/celda2022_202207l021

Akhtar, H., & Firdiyanti, R. (2023). Test-taking motivation and performance: Do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? *Learning and Individual Differences*, *106*, 102323. https://doi.org/10.1016/j.lindif.2023.102323

Akhtar, H., & Kovacs, K. (2023a). *Five decades of research on computerized adaptive testing: A bibliometric analysis* [Manuscript submitted for publication].

Akhtar, H., & Kovacs, K. (2023b). Which tests should be administered first, ability or non-ability? The effect of test order on careless responding. *Personality and Individual Differences*, *207*, 112157. https://doi.org/10.1016/j.paid.2023.112157

Akhtar, H., & Kovacs, K. (2024). Measuring Process Factors of Fluid Reasoning Using Multidimensional Computerized Adaptive Testing. *Assessment*, 10731911241236351. https://doi.org/10.1177/10731911241236351

Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology*, *63*, 346–350. https://doi.org/10.1037/h0046719

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Araci, F. G. İ., & Tan, Ş. (2022). Multidimensional Computerized Adaptive Testing Simulations in R. *International Journal of Assessment Tools in Education*, *9*(1), Article 1. https://doi.org/10.21449/ijate.909616

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational Components of Test Taking. *Personnel Psychology*, *43*(4), 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

Asseburg, R., & Frey, A. (2013). Too hard , too easy , or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*(1), 92–104.

Attali, Y., & Powers, D. (2010). Immediate Feedback and Opportunity to Revise Answers to Open-Ended Questions. *Educational and Psychological Measurement*, *70*(1), 22–35. https://doi.org/10.1177/0013164409332231

Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, *102*(1), 1–8. https://doi.org/10.1007/s10649-019-09908-4

Bard, G., & Weinstein, Y. (2017). The Effect of Question Order on Evaluations of Test Performance: Can the Bias Dissolve? *Quarterly Journal of Experimental Psychology*, *70*(10), 2130–2140. https://doi.org/10.1080/17470218.2016.1225108

Barry, C. L., & Finney, S. J. (2016). Modeling Change in Effort Across a Low-Stakes Testing Session: A Latent Growth Curve Modeling Approach. *Applied Measurement in Education*, *29*(1), 46–64. https://doi.org/10.1080/08957347.2015.1102914

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342–363. https://doi.org/10.1080/15305058.2010.508569

Barton, M. A., & Lord, F. M. (1981). An Upper Asymptote for the Three-Parameter Logistic Item-Response Model. *ETS Research Report Series*, *1981*(1), i–8. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462. https://doi.org/10.1007/BF03173192

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

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In *Innovations in computerized assessment* (pp. 67–91). Lawrence Erlbaum Associates Publishers.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the Level of Difficulty in Computer Adaptive Testing. *Applied Measurement in Education*, *5*(2), 137–149. https://doi.org/10.1207/S15324818AME0502_4

Bergstrom, B., Gershon, R., & Lunz, M. E. (1994). *Computerized Adaptive Testing Exploring Examinee Response Time Using Hierarchical Linear Modeling*. Paper presented at: the Annual Meeting of the National Council on Measurement in Education; April 4-8, 1994, New Orlenas, USA. https://eric.ed.gov/?id=ED400287

Betz, N. E., & Weiss, D. J. (1976). *Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing.* (Research Report 76–4).

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In *Statistical Theories of Mental Test Scores* (pp. 397–479). Addison-Wesley.

Blum, D., & Holling, H. (2018). Automatic Generation of Figural Analogies With the IMak Package. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01286

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2015). Regression in Meta-Analysis. In *Comprehensive meta analysis*. https://www.meta-analysis.com/pages/cma_manual.php

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Boring, E. G. (1923). Intelligence as the tests test it. *The New Republic*, *35*, 35–37.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2010). *Test of nonverbal intelligence (4th ed.)*. PRO-ED.

Buchholz, J., Cignetti, M., & Piacentini, M. (2022). *Developing measure of engagement in PISA* [OECD Education Working Paper]. https://one.oecd.org/document/EDU/WKP(2022)17/en/pdf

Burr, S. A., Gale, T., Kisieleweska, J., Milin, P., Pego, J. M., Pinter, G., Robinson, I. M., & Zahra, D. (2023). A narrative review of adaptive testing and its application to medical education. *MedEdPublish*, *13*, 221. https://doi.org/10.12688/mep.19844.1

Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Carroll Theory: Empirical, Clinical, and Policy Implications. *Applied Measurement in Education*, *32*(3), 232–248. https://doi.org/10.1080/08957347.2019.1619562

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. In *Cambridge University Press*. Cambridge University Press. https://doi.org/10.1177/001698629904300207

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. North-Holland.

Cattell, R. B., Krug, S. E., & Barton, K. (1973). *Technical supplement for the Culture Fair Intelligence Tests, Scales 2 and 3*. Institute for Personality and Ability Testing.

Chae, Y., Park, S. G., & Park, I. (2019). The relationship between classical item characteristics and item response time on computer-based testing. *Korean Journal of Medical Education*, *31*(1), 1–9. https://doi.org/10.3946/kjme.2019.113

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software*, *71*(5), 1–38. http://dx.doi.org/10.18637/jss.v071.i05

Chang, H.-H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement*, *20*(3), 213–229. https://doi.org/10.1177/014662169602000303

Chien, T. W., & Wang, W. C. (2017). An online multidimensional computerized adaptive testing (MCAT) module using APP. *Rasch Measurement Trans;*, *31*(1), 1625-6.

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S.-J. (2019). The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science*, *6*(10), 190232. https://doi.org/10.1098/rsos.190232

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*(1), 52–64. https://doi.org/10.1016/j.intell.2014.01.004

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*(1), 160. https://doi.org/10.1186/1471-2288-11-160

de Beer, M. (2013). The Learning Potential Computerised Adaptive Test in South Africa. In S. Laher & K. Cockcroft (Eds.), *Psychological Assessment in South Africa* (pp. 137–157). Wits University Press. https://doi.org/10.18772/22013015782.15

DeMars, C. E. (2010). Test Stakes and Item Format Interactions. *Applied Measurement in Education*, *13*(1), 55–77. https://doi.org/10.1207/S15324818AME1301_3

Deville, C. (1993). Flow as a testing ideal. *Rasch Measurement Transactions*, *7*(3), 308.

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006–1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7716–7720. https://doi.org/10.1073/PNAS.1018601108/SUPPL_FILE/SD01.XLS

Eccles, J. S., & Wigfield, A. (2002). Motivational Beliefs, Values, and Goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A Cross-National Comparison of Reported Effort and Mathematics Performance in TIMSS Advanced. *Applied Measurement in Education*, *27*(1), 31–45. https://doi.org/10.1080/08957347.2013.853070

Elbarbary, R. (2020). Impact of fixed and variable length computer adaptive test designs in reducing test anxiety and developing attitudes towards online exams: Faculty of education students’ case. *Educational Technology: Studies and Research Series*, *30*(1), 23–87. https://doi.org/10.21608/tesr.2020.91492

Emberton, S. E., & Reise, S. P. (2000). *Multivariate Applications Books Series. Item response theory for psychologists.* Lawrence Erlbaum Associates Publisher.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. https://doi.org/10.1037/a0015808

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An Item Response Theory Model for Incorporating Response Time Data in Binary Personality Items. *Applied Psychological Measurement*, *31*(6), 525–543. https://doi.org/10.1177/0146621606295197

Flanagan, D. P., & Dixon, S. G. (2014). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In *Encyclopedia of Special Education*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118660584.ese0431

Flanagan, D. P., Ortiz, S., & Alfonso, V. (2013). *Essentials of cross-battery assessment (3rd ed.)*. John Wiley & Sons.

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effects of adaptive testing on test-taking motivation with the example of the Frankfurt Adaptive Concentration Test. *Diagnostica*, *55, 20–28*, *55*, 20–28. https://doi.org/10.1026/0012-1924.55.1.20

Fritts, B. E., & Marszalek, J. M. (2010a). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education*, *13*(3), 441–458. https://doi.org/10.1007/s11218-010-9113-3

Fritts, B. E., & Marszalek, J. M. (2010b). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education*, *13*(3), 441–458. https://doi.org/10.1007/s11218-010-9113-3

Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortexin inductive and deductive reasoning. *Cognition*, *93*(3), B109–B121. https://doi.org/10.1016/j.cognition.2004.03.001

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Goto, T., Kano, K., & Shiose, T. (2023). Students' acceptance on computer-adaptive testing for achievement assessment in Japanese elementary and secondary school. *Frontiers in Education*, *8*. https://www.frontiersin.org/articles/10.3389/feduc.2023.1107341

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence Predicts Health and Longevity, but Why? *Current Directions in Psychological Science*, *13*(1), 1–4. https://doi.org/10.1111/j.0963-7214.2004.01301001.x

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*(3), 179–203. https://doi.org/10.1016/0160-2896(84)90008-4

Gustafsson, J. E., & Wolff, U. (2015). Measuring fluid intelligence at age four. *Intelligence*, *50*(1), 175–185. https://doi.org/10.1016/j.intell.2015.04.008

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications.

Harrison, P. M. C., & Müllensiefen, D. (2018). Development and Validation of the Computerised Adaptive Beat Alignment Test (CA-BAT). *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-30318-8

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly*, *50*(1), 75–87.

Hayes, B. K., Stephens, R. G., Ngo, J., & Dunn, J. C. (2018). The dimensionality of reasoning: Inductive and deductive inference can be explained by a single process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1333–1351. https://doi.org/10.1037/xlm0000527

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*(2), 388–395. https://doi.org/10.1037/0033-2909.93.2.388

Heydasch, T., Haubrich, J., & Renner, K.-H. (2013). The Short Version of the Hagen Matrices Test (HMT-S): 6-Item Induction Intelligence Test. *methods, data, analyses*, *7*(2), Article 2. https://doi.org/10.12758/mda.2013.011

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2021). *Cochrane Handbook for Systematic Reviews of Interventions* (version 6.). Cochrane. Available from www.training.cochrane.org/handbook.

Hofverberg, A., Eklöf, H., & Lindfors, M. (2022). Who Makes an Effort? A Person-Centered Examination of Motivation and Beliefs as Predictors of Students' Effort and Performance on the PISA 2015 Science Assessment. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.791599

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M. P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M. C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), Article 4.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review*, *79*, 242–259.

Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica*, *21*, 175–189.

Hornke, L. F., Küppers, A., & Etzel, S. (2000). Design and evaluation of an adaptive matrices test. *Diagnostica*, *46*(4), 182–188. https://doi.org/10.1026//0012-1924.46.4.182

Irribarra, D. T., & Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration* [Computer software]. https://github.com/david-ti/wrightmap

Jackson, A., & Greene, R. L. (2014). Impression formation of tests: Retrospective judgments of performance are higher when easier questions come first. *Memory & Cognition*, *42*(8), 1325–1332. https://doi.org/10.3758/s13421-014-0439-5

Kachergis, G., Marchman, V. A., Dale, P. S., Mankewitz, J., & Frank, M. C. (2022). Online Computerized Adaptive Tests of Children's Vocabulary Development in English and Mexican Spanish. *Journal of Speech, Language, and Hearing Research*, *65*(6), 2288–2308. https://doi.org/10.1044/2022_JSLHR-21-00372

Kamphaus, R., Winsor, A. P., Rowe, & Kim, S. (2018). A History of Intelligence Test Interpretation. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues, 4th ed* (pp. 73–163). The Guilford Press.

Kent, P. (2017). Fluid intelligence: A brief history. *Applied Neuropsychology: Child*, *6*(3), 193–203. https://doi.org/10.1080/21622965.2017.1317480

Kim, J., & McLean, J. E. (1995). The influence of examinee test-taking behavior motivation in computerized adaptive testing. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*.

Kim, S. Y., Lee, W.-C., & Kolen, M. J. (2020). Simple-Structure Multidimensional Item Response Theory Equating for Multidimensional Tests. *Educational and Psychological Measurement*, *80*(1), 91–125. https://doi.org/10.1177/0013164419854208

Kiskis, S. (1991). *Effects of test administrations on general, test, and computer anxiety, and efficacy measures.* [Theses Digitization Project. 579.]. https://scholarworks.lib.csusb.edu/etd-project/579

Klein, B., Raven, J., & Fodor, S. (2018). Scrambled Adaptive Matrices (SAM) – a new test of eductive ability. *Psychological Test and Assessment Modeling*, *60*(4), 451–492.

Knekta, E., & Eklöf, H. (2015). Modeling the Test-Taking Motivation Construct Through Investigation of Psychometric Properties of an Expectancy-Value-Based Questionnaire. *Journal of Psychoeducational Assessment*, *33*(7), 662–673. https://doi.org/10.1177/0734282914551956

Koch, M., Spinath, F. M., Greiff, S., & Becker, N. (2022). Development and Validation of the Open Matrices Item Bank. *Journal of Intelligence*, *10*(3), 41. https://doi.org/10.3390/jintelligence10030041

Kovacs, K., & Conway, A. R. A. (2019). A Unified Cognitive/Differential Approach to Human Intelligence: Implications for IQ Testing. *Journal of Applied Research in Memory and Cognition*, *8*(3), 255–272. https://doi.org/10.1016/j.jarmac.2019.05.003

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kroehne, U., Goldhammer, F., & Partchev, I. (2014). Constrained Multidimensional Adaptive Testing without intermixing items from different dimensions. *Psychological Test and Assessment Modeling*, *56*(4), 348–367.

Kvist, A. V., & Gustafsson, J. E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment theory. *Intelligence*, *36*(5), 422–436. https://doi.org/10.1016/j.intell.2007.08.004

Kyllonen, P., Hartman, R., Sprenger, A., Weeks, J., Bertling, M., McGrew, K., Kriz, S., Bertling, J., Fife, J., & Stankov, L. (2019). General fluid/inductive reasoning battery for a high-ability population. *Behavior Research Methods*, *51*(2), 507–522. https://doi.org/10.3758/s13428-018-1098-4

Lakin, J. M., & Gambrell, J. L. (2012). Distinguishing verbal, quantitative, and figural facets of fluid intelligence in young students. *Intelligence*, *40*(6), 560–570. https://doi.org/10.1016/j.intell.2012.07.005

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211–240.

Lee, N. Y. L., Goodwin, G. P., & Johnson-Laird, P. N. (2008). The psychological puzzle of Sudoku. *Thinking and Reasoning*, *14*(4), 342–364. https://doi.org/10.1080/13546780802236308

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Erlbaum.

Lunz, M. E., Bergstrom, B. A., & Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research*, *21*(6), 623–634. https://doi.org/10.1016/0883-0355(94)90015-9

Luo, X. (2016). *xxIRT: R Package for Item Response Theory* (2.1) [Computer software]. https://github.com/xluo11/xxIRT

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, *47*(4), 715–738. https://doi.org/10.1111/J.1744-6570.1994.TB01573.X

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.c01

Makransky, G., & Glas, C. A. W. (2013). The Applicability of Multidimensional Computerized Adaptive Testing for Cognitive Ability Measurement in Organizational Assessment. *International Journal of Testing*, *13*(2), 123–139. https://doi.org/10.1080/15305058.2012.672352

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205

Martín-Fernández, M., Ponsoda, V., Olea, J., Shih, P.-C., & Revuelta, J. (2016). A multistage adaptive test of fluid intelligence. *Psicothema*, *28*(3), 346–352. https://doi.org/10.7334/psicothema2015.287

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Matzke, D., Dolan, C. V., & Molenaar, D. (2010). The issue of power in the identification of "g" with lower-order factors. *Intelligence*, *38*(3), 336–344. https://doi.org/10.1016/J.INTELL.2010.02.001

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Mohd Ali, S., Norfarah, N., Ilya Syazwani, J. I., & Mohd Erfy, I. (2019). The effect of computerized-adaptive test on reducing anxiety towards math test for polytechnic students. *Journal of Technical Education and Training*, *11*(4), 27–35. https://doi.org/10.30880/jtet.2019.11.04.004

Naglieri, J. A. (2016). *Naglieri Nonverbal Ability Test–Third Edition (NNAT3)*. Pearson.

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2022). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*, 1–16. https://doi.org/10.1111/jcal.12719

Nease, A. A., Mudgett, B. O., & Quiñones, M. A. (1999). Relationships among feedback sign, self-efficacy, and acceptance of performance feedback. *Journal of Applied Psychology*, *84*, 806–814. https://doi.org/10.1037/0021-9010.84.5.806

Olea, J., Revuelta, J., Ximénez, M. C., & Abad, F. J. (2000). Psichometric and psychological effects of review aon computerized fixed and adaptive test. *Psicológica: Revista de Metodología y Psicología Experimental*, *21*(1), 157–174.

Oppl, S., Reisinger, F., Eckmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*, *14*(1), 2. https://doi.org/10.1186/s41239-017-0039-0

Orive, R., & Gerard, H. B. (1987). The familiar stimulus as a reducer of anxiety: An experimental study. *Journal of Social and Clinical Psychology*, *5*(3), 330–338. https://doi.org/10.1521/jscp.1987.5.3.330

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, *27*(3), 157–163. https://doi.org/10.1027/1015-5759/a000062

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I will probably fail: Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment*, *30*(1), 48–56. https://doi.org/10.1027/1015-5759/a000168

Otero, T. M. (2017). Brief review of fluid reasoning: Conceptualization, neurobasis, and applications. *Applied Neuropsychology: Child*, *6*(3), 204–211. https://doi.org/10.1080/21622965.2017.1317484

Ozdemir, B., & Gelbal, S. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. *Education and Information Technologies*, *27*(5), 6273–6294. https://doi.org/10.1007/s10639-021-10853-0

Paap, M. C. S., Born, S., & Braeken, J. (2019). Measurement Efficiency for Fixed-Precision Multidimensional Computerized Adaptive Tests: Comparing Health Measurement and Educational Testing Using Example Banks. *Applied Psychological Measurement*, *43*(1), 68–83. https://doi.org/10.1177/0146621618765719

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of Solution Behavior across Items in Low-Stakes Assessments. *Educational Assessment*, *24*(3), 189–212. https://doi.org/10.1080/10627197.2019.1615373

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1), 55–79. https://doi.org/10.1007/s11092-016-9248-7

Pitkin, A. K., & Vispoel, W. P. (2001). Differences Between Self-Adapted and Computerized Adaptive Tests: A Meta-Analysis. *Journal of Educational Measurement*, *38*(3), 235–247. https://doi.org/10.1111/j.1745-3984.2001.tb01125.x

Powell, Z. H. E. (1994). The Psychological Impacts of Computerized Adaptive Testing Methods. *Educational Technology*, *34*(8), 41–47.

Powers, D. E. (2001). Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the graduate record examinations (GRE©) general test. *Journal of Educational Computing Research*, *24*(3), 249–273. https://doi.org/10.2190/680W-66CR-QRP7-CL1F

Preckel, F., & Freund, P. A. (2005). Accuracy, latency, and confidence in abstract reasoning: The influence of fear of failure and gender. *Psychology Science*, *47*(2), 230.

R Core Team. (2012). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press.

Raven, J. C., Court, J. H., & Raven, J. (1988). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. H. K. Lewis.

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Reise, S., & Henson, J. (2003). A Discussion of Modern Versus Traditional Psychometrics As Applied to Personality Assessment Scales. *Journal of Personality Assessment*, *81*, 93–103. https://doi.org/10.1207/S15327752JPA8102_01

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement*, *63*(5), 791–808. https://doi.org/10.1177/0013164403251282

Rheinberg, F., Vollmeyer, R., & Burns, B. (2001). QCM: A questionnaire to assess current motivation in learning situations. *Diagnostica*, *47*(2), 57–66. https://doi.org/10.1026//0012-1924.47.2.57

Rios, J. (2021). Improving Test-Taking Effort in Low-Stakes Group-Based Educational Testing: A Meta-Analysis of Interventions. *Applied Measurement in Education*, *34*(2), 85–106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the Impact of Careless Responding on Aggregated-Scores: To Filter Unmotivated Examinees or Not? *International Journal of Testing*, *17*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Rios, J. A., & Soland, J. (2021). Parameter Estimation Accuracy of the Effort-Moderated Item Response Theory Model Under Multiple Assumption Violations. *Educational and Psychological Measurement*, *81*(3), 569–594. https://doi.org/10.1177/0013164420949896

Robitzsch, A. (2023). *sirt: Supplementary Item Response Theory Models.* (R package version 3.13-228) [Computer software]. https://CRAN.R-project.org/package=sirt

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules. R package version 4.1-4,* [Computer software]. https://CRAN.R-project.org/package=TAM

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Ruiz, P. E. (2009). Measuring fluid intelligence on a ratio scale: Evidence from nonverbal classification problems and information entropy. *Behavior Research Methods*, *41*(2), 439–445. https://doi.org/10.3758/BRM.41.2.439

Rulison, K. L., & Loken, E. (2009). I've Fallen and I Can't Get Up: Can High-Ability Students Recover From Early Mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83–101. https://doi.org/10.1177/0146621608324023

Sahin, A., & Anil, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

ŞahiN, M. D., & Gelbal, Prof. Dr. S. (2020). Development of a Multidimensional Computerized Adaptive Test based on the Bifactor Model. *International Journal of Assessment Tools in Education*, 323–342. https://doi.org/10.21449/ijate.707199

Santarnecchi, E., Emmendorfer, A., & Pascual-Leone, A. (2017). Dissecting the parieto-frontal correlates of fluid intelligence: A comprehensive ALE meta-analysis study. *Intelligence*, *63*, 9–28. https://doi.org/10.1016/J.IN℡L.2017.04.008

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 478–496. https://doi.org/10.1111/bmsp.12057

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance*, *15*(1–2), 187–210. https://doi.org/10.1207/s15327043hup1501&02_12

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-HornCarroll Model of Intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99–144). Guilford Press.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In *Contemporary intellectual assessment: Theories, tests, and issues, 4th ed* (pp. 73–163). The Guilford Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Segall, D. (1996). Multidimensional Adaptive Testing. *Psychometrika*, *61*(2), 331–354. http://dx.doi.org/10.1007/s11336-010-9163-7

Seo, D. G., & Weiss, D. J. (2015). Best Design for Multidimensional Computerized Adaptive Testing With the Bifactor Model. *Educational and Psychological Measurement*, *75*(6), 954–978. https://doi.org/10.1177/0013164415575147

Sheng, Y., & Wikle, C. K. (2007). Comparing Multiunidimensional and Unidimensional Item Response Theory Models. *Educational and Psychological Measurement*, *67*(6), 899–919. https://doi.org/10.1177/0013164406296977

Silm, G., Must, O., & Täht, K. (2019). Predicting performance in a low-stakes test using self-reported and time-based measures of effort. *Trames*, *23*(3), 353–376. https://doi.org/10.3176/tr.2019.3.06

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*(May), 100335. https://doi.org/10.1016/j.edurev.2020.100335

Simms, L. J., & Clark, L. A. (2005). Validation of a Computerized Adaptive Version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment*, *17*(1), 28–43. https://doi.org/10.1037/1040-3590.17.1.28

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education*, *9*(1). https://doi.org/10.1186/s40536-021-00100-w

Spearman, C. (1904). “General Intelligence,” Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201. https://doi.org/10.2307/1412107

Spielberger, C. D. (1972). *Anxiety: Current trends in theory and research*. Academic Press.

Stemler, S., & Naples, A. (2021). Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line. *Practical Assessment, Research, and Evaluation*, *26*(1), 11.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychological Review*, *125*(2), 218–244. https://doi.org/10.1037/rev0000088

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, d4002. https://doi.org/10.1136/bmj.d4002

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods 2010 42:4*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-talking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*(1), 6–26. https://doi.org/10.1016/S0361-476X(02)00063-2

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*(1), 8–9.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (6th ed).* Pearson.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation Matters: Using the Student Opinion Scale to Make Valid Inferences About Student Performance. *The Journal of General Education*, *58*(3), 129–151. https://doi.org/10.1353/jge.0.0047

Thompson, N. (2011). *Advantages of Computerized Adaptive Testing (CAT)*. Assessment System. https://assess.com/docs/Advantages-of-CAT-Testing.pdf

Tonidandel, S., & Quiñones, M. A. (2000). Psychological Reactions to Adaptive Testing. *International Journal of Selection and Assessment*, *8*(1), 7–15. https://doi.org/10.1111/1468-2389.00126

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320–332. https://doi.org/10.1037/0021-9010.87.2.320

van der Linden, W. J. (1999). Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion. *Journal of Educational and Behavioral Statistics*, *24*(4), 398–412. https://doi.org/10.3102/10769986024004398

van der Linden, W. J., & Hambleton. (1997). *Handbook of modern item response theory*. Springer.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*(4), 575–588. https://doi.org/10.1007/BF02295132

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting Answer Review and Change on Computerized Adaptive Vocabulary Tests: Psychometric and Attitudinal Results. *Journal of Educational Measurement*, *37*(1), 21–38. https://doi.org/10.1111/J.1745-3984.2000.TB01074.X

Wainer, H. (1993). Some Practical Considerations When Converting a Linearly Administered Test to an Adaptive Format. *Educational Measurement: Issues and Practice*, *12*(1), 15–20. https://doi.org/10.1111/j.1745-3992.1993.tb00519.x

Wainer, H. (2000). *Computerized Adaptive Testing: A Primer (Second Edition)*. Lawrence Erlbaum Associates.

Wang, C., Chang, H.-H., & Boughton, K. A. (2013). Deriving Stopping Rules for Multidimensional Computerized Adaptive Testing. *Applied Psychological Measurement*, *37*(2), 99–122. https://doi.org/10.1177/0146621612463422

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*(5), 295–316. https:// doi. org/ 10. 1177/ 01466 21604 265938

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models. *Psychological Methods*, *9*(1), 116–136. https://doi.org/10.1037/1082-989X.9.1.116

Ware Jr., J. E., Gandek, B., Sinclair, S. J., & Bjorner, J. B. (2005). Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. *Rehabilitation Psychology*, *50*(1), 71–78. https://doi.org/10.1037/0090-5550.50.1.71

Weiss, D. J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WAIS-IV and Clinical Validation of the Four- and Five-Factor Interpretative Approaches. *Journal of Psychoeducational Assessment*, *31*(2), 94–113. https://doi.org/10.1177/0734282913478030

Wigfield, A., & Eccles, J. S. (2000a). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J. S. (2000b). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wilhelm, O. (2005). Measuring Reasoning Ability. In *Handbook of Understanding and Measuring Intelligence* (pp. 373–392). SAGE Publications, Inc. https://doi.org/10.4135/9781452233529.n21

Wise, S., & Kuhfeld, M. (2021). A Method for Identifying Partial Test-Taking Engagement. *Applied Measurement in Education*, *34*(2), 150–161. https://doi.org/10.1080/08957347.2021.1890745

Wise, S. L. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L. (2014). The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees. *Journal of Computerized Adaptive Testing*, *2*(1), 1–17. https://doi.org/10.7333/1401-0201001

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An Application of Item Response Time: The Effort-Moderated IRT Model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Kuhfeld, M. R. (2020). Using Retest Data to Evaluate and Improve Effort-Moderated Scoring. *Journal of Educational Measurement*, *58*(1), 130–149. https://doi.org/10.1111/jedm.12275

Wise, S. L., & Ma, L. (2012). *Setting Response Time Thresholds for a CAT Item Pool: The Normative Threshold Method*. The 2012 annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In *High-stakes testing in education: Science and practice in K–12 settings* (pp. 139–153). American Psychological Association. https://doi.org/10.1037/12330-009

Wolf, L. F., & Smith, J. K. (1995). The Consequence of Consequence: Motivation, Anxiety, and Test Performance. *Applied Measurement in Education*, *8*(3), 227–242. https://doi.org/10.1207/S15324818AME0803_3

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of Performance, Test, Motivation, and Mentally Taxing Items. *Applied Measurement in Education*, *8*(4), 341–351. https://doi.org/10.1207/s15324818ame0804_4

Wolgast, A., Schmidt, N., & Ranger, J. (2020). Test-Taking Motivation in Education Students: Task Battery Order Affected Within-Test-Taker Effort and Importance. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/FPSYG.2020.559683

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Yang, C.-L., O'Neill, T. R., & Kramer, G. A. (2002). Examining item difficulty and response time on perceptual ability test items. *Journal of Applied Measurement*, *3*(3), 282–299.

Yao, L., Pommerich, M., & Segall, D. O. (2014). Using multidimensional CAT to administer a short, yet precise, screening test. *Applied Psychological Measurement*, *38*(8), 614–631.

Appendices

Figure A1.

Funnel Plot of Publication Bias Regarding the Overall Effect of Test Type on Anxiety and Motivation

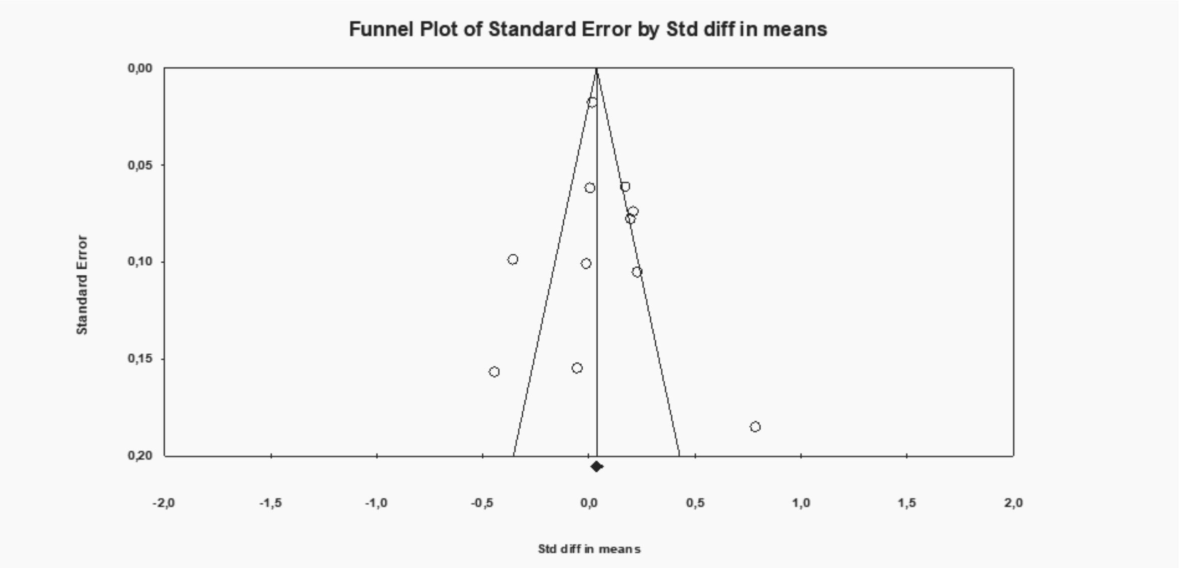
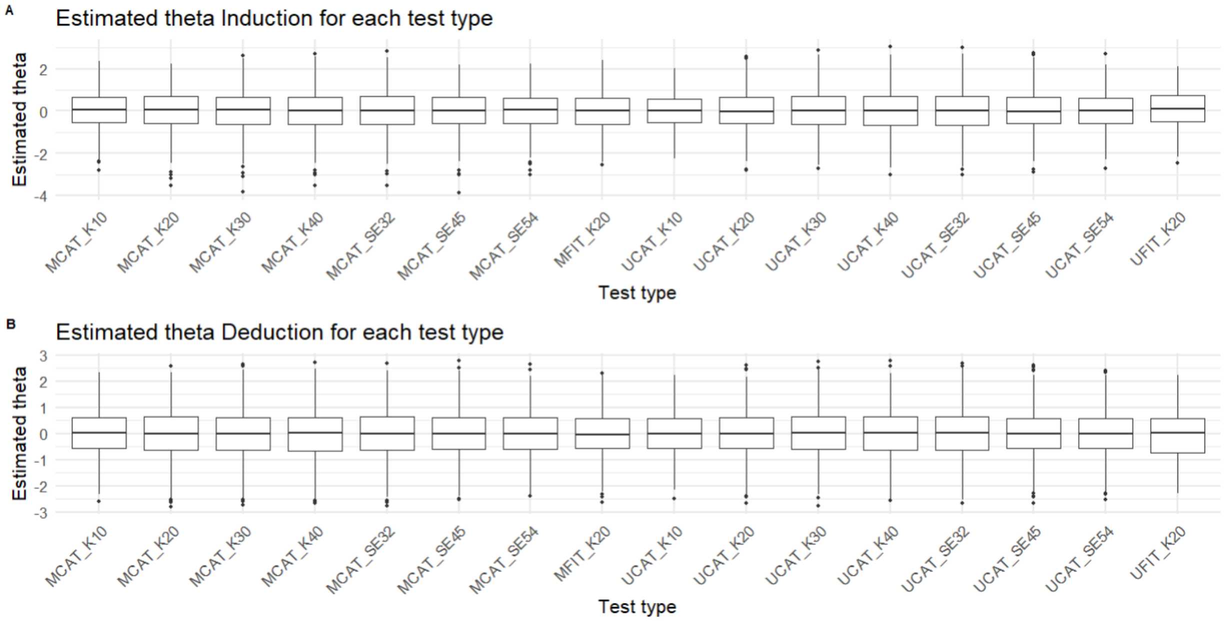


Figure A2.

Estimated thetas for each test type



Research instruments

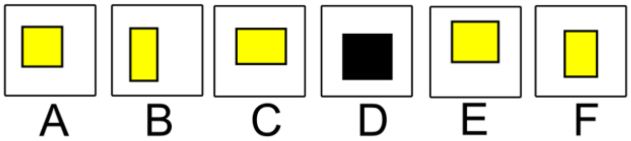
Multidimensional Induction-Deduction Tests

Subtest 1: Induction

Instruction

Five out of six images share similar characteristics based on a certain principle. Your task is to find one image that is most different from the others

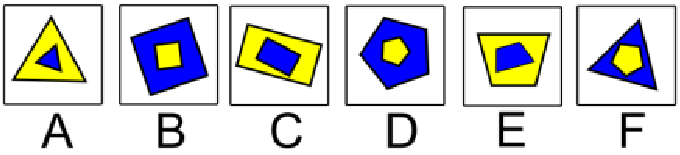
Example 1



Which image is different from the others?

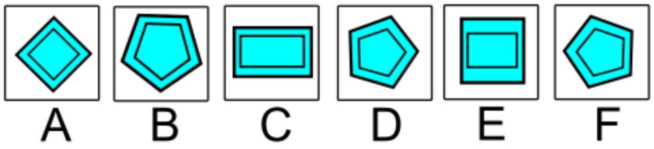
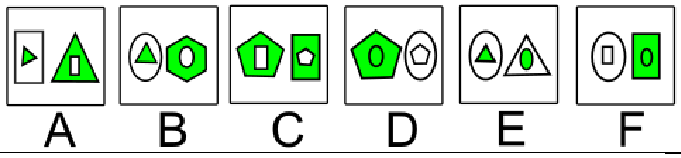
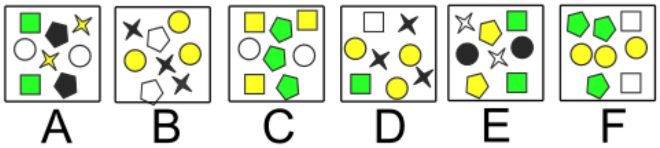
In the picture above, all the pictures have the same principle, that is, all the rectangles are yellow, except for picture D which is black. Thus, the answer to this question is D.

Example 2



Which image is different from the others?

Sampel items

Difficulty level	Sample item
Easy	
Medium	
Hard	

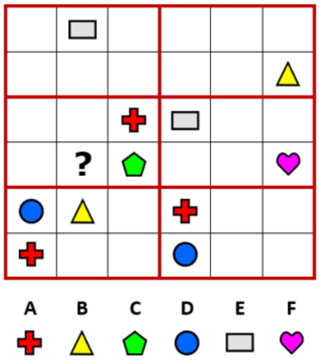
Subtest 2: Deduction

Instruction

Six unique shapes (✚ △ ⬟ ● ▭ ♥) should be placed into each row, column, and 2x3 grid without repeating any shapes within each row, column, or grid section; based on this principle, which shape can replace the question mark?

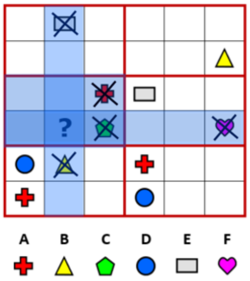
Example 1

Which shape can replace the question mark?



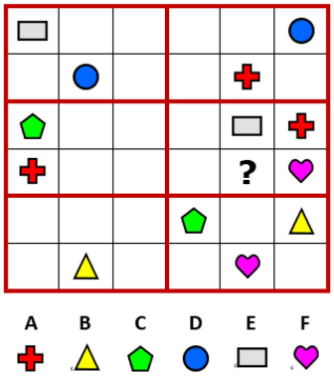
1. In the same column as the question mark, there are a shape ▭ and △, so the answer cannot be either of them.
2. In the same row as the question mark, there are shapes ♥and ⬟, so the answer cannot be either of them.
3. In the same 3x2 grid as the question mark, there is a shape✚, so the answer cannot be✚
4. Therefore, the shape most likely to fill in the question mark is ● (D)

Solution



Example 2

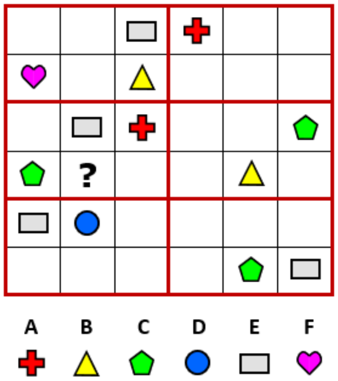
Which shape can replace the question mark?

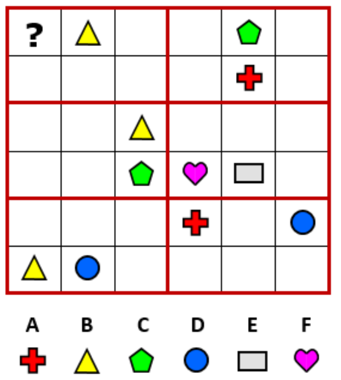
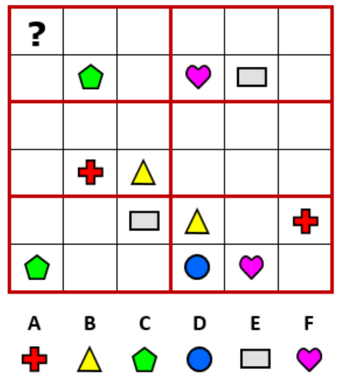


Please be noted!

- Each row, column, and 2x3 grid must contain six unique shapes.
- No shape is repeated within any row, column, or grid section.
- There is only ONE solution.

Sample items

Difficulty level	Sample items
Easy	

Medium	
Hard	

Note: Details about items, specification, and parameters for Multidimensional Induction-Deduction Tests is accessible at https://osf.io/h74wd/

Test-taking motivation instrument

Please indicate the extent to which the following statements are true for you, using the scale on bellow.

1 (strongly disagree), 2 (disagree), 3 (agree), 4 (strongly agree)

EF1. I did my best on this test.
EF.2 I worked with all items in the test without giving up, even when an item was difficult.
EF3. I felt motivated to do my best on this test.
EF4. I spent more effort on this test than I do on other tests.
EX1. I did well on this test.
EX2. Compared with other test-takers, I think I did well on this test.
IN1. I am very curious about the result I received on this test.
IN2. I looked forward to doing this test.
IN3. It was fun to do this test.
IN4. I learned something new by doing this test.

Note: EF = effort, EX = expectancy, IN = interest

State anxiety questionnaire

How well do the following adjectives describe your feelings during the part of the test that you just completed?

1=not at all, 2=a little, 3=moderately, 4=very much

1. Calm
2. Tense
3. Worried
4. Secure
5. Frightened
6. Anxious
7. at ease
8. nervous
9. content
10. jittery
11. pleasant
12. confused

Perceived performance

Out of 40 items, how many items do you think you answered correctly?

Post-feedback acceptance

Please indicate the extent to which the following statements are true for you, using the scale on bellow.

1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), 5 (strongly agree)

1. The feedback I received is an accurate evaluation of my performance.
2. I agree with the feedback provided.
3. It is hard to take feedback seriously.