EÖTVÖS LORÁND UNIVERSITY
FACULTY OF EDUCATION AND PSYCHOLOGY

Doctoral Dissertation Summary

Hanif Akhtar

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

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

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

Supervisor: Dr. Kristóf Kovács

Budapest, 2024

Doctoral candidate's list of publications

Publications related to the dissertation topic:

Akhtar, H., & Kovacs, K. (2024). Measurement precision and user experience with adaptive versus non-adaptive psychometric tests. *Personality and Individual Differences,* 225, 112675 https://doi.org/10.1016/j.paid.2024.112675

Akhtar, H., & Kovacs, K. (2024). Measuring process factors of fluid reasoning using multidimensional computerized adaptive testing. *Assessment.* Advance online publication. https://doi.org/10.1177/10731911241236351

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

Akhtar, H., & Kovacs, K. (2023). 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. (2023). Five Decades of Research on Computerized Adaptive Testing: A Bibliometric Analysis. [Manuscript submitted for publication]

Akhtar, H. (2022). Measuring fluid reasoning and its cultural issues: A review in the Indonesian context. *Buletin Psikologi*, 30(2), 276-288. https://doi.org/10.22146/buletinpsikologi.74475

Other publications:

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., & Firdiyanti, R. (2023). Predicting academic dishonesty based on competitive orientation and motivation: Do learning modes matter?. *International Journal of Cognitive Research in Science, Engineering and Education*. 11(3), 439–447. https://doi.org/10.23947/2334-8496-2023-11-3-439-447

Akhtar, H., & Silfiasari. (2022). A brief measure of self-reported cognitive abilities: Are males and females different?. *Testing, Psychometrics, Methodology in Applied Psychology*. 29(4), 475-493. https://doi.org/10.4473/TPM29.4.6

Akhtar, H. (2022). The pattern of test-taking effort across items in cognitive ability test: A latent class analysis. In D. G. Sampson, D. Ifenthaler, & P. Isaías (Eds.), *Proceeedings of the 19th International Conference on Cognition and Exploratory Learning in the Digital Age*. IADIS Press. http://doi.org/10.33965/celda2022_202207l021

Akhtar, H., & Sumintono, B. (2023). A Rasch analysis of the International Personality Item Pool Big Five Markers Questionnaire: Is longer better? *Primenjena Psihologija*, 16(1), 3–28. https://doi.org/10.19090/pp.v16i1.2401

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 FITs. 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

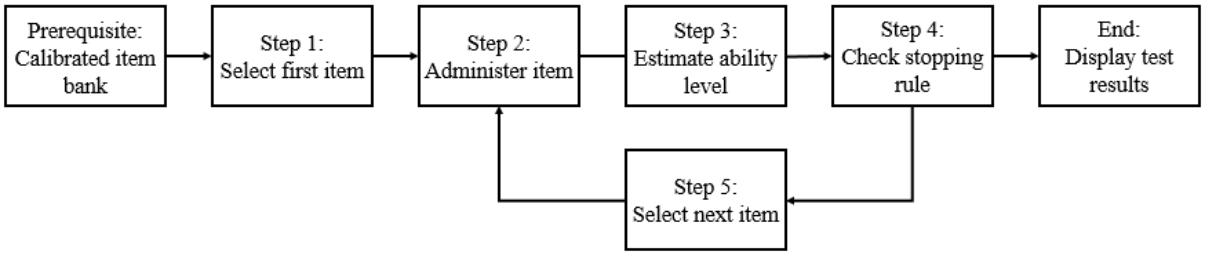
addition, several features of CAT, such as the inability to skip items and return to them later are disliked by examinees (Tonidandel & Quiñones, 2000).

Computerized Adaptive Testing

CAT is a methodology designed to increase the measurement efficiency of the tests. 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. Figure 1 depicts the basic procedures of CAT (adapted from Oppl et al., 2017).

Figure 1

The procedure of computerized adaptive testing (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), CATs aim 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. CAT enhances measurement efficiency and precision by selecting the most informative items for each examinee, thereby reducing the number of necessary items (Lunz et al., 1994; Wainer, 2000). It also minimizes floor and ceiling effects, as demonstrated by Ware et al. (2005), and provides an effective means to measure growth over time using a *test-train-retest* approach

(de Beer, 2013). Additionally, CAT increases test security and offers significant flexibility in item types and test administration.

Most CATs are grounded in unidimensional Item Response Theory (IRT) models, yet Van der Linden and Hambleton (1997) advocate for the use of multidimensional IRT (MIRT) models to capture interrelated abilities, enhancing CAT's efficacy. Multidimensional CAT (MCAT) employs MIRT to assess multiple correlated abilities simultaneously, offering advantages over 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 UCATs 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 (Wang & Chen, 2004).

Psychological aspects of CAT

There are special characteristics of CAT that differentiate this kind of testing from traditional fixed-item testing. *Firstly*, the test items given to the examinees are tailored to their level 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. *Second*, in CAT, when the item difficulty is equal to the estimated ability (theta), the probability of correctly answering is 50%, regardless of the examinees' ability. This differs from FIT, where the number of correct answers depends on the examinees' ability; higher ability examinees have more items correct. However, such a success rate might be perceived as too low, particularly by those with high abilities, potentially impacting their overall test experience. *Third*, several features common in FITs, such as the inability to skip or review previously answered items, are not well-implemented in CATs. *Fourth*, the number of correct answers in CAT does not solely determine the final scores. Even though two examinees have the same number of correct answers, their final scores might differ significantly, depending on which items they answered. In contrast, in FIT, the number of correct answers highly determines

the final scores. For these reasons, the test-taking experience in CATs might be different from that in FITs.

It has been frequently claimed that CAT provided a better experience than FIT (e.g., Thompson, 2011). The argument for the claim is that CAT will offer an appropriate challenge for each test-taker, so they are not administered items that are not too easy or too difficult. However, the support for the claim is mixed and not unequivocal. While some research indicates that CAT offers a better experience (e.g., Fritts & Marszalek, 2010), other studies suggest the contrary (e.g., Ortner et al., 2014), and additional research has not definitively shown a preference for either CAT or FIT (e.g., Ling et al., 2017). The probability of correctly answering of 50% in CAT is considered too low to retain test-taking motivation (Bergstrom et al., 1992). Additionally, test-takers often dislike certain features of CAT, such as the inability to skip items (Tonidandel & Quiñones, 2000).

The psychological impact of CAT on test-takers is an underresearched area, particularly regarding MCAT. MCAT differs from unidimensional CAT in significant ways. For instance, in MCAT, the item presented in a subsequent test is greatly influenced by the test-taker's performance on a preceding test. If a test-taker excels in the initial test, they face more challenging items at the beginning of the next test, a departure from unidimensional tests that typically start with simpler items to ease test-takers into the exam (Bergstrom & Lunz, 1999). Moreover, 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 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).

The Expectancy-value theory (Wigfield & Eccles, 2000) serves as a framework for understanding the motivation behind test-taking. This theory posits that individuals are more inclined to engage in testing when they anticipate success (high *expectancy*) and find *value* in the assessment. In expectancy-value theory, expectancy reflects the test-taker's perception of how they will perform. However, I believe that the same dimensions that drive expected performance are also relevant for evaluating past performance. The value components examined in this dissertation were interest and cost (i.e., anticipated *anxiety* and *effort* required to complete the task). Test taking-motivation can motivation can fluctuate based on the item's difficulty within the test (Wise & Smith, 2011). Given the inherent differences in

item adaptivity between CAT and FIT, these modalities may affect expectancy, interest, anxiety, and effort in distinct ways. Consequently, this study seeks to explore how these motivational components vary between CAT and FIT environments.

Another potential aspect differentiating MCAT from FIT could be test-takers' feedback acceptance, which is linked to perceived fairness (Tonidandel et al., 2002). 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 individual estimated their own performance (i.e., self-estimated performance) is a central aspect when comparing test-takers' experience in adaptive and non-adaptive tests.

Overview of the dissertation

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 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. 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?

This dissertation comprises four separate yet interconnected studies, each with distinct objectives. Briefly, the first study aims to synthesize the literature comparing psychological aspects of CAT and FIT. The second and third studies focus on developing a new multidimensional CAT for measuring general fluid intelligence (Gf) and evaluating its psychometric properties. The fourth study compares the psychometric and psychological aspects of CAT and FIT in an actual testing environment. Research Question #1 is explored through Studies 2, 3, and 4, while Research Question #2 is examined in Studies 1 and 4. Below is a concise overview of each study.

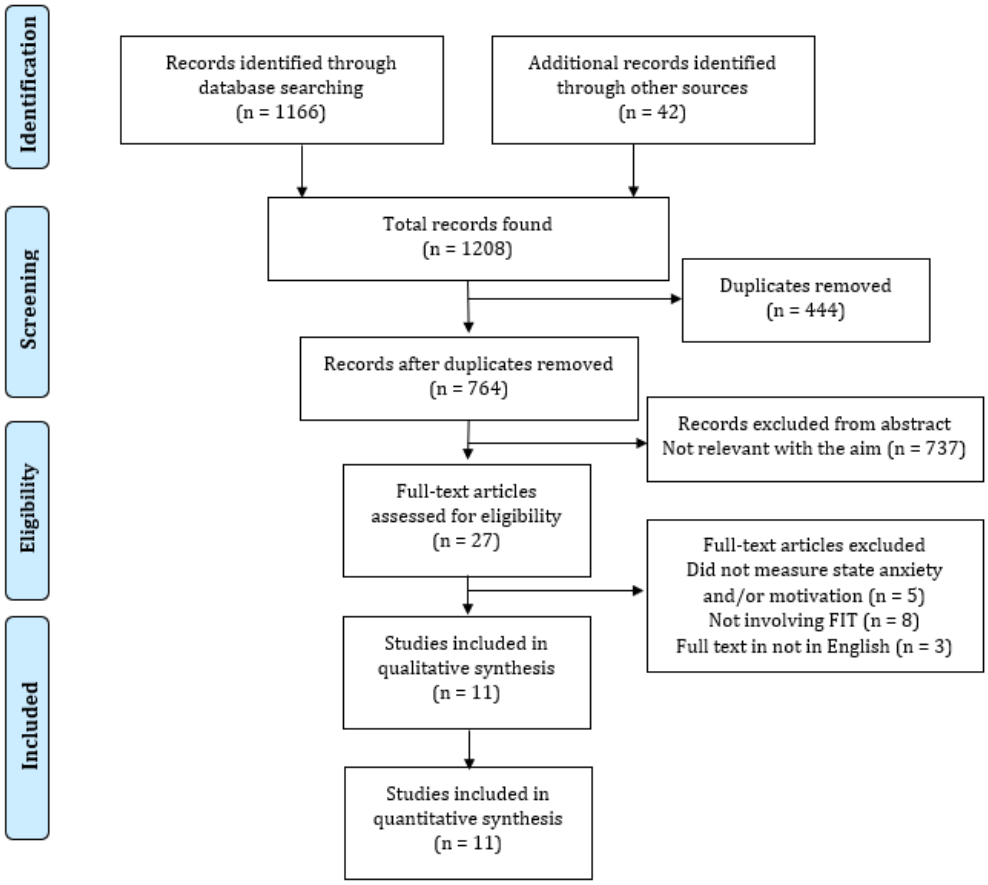
Study 1: The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis

Although many studies have been carried out on the psychometric aspects of 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. 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 supposed positive 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.

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. 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.

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. 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). Figure 2 illustrates the phases of article selection in accordance with PRISMA guidelines.

Figure 2
PRISMA Flowchart of the current study



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. 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. The overall effect was significantly heterogeneous, with a high proportion of observed variance (84%) reflecting real differences in effect size. Overall, there is no effect of test type on anxiety and motivation (k = 11, g+ = 0.06, p = .28). However, easier CAT (i.e., a CAT targeted at higher success rate), demonstrates a positive effect compared with a FIT (see Table 1).

Table 1

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

Study 2: Development of the item bank measuring process factor of fluid reasoning

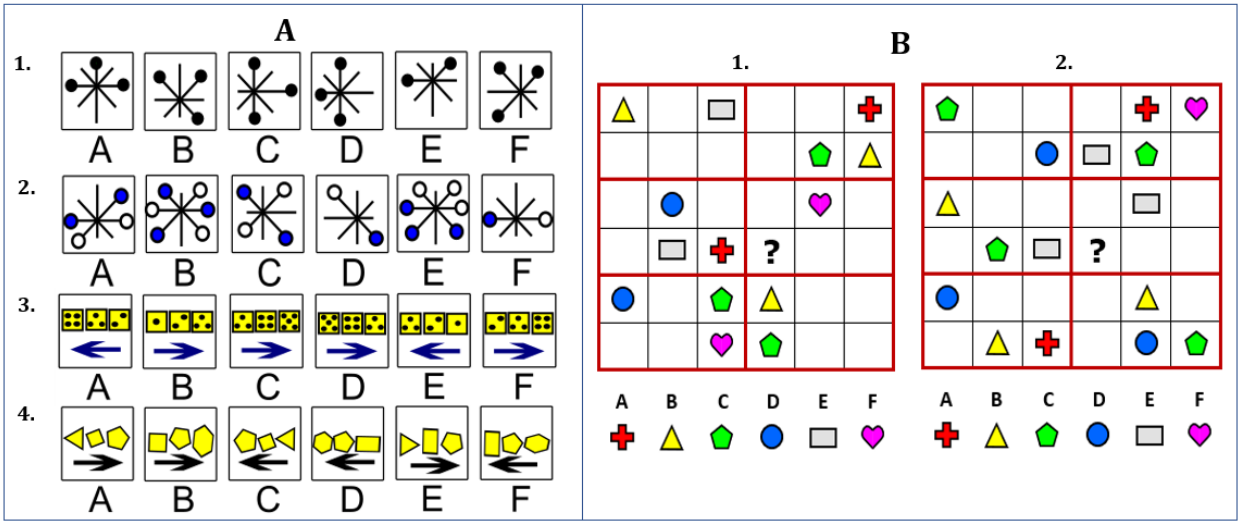
Although many Gf tests have been developed, there is a lack of figural tests measuring two narrow factors simultaneously. From a CHC perspective, to adequately represent a measure of Gf, it is essential to assess at least two narrow abilities (Flanagan et al., 2013; Schneider & McGrew, 2018). In addition, 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 (Wang et al., 2004). 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. 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. The purpose of this study was to create fluid reasoning items and investigate the psychometric properties of the item pool. The main issues we wanted to address in this study 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.

The total number of participants in this study was 2247 Indonesian adults. Data were collected in two waves (study 2a and 2b). In both studies tests were administered in an unproctored online environment using on the Psytoolkit platform (Stoet, 2017). The test administration was self-paced; participants used their own devices (PCs or laptops) to complete the test. The data from non-effortful test-takers were excluded as they might have negatively impacted estimates of item parameters (Rios & Soland, 2021). 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).

A total of 530 items were created. The test consisted of two subtests: the odd-one-out tasks (hereinafter called “induction test”) for measuring inductive reasoning, and Sudoku-like task (hf. deduction test) for measuring general sequential (deductive) reasoning. 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. Figure 3 depicts sample items of the tests.

Figure 3
Sample items of the induction (A) and deduction (B) test.



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. 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.

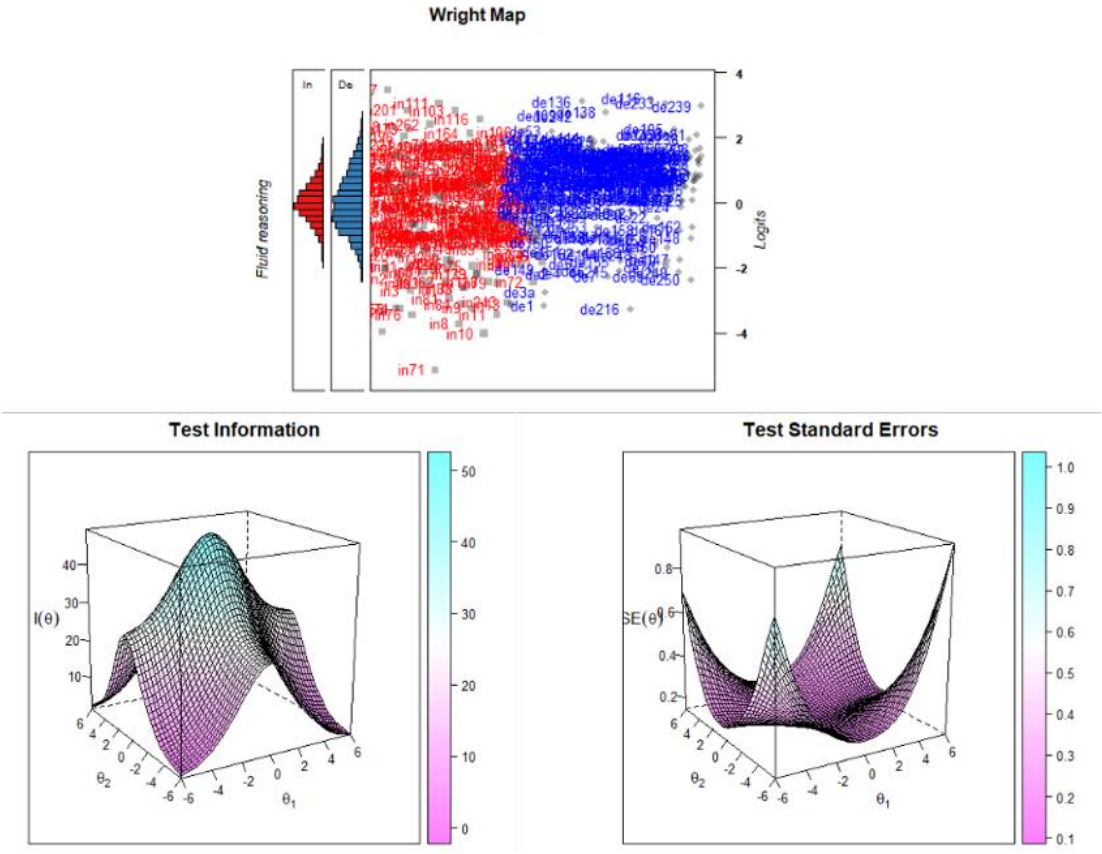
Model comparison was conducted to investigate which model fits the data better: unidimensional, separate-unidimensional, or multidimensional. Model comparison in shows that the multidimensional model has the lowest AIC and BIC values, indicating that this model fits the data better than the unidimensional 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.

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 4. 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 very high ability (e.g., θ = -2.0 or θ = 2.0), the standard error remains below 0.3.

Figure 4

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



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

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.

Study 3: A simulation study of MID-CAT

The purpose of study 3 was to conduct Monte-Carlo simulations to evaluate the potential performance of MCAT in comparison with separate-unidimensional CATs or non-CAT. The main issue we want to address in this study was to determine (a) whether MCAT was more efficient compared to UCATs or FIT, and (b) the number of items needed to be administered in high-stakes and low-stakes testing.

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 the previous study. 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.

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.

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.

As a stopping rule, the precision-based termination rules were utilized with three conditions: SE < 0.32 (equivalent of a reliability of 0.90[1]), 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 ware 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 UCATs 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).

The complete findings of the simulation study are shown in Table 2. As shown in Table 2, 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. 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 CATs or FITs.The benefits of MCAT over FIT was also varied for different test-takers with different ability levels. For example, when the test length was fixed at 20 items, MCAT resulted in lower SEs than MFIT, especially for test-takers with extreme theta scores (i.e., theta < -2 or > 2). 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

[1] 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.

Table 2

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.

Study 4: Psychometric and Psychological Evaluation of Multidimensional Computerized Adaptive Testing

This study aims to investigate the psychometric and psychological impact of CAT in the context of a multidimensional fluid reasoning test. Three questions were examined: (a) Is measurement precision different under MCAT and FIT? (b) Is test-taking experience different under MCAT and FIT? (c) Do participants show different patterns of rapid guessing behaviour under MCAT and FIT?

Measurement precision is operationalized as the standard error of the ability estimate (SE) after completing 20 items of each subtest. As studies have consistently found that CAT outperformed FIT in terms of precision, we expected SE in MCAT to be significantly lower than in FIT. 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 more objective measure of effort based on participants' response time to each question. RTE is based on the assumption that

unmotivated participants will answer items too quickly (i.e., before they have sufficient time to read and consider the correct answer). RTE makes it possible to investigate changes in participants' effort during a test session because response time data is available for each item.

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.

Participants registered in January 2023, and upon giving consent and providing demographic details, were randomly assigned to one of two groups. They received a link via email to complete the test in an unproctored online setting throughout February 2023. Participants in the MCAT group were informed about the adaptivity of item selection, whereas no such information was provided in the FIT group. After finishing the tests, participants filled out questionnaires on their experience, received their scores, and then completed a feedback acceptance scale.

Participants completed either MID-CAT or MID-FIT. MID-CAT is a multidimensional computerized adaptive test measuring two process factors of fluid reasoning: induction and deduction (see previous study for details). MID-FIT is a non-adaptive version of MID-CAT, consisting of 20 items for each subtest. To examine participants motivation, Test-taking motivation questionnaire (Knekta & Eklöf, 2015) was used. We used the three relevant subscales: *effort* (hf. SRE for Self-Reported Effort, to differentiate from RTE, Response Time Effort), *expectancy*, and *interest*. State Anxiety Questionnaire (Attali & Powers, 2010) was used to measure state anxiety during the test. Test acceptance was assessed using a three-item scale from Nease et al. (1999). Additionally, Self-estimated performance was measured with a single item: *"out of 40 items, how many items do you think you answered correctly?"*.

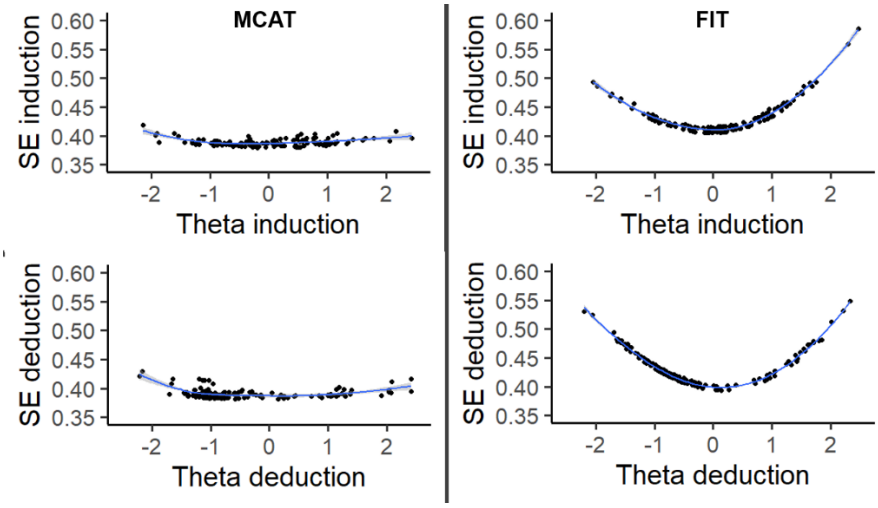
In addition to self-report, RTE (i.e., a time-based measure of effort) was used. The RTE index is based on the notion that answers provided below a particular time threshold

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 (RGB). 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.

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. Test-taking experience scores, which include expectancy, effort, interest, self-estimated performance, test acceptance, and anxiety, were moderately correlated with test performance. Specifically, there was a positive correlation with all variables except for anxiety, which showed a negative correlation. The two measures of test-taking effort, SRE and RTE, correlated weakly (r = 0.18).

Figure 5

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



The analysis of SE using mixed Anova revealed a significant main effect of test type ($F(1, 284) = 273.31$, $p < 0.001$, $\eta p2 = 0.490$): the MCAT had lower SE than the FIT. In addition, the main effect of time (SE1 vs SE2) was also significant ($F(1, 284) = 9.30$, $p = 0.03$, $\eta p2 = 0.03$), showing that participants' SE increased in the second subtest. The relationship between participants' abilities (thetas) and SE is shown in Figure 5.

The effects of test type and ability on six dependent variables (expectancy, SRE, interest, anxiety, self-estimated performance, and acceptance) were examined using ANCOVAs. 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. The analysis of RTE using 2X2 mixed Anova revealed a significant main effect of test type ($F(1, 283) = 12.10$, $p = 0.005$, $\eta p2 = 0.041$) and ability ($F(1, 283) = 76.35$, $p < 0.001$, $\eta p2 = 0.212$): taking the MCAT resulted in higher effort than taking the FIT. In addition, the main effect of time (T1 vs T2) was also significant ($F(1, 283) = 3.88$, $p = 0.049$, $\eta p2 = 0.013$), showing that participants' effort decreased in the second subtest.

The proportion of rapid guessing behavior 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. Spearman's correlation between item position and proportion of rapid guessing behaviour in the MCAT group was $r = 0.64$, $p < 0.001$, while in the FIT group, $r = 0.65$, $p < 0.001$.

Discussion

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. The following is an overview of the main research questions and empirical findings.

Research Question 1 examines whether measurement precision varies between adaptive and non-adaptive testing modalities, revealing significant distinctions. Adaptive tests maintain consistent precision across all ability levels, whereas FIT typically concentrates on median ability levels, resulting in variable precision across the ability

spectrum. This dissertation corroborates findings from both simulation studies and empirical data, highlighting the psychometric superiority of MCAT over unidimensional CAT and FIT. MCAT demonstrates enhanced measurement efficiency, either providing greater precision within set test lengths or requiring shorter lengths to achieve a predetermined level of precision, particularly under conditions employing a precision-based stopping rule. Notably, while the efficiency and reliability of MCAT are evident, the type of test—MCAT, UCAT, or FIT—does not significantly influence the estimation of test-taker abilities, ensuring that ability estimates remain consistent across different testing formats.

Research Question 2 delves into the comparative test-taking experiences under adaptive and non-adaptive formats, employing a combined approach of meta-analysis and direct empirical evaluation. The meta-analysis did not identify any significant differences in test-taker motivation or anxiety between CAT and FIT, although individual study outcomes varied based on specific testing conditions and measures. Direct comparison between MID-CAT and MID-FIT revealed negligible differences in test-taker expectancy, interest, and anxiety levels, suggesting similar acceptance rates for both testing types. At the same time, FIT resulted in a more optimistic evaluation: when completing FIT, participants believed they had answered more items correctly than when completing CAT. Yet response time analysis indicates that participants invested more effort when working in CAT. Rapid guessing increased as the test progressed in both CAT and FIT, particularly in the FIT condition with the most difficult items.

This dissertation provides two key contributions: a multidimensional Gf test – MID-CAT – that is flexible, efficient, and accessible, and empirical evidence highlighting the advantages of CAT over FIT regarding psychometric properties and test-taking experience. The MID-CAT, with its extensive item bank, stands out as a novel resource in the adaptive testing field. To our 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, script analysis, test specification, and item properties, are available in the online repository (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. 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. 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.

Future research directions emerging from this dissertation suggest several areas for further exploration, including modifying the CAT algorithm to select easier items, which could enhance the test-taking experience by adjusting the item difficulty level to optimize psychological outcomes. Future studies should also investigate the impact of CAT in high-stakes assessments to determine if the findings extend beyond low-stakes contexts and explore how varying test features might influence test-taking motivation. Additionally, as the current study was conducted in a region where CAT is novel, ongoing research is needed to track how familiarity with CAT might change test-takers' attitudes over time, especially as CAT becomes more integrated into various educational and assessment settings.

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

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

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

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

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

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

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

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

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

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

Fritts, B. E., & Marszalek, J. M. (2010). 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

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

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

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

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

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

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

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*, 715–738. https://doi.org/10.1111/j.1744-6570.1994.tb01573.x

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

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

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

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

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

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.

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

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

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. (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.

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

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

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., & 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

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

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., & 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

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