Summer 8-2021

Exploring the Effect of Practice on Adverse Impact using a Measure of Cognitive Ability

Derrick McDonald

Exploring the Effect of Practice on Adverse Impact using a Measure of Cognitive Ability

by

Derrick McDonald, M.A., B.S.

A Dissertation Presented in Partial Fulfillment
of the Requirements of the Degree
Doctor of Philosophy

August 2021

LOUISIANA TECH UNIVERSITY

GRADUATE SCHOOL

June 21, 2021
Date of dissertation defense

We hereby recommend that the dissertation prepared by

Derrick McDonald M.A., B.S.

entitled **Exploring the Effect of Practice on Adverse Impact using a Measure of Cognitive Ability**

be accepted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Industrial/Organizational Psychology

Tilman Sheets
Supervisor of Dissertation Research

Donna Thomas

Donna Thomas
Head of Psychology and Behavioral Science

Doctoral Committee Members:
Jane Jacob
Frank Igou

Approved:

Don Schillinger

Don Schillinger
Dean of Education

Approved:

Ramu Ramachandran

Ramu Ramachandran
Dean of the Graduate School

ABSTRACT

Cognitive ability testing and cognitively loaded measures in employee selection have been utilized, developed, and improved upon for over a century; however, it is not without its faults. Two major problems facing cognitive ability tests are their tendency to produce adverse impact when used in selection systems and the costs associated with creating a well-constructed measure. This paper proposed that Automated Item Generation (AIG) may provide a solution to both of those problems. The first study focused on the construct validation of the Katyem Object Tracking Assessment (KOTA), a nonverbal AIG measure of fluid intelligence, that would allow test takers to practice as much as they want, comparing it to the emotionality portion of the HEXACO and to the short form of the Hagen progressive Matrices. After cleaning and removing careless responders from the sample of 458 participants, 89 remained, far below the 200-participant sample size needed to find a medium effect size. The data were analyzed using the Multitrait-multimethod matrix. Support for the hypotheses were not found. Afterward, the measure was used in a second study to determine if allowing participants to practice reduces adverse impact in a hypothetical employment situation. After cleaning and removing careless responders from the sample of 172 participants, 56 remained and were analyzed using two-way repeated measures ANOVA, Chi-squared goodness of fit test, Fisher's Exact test, and the four-fifths rule. The hypotheses concerning group differences and practice effects were unsupported, however, the hypothesis for the KOTA not having adverse impact was supported. Directions for future research are also provided.

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It is understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author D. ______________________

Date 7/26/2021

DEDICATION

This dissertation is dedicated to everyone, family, friends, strangers, peers, and mentors, who helped me on my journey here by following my ridiculous methodology and taking part in my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Charletta, Derrick, Victoria, Tilman, Frank, Jane, Clover, and everyone else who took part in this study. I literally could not have made it to this point without your contributions. From the bottom of my heart, thank you!

CHAPTER 1

INTRODUCTION

Intelligence testing has been around for over a century with the contributions of William Stern (e.g., the intelligence quotient or IQ) standing at the forefront (Stern, 1912). Researchers, academics, practitioners, and lay-people all benefit in various ways from the study of intelligence; whether it is using a test to hire the most qualified candidate for a job, or to see where one stands amongst their peers. Intelligence testing, also sometimes referred to as cognitive ability testing, is not a perfect science; different researchers have provided different operational definitions and measures, some of which I will discuss later in this document. There are several areas within the domain of intelligence testing that could be improved upon, one of which is the adverse impact associated with the use of such tests in employee selection, others are test security.

Historically, the definition of cognitive ability and ways to measure it has changed over time. The first modern form of what we would consider an intelligence test was developed by Binet and Simon in 1904, the goal of which was to distinguish between the cognitive capabilities of children. Over time many other researchers would modify or improve upon their works, redeveloping it for different languages or adding sections. Based on the works of Binet & Simon, the first mainstream use of mental testing was done by the United States military with the Army Alpha test developed in World War I with the goal of classifying applicants by their *mental standings* and to aid in the selection of competent officers (Yoakum & Yerkes, 1920). In 1955, Weschler developed multiple intelligence scales, the Weschler Adult Intelligence scale, the Weschler intelligence scale for children, and the Weschler preschool and primary scale of intelligence, based on his dissatisfaction with the limitations of the Stanford Binet test, such as its focus on children. These groups of tests are scored by comparing the score of the test taker to that of others within their age group and set the precedent for scoring intelligence tests in the modern day.

One of the more recent advancements in the field of assessment came about as a result of the ease and prevalence of computers. Computer adaptive testing (CAT) allows for a more dynamic testing scenario in which the test takers' performance influences the subsequent items they receive, thereby presenting a much more accurate depiction of their cognitive ability (Wainer, 2002). With each passing generation, what we defined as intelligence and how we measured it improved with the introduction of new technologies and methodologies, e.g., multiple-choice formats, item banks, non-verbal items, factor analytic theory, CAT and, most recently, Automated Item Generation (AIG) (Gierl et al., 2012; Wainer, 2002). AIG is poised to be the next important tool in test development given that the methodology combats many of the issues in testing (Gierl et al., 2019).

A major problem with cognitive ability tests that are used in employee selection, where appropriate, is the potential of adverse impact (Hough et al., 2001). Adverse impact, outlined in Title VII of the Civil Rights Act of 1964, occurs when members of a protected group score differently on a selection assessment than members of another group. For example, asking the question “Have you ever been convicted of a non-violent felony” on a selection test and assigning a score of one for those answering yes and zero for those answering no. A question such as this may show noticeable differences in responses across racial lines, as a greater proportion of whites may respond no to it than blacks (Hough et al., 2001). The problem arises when such a question is used for selection; if people who answered no are not extended the job offer and the largest proportion of people who say no are of a specific race, then the company may be opening itself up for legal action on the part of the applicants. The burden of proof then falls on the company to provide evidence that its questions are job relevant. The company could, in response, provide evidence that would suggest that the questions that they ask address bona fide occupational qualifications (BFOQ) and that there is no way to achieve the same purpose in an alternative way (EEOC, 1981). By establishing that the question addresses a BFOQ, the company would then have the leverage to win the court case. But just how frequent are these problems? A study by Schuster & Miller (1981) found that only two of 151 fortune 500 companies had applications that were completely fair, meaning there was no threat of the applications causing adverse impact. Thirty-eight percent of the firms had more than 10 items that could be deemed inadvisable by a court, indicating that this is a prevalent problem. Questions like “have you ever been arrested for a misdemeanor or felony” appeared 64.7% of the time. A question such as this is much more likely to have some adverse impact on minorities applying to those positions. Evidence for adverse impact may be obtained through various statistical processes such as a Fisher's exact test or through the 4/5ths rule.

Cognitive ability testing, which is one of the most consistent predictors of job performance, has also been shown to have adverse impact (Schmidt & Hunter, 1998). A meta-analysis by Roth et al. (2001), looked at ethnic group differences in cognitive ability using both educational and employment tests. They, along with Hunter and Hunter (1984), found evidence that there is approximately one standard deviation difference in the means of cognitive ability scores between blacks and whites. Roth et al. (2001); however, challenged the findings of the generally accepted large effect size for the black-white cognitive ability differences since most of the findings are from limited narrative reviews. The authors point out that while there is evidence to support the one standard deviation difference, much of that research fails to account for many of the complexities associated with the measurement of the group differences. The authors explored moderators of the relationship, sampling errors, and study design and found that issues like job complexity and which areas of cognitive ability were measured influenced the validity of the results.

Along these same lines, Jensen (1998) posits that there is approximately a 1.2 standard deviation difference in population scores of IQ tests between Blacks and Whites and that little change in the size of that gap had occurred over the 80 years prior. However, there is evidence that such differences are not independent of testing situations. Stereotype threat is one such factor that influences individual's performance in a situation, if a person is reminded of a negative stereotype regarding the performance of the group they belong to while working on that task their performance tends to reflect that of the group as a whole (Steele & Aronson, 1995). Brown and Day (2006) found that the stereotype threat associated with information regarding blacks and whites having cognitive differences can influence their comparative scores. The authors found no significant difference in Black and White scores on the Ravens Progressive Matrices in conditions where the test was not presented as an IQ test (High threat) or a standard test

(Standard threat) but rather as a puzzle on which the researchers wanted opinions (Low threat). In fact, the Black scores in the low threat condition were not significantly different from the White scores in the conditions in which they performed their best. Based on the works by Roth et al. (2001), there is evidence that in many employment settings that these group differences are less than 1 standard deviation, typically moderated by study design considerations like the exact construct being measured as well as if the sample consists of applicants or job incumbents.

One particular and common method for improving a participant's score on a measure is allowing them to practice or retake the measure. The effects of practice on cognitive ability scores have a long history. Anastasi (1934) reported practice effects between $d = 0.2$ to $d = 1.1$. A meta-analysis on practice effects by Kulik et al. (1984) found that practice can indeed raise scores on achievement or aptitude tests, the magnitude of which are influenced by the ability level of the subject, the number of practice tests, and whether the next administration of a test was parallel or identical. The highest increases are typically between the first and second administrations (Hausknecht et al., 2002). Exposure to the test materials or similar items increase the participants' scores and are influenced by a number of moderators (e.g., test anxiety, rote memorization of answers, new cognitive strategies) (Kulik et al., 1984; Messick & Jungeblut, 1981). There are several factors that have been found to influence likelihood to retest a particular measure, with previous score, age, and gender all having the strongest influences: if a previous score is satisfactory, the participant is less likely to retest; if the participant is older they are more likely to retest; and African-American females are more likely to retest than African-American males (Boyte-Eckis et al., 2018).

Additionally, Dixon et al. (1993) found evidence that between younger and older groups on a cognitive task involving handwriting speed, familiarity with the task being administered reduced group differences. The authors also found evidence that there are differential increases in scores as a result of uninstructed practice, with older adults improving faster than the younger adults. Hausknecht et al. (2007) found evidence that the strength of practice effects has a positive relationship with the amount of time spent practicing via formal coaching. A meta-analysis by Trahan et al. (2014) on the magnitude of the Flynn effect found that mean scores on IQ tests such as the Stanford-Binet and Wechsler increase by about 3 points per decade, providing additional evidence that over time IQ scores tend to increase.

A common finding in the practice literature is that the greatest increase in performance is between the 1st and 2nd administration, with each subsequent administration providing diminishing returns (Falleti et al., 2003). Their specific research explored practice effects across brief intervals. In their meta-analysis of coaching and practice effects for cognitive ability tests, Hausknecht et. al (2007) reported increases of $d = .25$ for second administrations and $d = .20$ for the third administration. There is also evidence that the amount of time between attempts and score increase on spatial reasoning tests are positively related, implying the potential impact of learning (Olenick et al., 2016). Bors and Vigneau (2003) provide additional evidence that score increases due to practicing or retaking spatial reasoning tests are a result of learning effects rather than rote memorization of item responses or strategies to respond to more items. Several other authors have provided hypotheses regarding the reason for change due to practice that are not construct-related such as test anxiety, stereotype threat, and, relatedly, a lack of familiarity with the particular test (Anastasi, 1981; Lievens et al., 2007; Matton et al., 2009; Reeve & Lam, 2007). Additionally, researchers have concluded that score gains due to practice were also due to strategy refinement e.g., improvements in problem-solving strategies (Hayes et al., 2015; Lievens et al., 2007; te Nijenhuis et al., 2007). Indeed, results from a longitudinal study by Estrada et al. (2015) also supports the idea that practice leads to such changes in test-taking strategy. These findings lend credence to the idea that improvement in scores is due to factors other than changes at the construct-level.

Given that there are socio-economic and racial differences regarding opportunity and familiarity with test taking (Grodsky et al., 2008), the associated development of problem-solving strategies related to tests of cognitive ability along with the anxiety associated with such tests may also differ between groups. Such difference could explain at least part of the observed differences in scores between groups. If such is the case, providing test takers with the opportunity to develop and refine their test-taking strategies should aid in reducing score differences between groups.

The validation of an original non-verbal test of cognitive ability that has a vast potential item bank that would allow participants to practice is one outcome of the study that may benefit other researchers. Drawing from that, if similar auto-generated tests are designed and validated, perhaps allowing participants to practice measures will become more commonplace, without running into the age-old problems of test security and limited item banks. In addition, the allowance of practice may have an impact on reducing differences between minorities and non-minorities. The purpose of this study is to see if such a reduction is possible using an AIG approach to assessing cognitive ability.

Review of Literature

Cognitive Ability

While in layman's term's cognitive ability is often seen as a *know-it-when-you-see-it* phenomenon, researchers have spent years trying to pin down what it is and how to measure it. Binet & Simon (1916) used words like judgment, initiative, adaptation to circumstances, and "practical sense" when they defined it. Common themes found across definitions include problem-solving, and reasoning (Sternberg et al., 1981). Alternatively, Gottfredson (2004) forwards the idea that cognitive ability is a person's ability to reason, think abstractly, problem solve, understand complex ideas, plan, and integrate new information. Legg & Hutter (2007), further point out that the definition of intelligence, used interchangeably here with cognitive ability, is controversial and even experts in the field may disagree, hence their compiling of over 70 different definitions. One can tell a lot about what a construct is by identifying what it isn't. The same article differentiates those two aforementioned common traits with traits that are not related to intelligence, namely personality traits with examples revolving around dishonesty, unreliability, and apathy. Based on this, intelligence can at least be distinguished as a construct that is orthogonal to personality.

The idea of separating intelligence from other constructs to help define it finds its origins in Spearman's (1904) then-new methodology of factor analysis, which he used to help identify the components of intelligence. Spearman discovered that school children's scores on a wide variety of seemingly unrelated subjects were positively correlated, which led him to propose the existence of a general mental ability that underlies human cognitive performance.

Using a rudimentary form of factor analysis on various measures of intellect, Spearman (1904) developed his Two-Factor Theory of Intelligence. The two factors he distilled from the data were *g* and *s,* a general and a test specific factor respectively. The *g* factor contributes to all cognitive processes, while on the other hand, the *s factor* contributes to how one scores on a given measure using their mechanical, logical, or arithmetical abilities for example. Most studies focus on *g* and it is often a go-to definition when it comes to intelligence (Gottfredson, 2002). Some tests like the Ravens Progressive Matrices and the WAIS III are more *g*-loaded than others (Jensen, 1998). Tests that require problem-solving and reading comprehension tend to be more g-loaded than those that just require simple computation or spelling (Jensen, 1992). Jensen (1998) provides evidence that *g* emerges across all mental test batteries even when people of different demographics are tested.

One key difference between *g* and *s* is how well tests that measure them overlap or diverge. If a test correlates in a strong positive manner with other cognitive ability tests it would be considered to have higher levels of *g* saturation. The more *g* saturated a test is the more it taps into a person's level of *g*. The *s* factor is more prevalent when a test doesn't correlate strongly with other tests of cognitive ability, so a person's level of *s* will have a greater influence on their score. An example of a high *g* saturated test is the Raven's Progressive Matrices, a nonverbal abstract reasoning test where subjects extrapolate the next object in a matrix given the other information contained therein (Spearman, 1938). The Raven's Progressive Matrices Test has been used around the world for a litany of purposes, including for use in the armed services since item translations were not necessary.

A rival theory emerged as a counter to the idea of *g* and *s* with the premise that there are seven primary mental abilities. Recall the section on the subjectivity of factor analysis, in order to make the data more interpretable researchers may use a series of orthogonal or oblique rotations. If researchers theorize variables to be correlated, they will use oblique rotations, and if they theorize them to be uncorrelated, they will use orthogonal rotations (DeVellis, 2003). The rotated factors are mathematically equivalent to their predecessors, and they are simply more interpretable post rotation (Gorsuch, 1990). By using this methodology, Thurstone (1938) initially concluded that rather than one superordinate factor, there were seven. These primary mental abilities were induction, perceptual speed, associative memory, number, verbal comprehension, word fluency, and space. According to Thurstone, the core of cognitive ability was a combination of those seven rather than only *g*. However, replications of their study showed that the factors were more correlated than they previously believed. This led them to come to terms with the idea that *g* was much more prevalent than they originally theorized.

Cattell (1941) posited that intelligence is made up of two, rather than seven primary mental abilities, calling them crystallized and fluid. Cattell viewed crystallized intelligence (*gc*) as the accumulation of knowledge from prior experiences and learning (Schneider & McGrew, 2012). It is based on information and our own experiences. An example of crystallized intelligence would be naming all 50 states in alphabetical order. On the other hand, he saw fluid intelligence as being able to solve problems and think abstractly outside of what a person has learned (Schneider & McGrew, 2012). The development of fluid intelligence (*gf*), with its focus on problem-solving in ways independent of knowledge that has been previously gained, typically peaks in young adulthood, while crystallized intelligence increases gradually until late adulthood before it begins to decline (Ashton et al. 2005).

A student of Cattell, John Horn, improved upon the theory, once again using factor analysis, to add several additional factors. The new eight-factor model, coined the Cattell-Horn theory, included visual perception, speed of processing, short- and long-term memory, auditory processing, reaction time and decision speed, reading and writing, and quantitative abilities (Horn, 1991). Within this new model, crystallized and fluid intelligence were viewed as overarching categories encompassing the other eight factors, and each of the eight factors can be assessed through individual tasks (Horn, 1991). The eight factors of this model specifically exclude *g,* making it easily distinguishable from another similar theory inspired by the same line of research.

Inspired by the research results of Thurstone, Cattell, and Horn, Carroll (1993) began his work on what would later be called the three-stratum theory in which he created a three-layered model of cognitive ability where the correlations of previous layers are accounted for by the higher layers with *g* being the top layer. It can be described as a melding of Spearman's model of *g* with Cattell and Horn's theory of fluid and crystallized intelligence. The hierarchy he chose for the stratum goes from general at the top in stratum III to broader abilities in stratum II and ending with specific abilities in stratum I. He placed *g* in stratum III since his research provided evidence that it accounts for the correlations in the next stratum down. Stratum II contained fluid and crystallized intelligence along with broad retrieval ability, auditory perception, cognitive speediness, processing speed, visual perception, and general memory and learning. The abilities in stratum II can be measured with different tasks and involve different processes from each other. The final stratum representing more specific factors than stratum II, 69 factors to be exact, that "…represent greater specializations of abilities, often in quite specific ways that reflect the effects of experience and learning, or the adoption of particular strategies of performance" (Carroll, 1993, p. 634). Similar to a lot of concepts in psychology, while they may not be completely orthogonal to each other, they do have enough of a difference to be differentiated from one another reliably (Keith & Reynolds, 2010). This relationship allows them to be stacked in this hierarchical manner.

This series of theories and studies culminated in a combined Cattell-Horn-Carroll theory (Flanagan, 2000), in which the strengths of each theory were combined to bring about what is currently the most influential model of modern cognitive ability to date. A major strength of the theory and what may account for its staying power is that it is continuously updated based on new research (Flanagan, 2000).

Practical uses for cognitive ability

Several researchers have found evidence that cognitive ability has a high predictive validity for job performance, although there are several moderators such as complexity which will be discussed later, making it one of the most consistent tools for employee selection (Schmidt, 2014; Schmidt & Hunter, 1998; Srikanth, 2020). Their findings, however, may be a result of range restriction due to researchers only having access to the data of those applicants who are hired; Hunter et al. (2006) estimated that the coefficient is actually closer to .6 when that range restriction is corrected. Both numbers are actually averages across multiple jobs and industries. There are several factors that influence the relationship between cognitive ability and job performance.

Familiarity and practice with a task have a positive influence on this relationship, with more practice and familiarity increasing the correlation; these findings have been replicated across various tasks (Fleishman & Fruchter, 1960; Fleishman & Hempel, 1955). Levels of complexity also play a major role in the relationship between performance and cognitive ability (Schmidt, 2002). Breaking the complexity of jobs into high, medium, and low levels of complexity, the validity coefficient is .57 for high complexity jobs, .51 for medium complexity jobs and .38 for low complexity jobs (Hunter & Hunter 1984). Schmidt (2002) presented additional evidence about the validity generalizability of cognitive ability with results that suggest that it varies similarly across completely different jobs, like cooks and welders, as it does within a single job; indicating that while complexity as a whole affects the validity, individual tasks have less of an impact. The individual tasks may vary across a job, but the overarching ability to perform those activities all relies on an employee's cognitive ability.

The use of validity coefficients in selection becomes more relevant when examining the utility of cognitive ability. If there is a way to discriminate between potentially high and low performers, then an organization can place themselves in a better position to reach their goals. Blatter et al. (2011) estimated that the average hiring costs for a position can range anywhere from 10-17 weeks of wage payments for that position, and that amount can change depending on how many people need to be hired. Therefore, incremental changes in validity can have a big influence on costs.

There are several moderators that influence the relationship between intelligence and job performance (e.g., practice, familiarity, job complexity), but there are also several mediators. One such mediator that influences the relationship, as found by Borman et al.

(1993), is the opportunity to obtain extra job experience. Oftentimes it is the employees who have proven they are capable that are given or seek out additional responsibilities and opportunities and through the process they gain job knowledge which helps increase their performance. These conclusions were also supported by Schmidt et al. (1986) when they found that job knowledge, which is influenced by job experience, has a direct influence on work sample performance. The more efficiently a person can gain job knowledge, the better their performance will be. One way to gain job knowledge is through successful training which, as Schmidt & Hunter (1998) found, is strongly influenced by cognitive ability. In other words, one-way job performance is influenced by cognitive ability is through the route of efficiently synthesizing and utilizing the information gained through training.

A high level of cognitive ability is not necessary for all positions, there may be an ideal level depending on the specific position after which it is unnecessary. In regard to the validity of cognitive ability tests, the most valuable instrument would also be the most valid instrument for selection in a given job. This validity coefficient changes in response to job complexity, as job complexity increases so too does the predictive validity of cognitive ability tests, but as complexity decreases, then psychomotor abilities and tenure tend to have better predictive validity (Gottfredson, 2002; Schmidt et al.,1981). While cognitive ability has its uses in selecting workers for complex and higher-level jobs, it is not always the best choice in every selection situation. A common visual used to describe organizational hierarchies is a pyramid, as one moves up the pyramid job complexity increases as work becomes more abstract and autonomous. As expected, individuals with higher levels of cognitive ability are found higher up in the pyramid and also tend to move higher as their tenure progresses, while individuals with lower levels of cognitive ability either remain where they are or move to less complex jobs (Wilk & Sackett, 1996).

A quick but important side note when mentioning job movement and complexity, is their relationship with satisfaction. While there are many factors that can motivate an employee to stay or leave (e.g., availability of other employment options and commitment with a job), one powerful predictor, even more so than commitment, are employees' satisfaction with their jobs (Tett & Meyer, 1993). Satisfaction can come from a myriad of sources but one of those is through a job that meets an employee's needs. In a study by Park et al. (2008), the authors found that job satisfaction could be predicted by employees' need for cognition and the complexity of their job tasks. Employees with a higher need for cognition sought more complex tasks and when they perceived those tasks to be sufficiently complex, they reported being more satisfied with their jobs.

Measuring cognitive ability

There are a myriad of limitations and considerations that need to be taken into account when measuring cognitive ability. First and foremost, we are human and, as is the case with many of the measure's psychologists use, we are limited by both technology and our own conceptualizations of what we are trying to measure. Unlike height, weight, strength, or speed, cognitive ability must be measured with a proxy such as theoretical concepts like *g* and *gf/gc* and tests designed to give a somewhat agreed-upon approximation of the targeted construct as there is no direct way to observe intelligence directly.

While historical records indicate that in some ways cognitive ability testing has been around since the civil service exams being used in 220 B.C. China (Cartwright, 2019), current theoretical perspectives on cognitive ability tests appear with the introduction of large-scale mental testing starting with the Army Alpha and Army Beta. In 1917, Robert Yerkes and Clarence Yoakum formed a committee whose goal was to develop a group test of intelligence for army recruits (Yoakum & Yerkes, 1920). The original version of the test they created, which contained ten subtests with ten different forms, was piloted on a range of individuals. To validate their test, they also administered the Stanford-Binet or an abbreviated form of it and found .9 and .8 correlations respectively (Wainer, 2002). After several revisions and the creation of a nonverbal alternate form (the Army Beta), approximately two million men were tested, constituting the first large-scale use of intelligence testing. Shortly after this first large-scale use of intelligence testing, Link (1919) provided evidence that by combining job analysis with tests that require the same abilities for the job, employers could better discriminate between good and bad applicants. Another large-scale intelligence test that saw a lot of use was the General Aptitude Test Battery (GATB). The GATB is a battery of 15 tests that measured applicants on several aptitudes like intelligence and dexterity which would assist in presenting the individuals propensity for success in thousands of occupations (Dvorak, 1947)

Mass scale intelligence testing was not without its problems; however, researchers sought solutions that would further improve upon testing. One of the main problems with mass intelligence testing was how it was administered. In its original form, it was a paper and pencil test, and individuals would have to take multiple items of multiple difficulty levels in order to obtain a score, potentially wasting valuable time (Wainer, 2002). Additionally, having to complete an entire examination of increasing difficulty when a subject has a low level of cognitive ability may cause additional frustrations or cause guessing that may introduce error into their scores (Barker, 1938). A solution to these problems was first proposed by Lord (1970) wherein he suggested the idea of an adaptive test. Starting from the middle, a question would be asked that would assess the participants' ability and based upon their answer, they would either move to a difficult item respectively. This would allow a participant's score to be calculated in a much more efficient manner since they would not need to take the entire test. From there, after technology made this type of adaptive testing much easier to implement, computerized adaptive testing was born. The graduate records examination is one such example of a computerized adaptive test that is in use today.

A final concern for testing, not specifically limited to computer adaptive testing, is the concern for cheating. Manipulation of a test score using outside sources or participants or having prior knowledge of the content of the test can all constitute test security issues which would, in turn, lead to inaccurate scores (Karim et al., 2014). There are many routes through which a test may be compromised with each having a different level of impact on the validity of the test. Foster (2010) lists six types of cheating that can potentially affect all forms of testing: copying off of another person, having someone else take the test in your place, having an inside-man help you in some way, gaining access and manually altering the scores, using unapproved materials during the test like notes, and obtaining the content of the test prior to taking it. The last of those, according to Foster, is the most detrimental given how easy it is to obtain the materials and how, from the scoring standpoint, it is nearly impossible to differentiate between someone who had access to the materials beforehand and someone who did not. Unlike coming into an exam with notes written on your hand, having the answers to an assessment memorized cannot be detected by a proctor. To combat that one might use expensive computer adaptive tests with large item banks, but this strategy is not without its faults, as was seen in the early 2000s when students collaborated and uploaded questions from the GRE online to help other students who would be taking the exam. If enough people collaborate and provide the items that they were given, they can, in theory, upload the entirety of the test for other participants to view (Hornby, 2011).

Some organizations cannot afford the aforementioned costs of developing their own cognitive ability tests or cognitively loaded tests but there are plenty of pre-made tests available for purchase and use. One of the most widely used measures of cognitive ability in personnel selection is the Wonderlic Personnel Test (Wonderlic, 2007). For this particular test, participants have 12 minutes to answer 50 quantitative, verbal, and spatial ability items. Matthews & Lassiter (2007) found a stronger correlation between the Wonderlic and crystalized intelligence than fluid intelligence, thus implying that it is better at testing for acquired knowledge rather than reasoning abilities. An alternative to the Wonderlic that focuses more on fluid intelligence, the arguably more important ability when it comes to assessing *g*, is the Raven's Progressive Matrices Test (RPMT) (Nisbett et al., 2012). In the RPMT, participants are given 20 minutes to complete a 60-item measure of cognitive ability that does not require verbal ability to respond to its items. The items of the RPMT require participants to look at a matrix of geometric and patterned figures and determine which among a series of options would complete the matrix following the rules established by the present figures. The non-verbal nature of this assessment allows administration across cultures and languages without the need for translation and provides the additional advantage of reducing adverse impact when compared to global intelligence measures (Hausdorf et al., 2003).

Automatic Item Generation

Automatic item generation (AIG) is a methodology that uses computers to generate items that follow specific rules and are based upon preset item models (Gierl et al., 2015). AIG can be used to create a large number of items from the rules it is given at a rate that dwarfs most traditional item-creation methodologies and thereby greatly reduces the costs as well as concerns about item exposure since, theoretically, it is possible that no two tests will have the same items (Geerlings et al., 2011; Kosh et al., 2019). This sets AIG apart from traditional test creation methods in terms of overall utility for creating cognitive ability tests (Poinstingl, 2009).

As Embretson & Yang (2006) point out, AIG item construction begins with the establishment of item models, a prototype item of sorts that is either uniquely created or imported from an existing measure that can spawn new items. These item models provide the foundation on which new items can be built and have two types of elements: radicals and incidentals. Radicals are the elements of an item that affect its difficulty and are related to the cognitive processes that a participant would need in order to solve that item (Irvine, 2002). An example of a radical would be how big the numbers are in a long division equation, as well as whether or not the answer will include a remainder. Changing the value of the numbers to be divided changes the difficulty of the problem. Doebler & Holling (2016) found that similar psychometric characteristics are found in items that have radicals of similar levels of difficulty. Multiple radicals may also be manipulated simultaneously to broaden the difficulty of a given item and thereby provide a clearer picture of the ability of a given subject (Alves et al., 2010). Incidentals, on the other hand, are any elements of the item that do not influence its difficulty (Irvine, 2002). So, keeping with the division example, an incidental would be the color used to print an item or perhaps its location on the page. While these are both aspects of the item, to be incidentals, they should have no influence on the difficulty of the item. It is important to note here that radicals and incidentals are all test-specific so what may be considered a radical in one test may be an incidental in another.

Constraining items to follow certain rules is an important part of using AIG to develop items since it assures the test creator that the processes that would be used to solve the item are related to the construct that they are trying to measure (Arendasy et al., 2008; Penfield & Camilli, 2007). An example of this when testing mathematical abilities would be avoiding common heuristics that people use to solve specific division problems. The test creator may set constraints on the division items such that the divisor is never one, zero, or the number itself.

As mentioned previously, there are a few ways items can be generated, and each has its pros and cons. The first involves modeling new items off of existing items, like taking items from the GRE to use as models, which results in the creation of item clones (Glas & Van der Linden, 2003). Within this process, the item clones will have similar psychometric properties as the parent items and can be produced and manipulated with the same radicals allowing many items to be created relatively quickly (Geerlings et al., 2011). This brings up one of the main advantages AIG has over traditional test creation, in that the large number of items that can be created would, in theory, prevent the creation of an illegal answer key and also negating memory effects. Its simplicity is also its downfall, as pointed out by Gierl et al. (2015), because even though the items themselves may be different, they would still follow a pattern that could be recognized and exploited by wary test-takers since there is a limit to the psychometric distinctness of each item.

The second way to develop items using AIG is through what Irvine (2002) calls a strong theory of item development or cognitive design system approach. Here, rather than relying on pre-existing items, you focus on a specific cognitive model and manipulate radicals in a systematic fashion that allows the researcher to predict difficulties of the items and ensure the cognitive model is being used. While this theory-backed design approach sounds good, the lack of cognitive theories to back them can hinder the number of potential applications (Lai, Alves, & Gierl, 2009). This could also result in the discarding of items due to insufficient model characteristics (Arendasy & Sommer, 2012).

With both of those previous methods proving to be insufficient, researchers searched for a more robust way to create items that would have sufficient psychometric characteristics. The automatic min-max approach provides a method of doing so that includes a cognitive model that would allow for more item types that assess the construct built right into the item construction process (Arendasy & Sommer, 2012). In classic test creation, both Hinkin (1998) and DeVellis (2003) suggest beginning with defining the latent construct that you intend to measure with the items you will write, a process which is mirrored in the min-max approach, followed by the specification of the cognitive model that the researcher believes will assess the latent trait. It is here where item constraints may be introduced in order to get a purer assessment of the latent trait. This allows the creation of more item types without having to discard them due to poor model characteristics (Arendasy & Sommer, 2007; Arendasy & Sommer, 2010; Arendasy & Sommer 2012). Through the automatic min-max approach, a much more effective AIG test can be created.

Practice Effects

The effects of practice on cognitive tasks have been well documented in previous research (Bartels et al., 2010; Hausknecht et al., 2007). Mere exposure to testing materials in either parallel or identical forms has often shown marked increases in participants' scores. The magnitude of such changes differs based upon a variety of factors such as similarity between the test and the materials practiced, the time intervals between practice sessions and testing sessions, or differences in methodology across research settings (Hausknecht et al., 2007).

There are many explanations as to how practice actually improves performance and in any given scenario there can be multiple influences operating at the same time. Messick and Jungeblut (1981) found that one explanation for practice effects was a reduction in anxiety pertaining to the testing situation. Anxiety can inhibit performance on novel tasks so familiarity with the task via exposure or through coaching would reduce said anxiety leading to an increase in score across subsequent administrations. Alternatively, strategies such as coaching may reduce group differences due to stereotype threat. However, if a measure does not change questions over different administrations, then a participant's memory of their correct and incorrect responses from previous testing scenarios may explain their increase in score on subsequent tests (Kulik et al., 1984). Another explanation for practice effects is regression to the mean: essentially an extreme score moving towards average on subsequent collections (Campbell & Kenny, 1999). Additionally, the effects of practice may be due to enhanced test-taking strategies based on participants’ past experiences with the items and concerted efforts to improve (Sackett et al., 1989). Finally, Hausknecht et al. (2007) identify mere repetition as an explanation of practice effects that is absent of any type of formal intervention or strategic undertaking. It is important to note; however, that not all score differences have the same source, e.g., stereotype threat, testing anxiety, actual ability differences, and as such the influence of practice may differ.

Since practice effects due to memorization of items are widely known, many restrictions have been put in place to combat it in a variety of settings (e.g., educational, occupational, research). The Educational Testing Service, for example, not only changes the items that are in each test administration, but they also have specific windows in which testing may occur in an attempt to combat memory effects. The Basic Attributes Test for the United States Air Force only allows candidates to take the test once in their careers (Carretta et al., 2000). According to the Society for Industrial-Organizational Psychology’s (2003) guidelines on employee selection, if it is technically and administratively feasible, employers should offer applicants opportunities for reassessment. This would allow participants the opportunity to showcase their best performance. While practice effects may lead to improved scores on cognitive ability tests, the relationship between this improvement and job performance has not been explored.

Adverse Impact

Adverse impact as a concept originated with the legal case Griggs v. Duke Power Co. in 1971, where, after the passing of Title VII of the Civil Rights Act of 1964, the Supreme Court ruled that tests that are not reasonably related to the job that disparately affects people falling under protected classes violates Title VII. If there is discriminatory evidence, the organization is vulnerable to charges. Adverse impact is often a major consideration regarding the use of selection processes within an organization.

Adverse impact is often assessed through the use of the 80% or 4/5ths rule as outlined by the Uniform Guidelines on Employee Selection Procedures. The calculation compares the passing rate of the group with the highest selection rate to that of the other groups. There is evidence for adverse impact if the comparison groups' passing rate is less than 80% of the highest group. If evidence of adverse impact of a particular test is found then it is up to the company to either remove the test or provide validity evidence (e.g., BFOQ) satisfying the Uniform Guidelines. A problem with calculating adverse impact in this manner; however, is that it is vulnerable to small sample sizes. A way to circumvent this is to use more rigorous methods, like a Fisher's Exact Probability Test or a chi-square goodness of fit test, which the Uniform Guidelines allows. It is important to note; however, that different tests may produce different conclusions, especially when small sample sizes are an influence. When this occurs, it is suggested that using a significance test combined with the 4/5ths rule will reduce type one error the most while simultaneously maximizing power (Collins & Morris, 2008). The Uniform Guidelines also add the provision with the 4/5ths rule that if switching one case would change the result to not having adverse impact then it is an acceptable ratio.

Study 1: Validation of an AIG Measure of Intelligence

The purpose of this study is (1) to validate a newly created measure of intelligence and (2) use that measure in a hypothetical selection scenario to see if practicing the measure will reduce adverse impact. The first study concerns the validation of the Katyem Object Tracking Assessment (KOTA) in two stages. The KOTA is an AIG test with figural matrix items. Such items are useful in assessing *g* (Freund et al., 2008). Similar tests, such as the Ravens Advanced Progressive Matrices (APM) and the Blox test of spatial ability have been used in employee selection scenarios (Kock & Schlechter, 2009) The test uses a series of geometric figures, circles, triangles, and rectangles, that change orientation, size, position, and border thickness to create a pattern that the participant must complete by choosing the correct option that follows the rules established by the previous three figures in the item. This test follows a somewhat similar format to the APM in that subjects are asked to find the correct answer for a progression of shapes across three examples. If executed properly, much like the APM, it should tap into *gf* given the nature of the item type. Figure one below provides an example item from the KOTA.

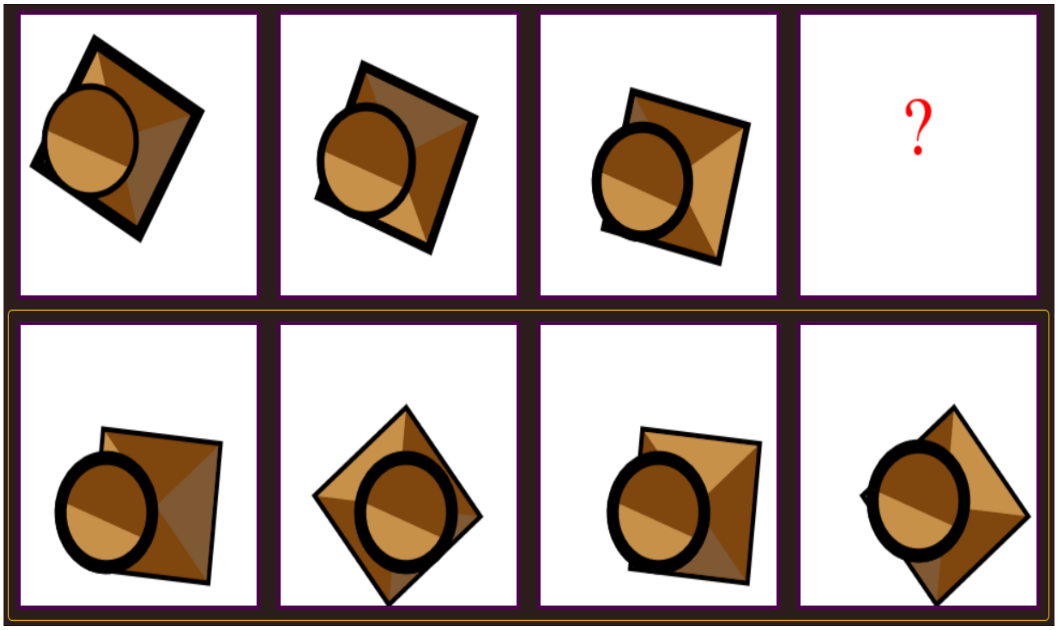


Figure 1: Example Item from the KOTA.

Construct validity is the degree to which a given instrument is measuring the specific construct it is intended to measure. Furthermore, construct validity can be established using two different approaches and both were be used to provide support in the current research: convergent and discriminant validity studies (Campbell & Fiske, 1959). One way to provide evidence for convergent and discriminant validity is through the use of the multitrait multimethod matrix, wherein scores from various measures are put into a matrix and their correlation coefficients are calculated allowing the researcher to see how the scores correlate.

With convergent validity, the instrument you are validating is compared to an existing instrument that assesses the construct in question. Ideally, the two instruments should be highly correlated (Campbell & Fiske, 1959). If it has a poor correlation with the other measure, then there is a lack of evidence that your test is tapping into the construct of interest. This study will use the Hagen Matrices Test (HMT) as the test of comparison. The HMT has a.53 correlation with *gf* as measured by the extended German Intelligence-Structure-Test 2000 R (Heydasch, 2014). The researchers would expect a similarly strong correlation between the HMT and the KOTA.

Hypothesis 1: Participants' scores on the KOTA will have a strong positive correlation (.5) with their scores on the HMT.

Discriminant validity is demonstrated when the instrument in question has a low or no correlation with instruments measuring constructs that should be theoretically unrelated (Campbell & Fiske, 1959). Constructs such as personality are typically weakly correlated with intelligence (Heydasch, 2014; Moutafi et al., 2003); Austin et al., (2002) found evidence that across multiple personality measures and intelligence measures, correlations between *g* and maladaptive personality traits, like neuroticism, on average are.1. Within the Five-Factor model of personality, they consistently found traits associated with neuroticism to have the highest negative correlations with *g*. The emotionality factor of the HEXACO-PI-60, which correlates strongly with the neuroticism factor in the five-factor model, should also have a low correlation with *g* (Ashton & Lee, 2009). Researchers would therefore expect a weak correlation between the KOTA and the emotionality dimension of the HEXACO-PI-60.

Hypothesis 2: Participants' scores on the KOTA will have weak correlations with their scores on the emotionality dimension of the HEXACO-PI-60, between -.3 and .3.

Study 2: Adverse Impact and Practice

The second study focused on using the KOTA in a hypothetical selection scenario. The scores on the KOTA will constitute the sole measure of the system in determining if the participant is "hired" or not in a strict top-down selection system based on the selection ratio of "hiring" the top 20% of applicants to fill 11 "vacancies". 20% was chosen as a realistic number of hires for a low-level job, in tandem with allowing enough participants to be hired to allow for adverse impact analyses to be assessed. Given the prevalence of adverse impact within cognitive ability measures and keeping in mind the legal issues of preferential treatment, all participants will take a pre-test and then practice the measure three times before taking the fifth and final version of the test. Three practice sessions were chosen as a balance between not burning out the participants with repetitive testing and also maximizing score gains given that after three sessions there is almost no significant change in score (Falleti et al., 2003). It has been well documented that there is adverse impact present within many cognitive ability tests, with Blacks and Hispanics often being the ones impacted the most (Loehlin et al.,1975; Roth, et al., 2001; Waschl et al., 2016). Adverse impact will be examined using the 4/5ths rule as well as a chi-squared goodness of fit test and the two standard deviation rule. The allowance of practice is one of many methods suggested by Ployhart and Holtz (2008); however, they found that it may not be the most effective given the constraints of having to create practice items along with the test. The use of an AIG-based test foregoes this concern and provides ample opportunities for practice. Given that cognitive ability tests tend to have group differences based on the race of participants, since the KOTA is a cognitive ability test, the researchers believe that there will be evidence of group differences. Additionally, since there is evidence that practice increases scores on measures, moderated by test familiarity, researchers expect there to be a reduction in the difference in mean scores between the two groups. Finally due to the score increases from practicing, the researchers believe that there will not be evidence of adverse impact in a simulated hiring scenario.

Hypothesis 3: There will be initial group differences in the scores on the KOTA based on the race of participants.

Hypothesis 4: The group differences in the scores on the KOTA will decrease due to practice.

Hypothesis 5: There will be no evidence of adverse impact between the groups in a simulation of a hiring scenario after the groups have practiced.

CHAPTER 2

METHOD

Study 1

Materials and Procedure

The purpose of this study was to examine hypotheses one and two concerning the validity of the measure. Participants were given the following three assessments: the Hagen Matrices Test Short form, The HEXACO-PI-60 (Emotionality portion only), and the KOTA. Participants were randomly assigned to either take the KOTA first or the HMT first followed by the other one and they took the HEXACO last. This counterbalanced design was implemented to prevent order effects. The tests were administered remotely through links distributed by the researcher sending the participant to the appropriate website to complete the tasks. Participants were given a one-week window to complete the assessments, with periodic reminders from the researcher. The participants were instructed to take at least a one-hour break in between tests to minimize test fatigue. The three assessments combined took under an hour to complete. The participants were given a short demographic survey recording age, race, vocation, education level, gender, and how often they engage in puzzles/ puzzle games. No identifying information was collected, and participants were assigned a participant code that they used for each of their assessments to allow for comparisons. Data was collected via the Qualtrics survey platform and the Katyem website. The data of participants who did not fully complete all three measures was not used.

Participants

458 participants, after a G*power analyses indicating only 200 participants for a validation study were needed to detect a medium effect size at $\alpha = .05$ with power = .80, were recruited via snowball sampling from Facebook, and LinkedIn, professional contacts of the researchers, Mturk participants as well as from university students from several universities. Participants were sent an email containing the instructions for completing the survey if they were university students. Facebook, LinkedIn, and Mturk participants received a similar set of instructions. After data cleaning steps were conducted, removing careless responders and incomplete cases, 89 participants remained.

Analysis Plan

The first two hypotheses concerning convergent and discriminant validity were assessed via multitrait multimethod matrix (MTMM) (Campbell & Fiske, 1959) with each of the three test scores. The MTMM provided a way to establish convergent and discriminant validity by showing how similar or dissimilar traits or constructs were based on the method specific variance. The outputted correlations of this test, specifically the correlations between the end final test scores, would provide some evidence of convergent and discriminant validity between the three measures.

Study 2

Materials and Procedures

The purpose of this study was to see if allowing participants to practice reduces potential adverse impact. Participants were administered the same demographic survey items as in the last study. The participants had one week to complete five sessions of the KOTA with a minimum of half an hour between sessions. The first and final session of the KOTA consisted of the same items in order to compare performance changes, while the middle three sessions consisted of parallel items. The researcher sent periodic reminders to participants. The demographics survey and KOTA were administered remotely through links distributed by the researcher. To simulate a realistic selection system for an entry level job, only 11 participants were *hired*. In this study, hiring was purely a hypothetical and no follow ups with the participants were made regarding their scores. The data of participants who did not fully complete the KOTA was not used for the hypotheses.

Participants

The number of participants for this study was estimated using the G*Power software. An estimated 40 participants were needed to detect a medium effect size at $\alpha = .05$ with power = .80. Participants were split into two groups: minority (Blacks and Hispanics) and non-minority (Whites). Participants were collected using online crowdsourcing websites.

Analysis Plan

A Repeated-Measures ANOVA was conducted to analyze the group differences between the minority and non-minority groups allowing for between and within group comparisons. To evaluate the effects of practice only, the scores of the first and fifth session of the KOTA were compared since they are identical forms. Post-hoc analyses were conducted as necessary. To calculate adverse impact, participants were ordered according to their final KOTA scores and a strict top-down selection method was used to determine the top 20% and adverse impact calculations were conducted using a chi-squared goodness of fit test as well as using the Fisher's Exact Test, and 4/5ths rule.

CHAPTER 3

RESULTS

Study 1

The data were cleaned, formatted, and explored for missing or careless responses. Data was collected from 458 participants in total, originating from a combination of MTurk, convenience, and snowball sampling methodologies. After removing any participant that did not complete the demographics survey, Hagen Matrices Test (HMT), HEXACO (emotionality only), and the Katyem Object Tracking Assessment (KOTA), only 160 participants remained. 36 additional cases were removed when it was discovered they were provided different items for three questions on the KOTA than the rest of the sample due to an error. Of the remaining 124, 35 additional participants were removed due to a pattern of careless responding, objectively defined and screened as having five or more incorrect items in the KOTA completed with average response times of less than three seconds or having seven or more items incomplete after the allotted 30 seconds per item had passed. Particularly swift or consistently lengthy times for within participant data has been suggested by a number of authors as evidence for inattentiveness (Hauser et al., 2018; Kittur et al., 2008). Three seconds was identified as the minimum amount of time it should take to answer each item based on the mean and standard deviation of item response time (mean = 11.76, sd = 6.28). Although there is the potential for user error in clicking rapidly and accidentally answering the next item correctly and moving on to the item, the likelihood of this is very low and so it was determined that this occurring for more than 5 items was indicative of careless responding. Since the responses taking 30 seconds or longer are coded as incorrect responses and automatically skipped, this negatively impacted participants scores and the distribution of the overall dataset. No additional cleaning steps were conducted on this dataset. The 89 remaining participants were analyzed, and the results are included in the Table 1 below.

Table 1: *Table of means and standard deviations for KOTA, HMT, and HEXACO*

Measure	Mean	SD	Range
KOTA	36.58	10.64	14-60
HMT	3.43	1.64	0-6
HEXACO	32.69	8.62	14-48

Next the assumptions for correlation were checked. Each of the variables were continuous and had related pairs from the same participant. There were no outliers in the dataset (i.e., no more than ±3.29 SD from the mean).

A Pearson's product-moment correlation was conducted between the scores on the KOTA and HEXACO, the scores on the KOTA and HMT, and between the items on the KOTA and HEXACO, and between the items on the KOTA and HMT to provide evidence of divergent and convergent validity respectively.

For the HMTand the KOTA, there was a significant moderate correlation $r(87) = .45$, $p < .01$, 95% CI [.28, .61]. While these results provide evidence that there is a moderate positive correlation between scores on the HMT and scores on the KOTA, they fail to fully support the hypothesis that participants' scores on the KOTA will have a strong positive correlation (>.5) with their scores on the HMT.

For the HEXACO and the KOTA, there was a non-significant correlation (*r*(87) = -.1), *p* = .32, 95% CI [-.31, .11]. This result fails to provide evidence that there is a weak correlation between the participants' scores on the emotionality dimension on the HEXACO and the KOTA. Table 2 displays the correlations.

Table 2: *Table of Correlations*

Test	Correlation	*p*	95% CI
HMT-KOTA	*r*(87) = .45	*p < .01*	[.28, .61]
HEXACO-KOTA	*r*(87) = -.1	*p* = .32	[-.31, .11]

In sum, the results of this study fail to adequately provide convergent and discriminant validity evidence for the KOTA as a measure of cognitive ability. There is insufficient evidence that the construct the KOTA measures is the same as the construct the HMT measures, and is different than the construct the HEXACO (emotionality) measures. Experimental pitfalls, potential modifications, and other points will be broached in the next chapter.

Study 2

The data were cleaned, formatted, and explored for missing or careless responses. Data was collected from 69 participants in total, originating from a combination of MTurk, convenience, and snowball sampling methodologies. After removing any participant that did not complete the demographics survey, Pre-Post and three practice KOTA measures, only 52 participants remained. 23 identified as Black or African

American and one identified as Hispanic, or Latino were placed into the minority group. 28 participants who identified as white/Caucasian were placed into the non-minority group. The researchers noticed some of those participants took the post-test multiple times, for example one specific participant took the post-test 29 times across two different days. The data that were analyzed only included the participants first pre-test score and their first post-test score; any additional submissions were discarded based on the time the submission was completed. The means and standard deviations of each group are included in Table 3 below.

Table 3: *Table of means and standard deviations for minority & non-minority pre and post test scores.*

Group	Test	Mean	SD	Maximum
Minority	Pre-Test	34.08	8.53	18-48
Minority	Post-Test	34.91	9.83	14-50
Non-Minority	Pre-Test	33.57	9.21	20-54
Non-Minority	Post-Test	37.42	12.54	12-58

To explore the hypotheses that there will be group differences in participants' scores on the KOTA and that practice will reduce these group differences, a two-way repeated measures ANOVA was conducted. First, the assumptions were checked. There were no outliers among the scores of the participants. A Shapiro-Wilkes test of normality indicated the data was normally distributed (p = .3); additionally, the results of the QQ plot supports the same conclusion (see Appendix F).

Based on the results of the two-way repeated ANOVA, there is no evidence that there are differences between the group scores. The hypothesis that there would be group differences in participants' scores on the KOTA based on their group was not supported (*f*(1,49) = .228, *p* = .63). This is a promising result in the context of adverse impact as this preliminary result indicates a lack of group differences on the measure based on race of participants of this study. The results also provided no evidence that practice reduces group differences on the measureand fail to support the hypothesis that practice will reduce the difference in scores between minorities and non-minorities (*f*(1,49) = .863, *p* = .35). Table 4 displays the results of the ANOVA

Table 4: *Two-Way Repeated Measures ANOVA Table*

Group	DF	Sum sq	Mean sq	F value	Pr(>F)
Race	1	26	25.85	.228	.635
Test	1	98	98	.863	.357
Residuals	49	5562	113.51		

To assess the presence, or absence, of adverse impact, the participants' scores were used in a hypothetical selection scenario that utilized strict top down selection based on the score of the participants. For the simulation, 11 of the 52 (20%) participants would be *hired*. Table 5 displays the score and race of the “hired” participants.

Table 5: *Hiring scenario results*

Placement	Minority Status	Score
1	Non-Minority	58
2	Non-Minority	56
3	Non-Minority	56
4	Non-Minority	54
5	Minority	50
6	Non-Minority	50
7	Minority	48
8	Minority	48
9	Minority	46
10	Non-Minority	46
11	Non-Minority	46

Multiple methods to calculate adverse impact were used. The results of the four-fifths rule indicated that adverse impact was present in this hiring scenario; however, this would not have been the case if just one additional minority had been selected. A chi-squared goodness of fit test was also used to determine if there was evidence of adverse impact, the results of which found that the observed proportions are not significantly different from the expected proportions (p = .36). As further evidence for a lack of adverse impact in this measure, Fisher's exact test of independence was not statistically significant (p =.51). These results support the hypothesis concerning the lack of adverse impact associated with the KOTA. Table 6 displays the results of the 4/5ths rule.

Table 6: *4/5ths Rule table*

	Hired	Applied
Minority	4	23
Non-Minority	7	28

CHAPTER 4

DISCUSSION

Power & Sample of Study 1

The main obstacle of study one was lack of power and sample size. While it was calculated that 200 participants were needed for statistical power and to detect a medium effect, the end result after cleaning the data was 89 participants. Achieving the size of 200 participants would have provided enough power to detect the low correlation between the KOTA and HEXACO, which the current 95% confidence interval indicated was between -.3 and .1 Bootstrapping was considered by the researchers as a method of evaluating the distributions of the data, however it would not have influenced the actual power of the study and therefore was not conducted. This highlights the constraints of unproctored data collection for multiple complex measures. The necessity of the participants completing all three measures, taking a break in between each, coming back to finish, and careless responding were difficult hurdles to surmount in a virtual environment when there was no identifying data collected to allow for the researchers to ping participants to complete their tests. After collecting more than twice the number of estimated participants needed, the researchers began analyzing the data, which prevented further data collection. Perhaps a better approach to this study would have been to use proctored lab setting with participants being offered breaks onsite before being encouraged to continue the measure, this way their completion would be reinforced; however, current pandemic restrictions prevented the researchers from utilizing this type of methodology.

Test Length

The use of the short version (6 items rather than 20) of the HMT might have influenced the results in some manner; however, the short version was specifically chosen to combat test fatigue. The decision to prioritize reduction of test fatigue over a potential increase in validity was made in order to account for the expectation that fewer participants would complete the study if they had to complete all 20 items of the HMT. This hypothesis was supported in the pilot testing of the study, wherein feedback from participants was negative towards the length and commitment for the HMT and many who received the HMT first did not return after their break to complete the KOTA and HEXACO.

Test Choice

Although only the emotionality portion of the HEXACO was used, there was a possibility that any other measure within it may have provided different and potentially significant results. Additionally, the choice of the HMT was due largely to availability as it was already designed to be distributed virtually. Additional research would need to be performed to see how the KOTA correlates with other measures. Although the hypotheses within this study were rejected, there is still potential for further research in the validation of the KOTA as a measure of cognitive ability, with better experimental conditions.

Future Directions

The lack of adverse impact in the KOTA is a promising finding given the prevalence of it in other cognitive ability tests. It should be noted, however, that the use of students in the participant pool introduces preselection effects into the study so it was not a truly random sample. Therefore, more data collection will be needed to truly make inferences about this test. More research will be needed to explore exactly what niche this test fills and how to validate it, however, that is beyond the scope of this current study. Future researchers are encouraged to utilize the KOTA in their validation studies but are cautioned against experimental designs with similar pitfalls as this one. Gathering participants and motivating them to complete a single measure without careless responding is somewhat difficult; gathering participants and motivating them to complete the same measure multiple times is very difficult. Measures like the KOTA, however, provide unique opportunities to explore the effects of practice and answer questions such as: what is the ideal number of times an individual can practice before there is no benefit, or does that peak amount of practice change in tandem with another variable such as race, or gender?

Conclusion

Advancements in technology are continually paving the way for researchers to test new and different methods of testing, one of which being AIG. Having access to a vast library of items that can be assembled and distributed remotely to participants provides new avenues for testing without issues of test security. Giving researchers the opportunity to further explore practice effects and cognitive ability are an additional benefit of these technological advancements. Although the findings of this study failed to support the four hypotheses concerning validation and group differences, they did support the hypothesis concerning the absence of adverse impact in the KOTA. There are still opportunities for additional studies to be conducted which can further explore the findings contained herein.

APPENDIX A

IRB FORM

Do you plan to publish this study?

□ **YES** □ NO

Will this study be published by a national organization?

□ YES □ **NO**

Are copyrighted materials involved?

□ YES □ **NO**

Do you have written permission to use copyrighted materials?

□ YES □ NO

Researchers must comply with all training requirements from their funding agency.

Are all Researchers Up to Date on Human Subjects Training? (attach certificates) □ **YES** □ NO

Training is on www.citiprogram.org □ YES □ NO

Do any Special Permissions Need to be attached? (School district, data holder, Agency) □YES □ **NO**

STUDY/PROJECT INFORMATION FOR HUMAN SUBJECTS COMMITTEE

Describe your study/project in detail for the Human Subjects Committee. Please include the following information.

TITLE: Exploring Practice as a Method for Reducing Adverse Impact in a Selection System

PROJECT DIRECTOR(S): Derrick McDonald & Dr. Tilman Sheets

EMAIL: mcdonad.derrick27@gmail.com

PHONE: 773-260-2869

DEPARTMENT(S): Psychology and Behavioral Sciences

PURPOSE OF STUDY/PROJECT: To see if allowing participants to practice a measure before taking it will remove any potential adverse impact present in a simulated selection system.

SUBJECTS: Amazon mTurk users, colleagues and connections of the Louisiana Tech I-O Psychology Doctoral Program.

PROCEDURE: Participants (n=200) will be recruited through social media snowballing methods, Amazon mTurk, and within Louisiana Tech University. Participants will be instructed to take an open sourced cognitive ability test and another one created by Dr. Tilman Sheets. The one created by Dr. Sheets will be the one they are going to practice, while the other cognitive ability test will be used in validation procedures. Comparisons will be made on the scores of both tests to determine construct validity.

INSTRUMENTS AND MEASURES TO INSURE PROTECTION OF CONFIDENTIALITY, ANONYMITY: Information regarding participants will be kept confidential. Participants will be randomly assigned an alphanumeric identifier. The data will be recorded on a secure private server. Using a flash drive I will download the data from the server and analyze it on a separate computer that is not connected to the internet. There will not be enough demographic data gathered to identify any of the participants.

RISKS/ALTERNATIVE TREATMENTS: There are no risks or alternative treatments related to this study.

BENEFITS/COMPENSATION: Only mTurk workers will be paid $7.25 per hour, shorter working hours will be paid according to the same ratio. Students may be offered extra credit at the discretion of their professor but alternate methods of compensation for those who do not participate will be recommended.

SAFEGUARDS OF PHYSICAL AND EMOTIONAL WELL-BEING: It will be stated to subjects that this test is an estimation of cognitive ability, and does not reflect their true intelligence and may not reliably reflect or measure their intelligence, it is a research instrument. Participants will not receive any meaningful or actionable feedback regarding their scores.

APPENDIX B

DEMOGRAPHIC ITEMS

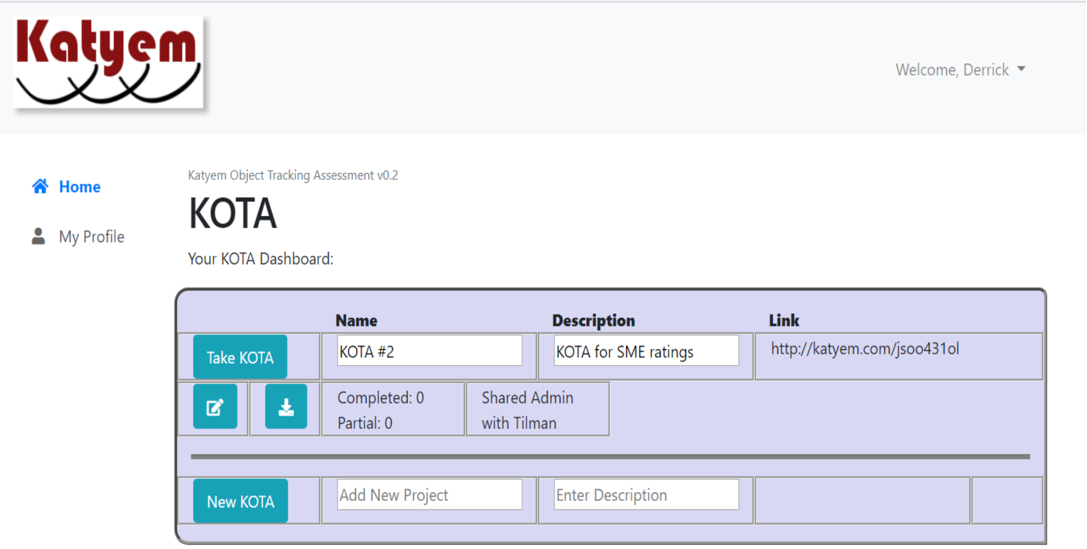
All the demographic items that participants were asked along with all possible response options.

- How old are you?
- Sex
    - Male, Female
- Race, national origin, ethnicity
    - White/Caucasian, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Hispanic, Latino, or Spanish origin, Other
- What is the highest level of school you have completed or the highest degree you have received
    - Less than high school degree, High school graduate (high school diploma or equivalent including GED), Some college but no degree, Associate degree in college (2-year), Bachelor's degree in college (4-year), Master's degree, Doctoral degree, Professional degree (JD, MD).
- Please indicate your occupation:
    - Management, professional, and related, Service, Sales and office, Farming, fishing, and forestry, Construction, extraction, and maintenance, Production, transportation, and material moving, Government, Retired, Unemployed, Student worker, Other
- How often do you engage in puzzles or puzzle games
    - Frequently, Often, Occasionally, Somewhat, Never

- Please enter your participant ID provided in the previous question.

APPENDIX C

KOTA ITEMS AND TEST

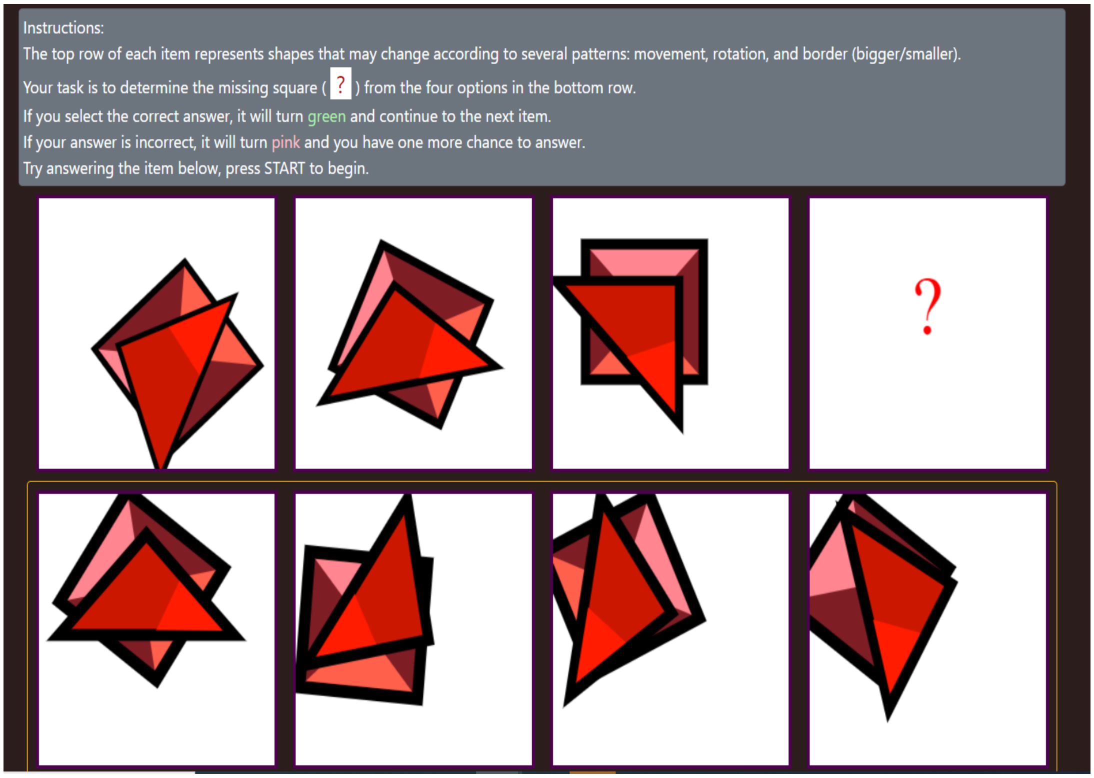


Katyem Object Tracking Assessment v0.2

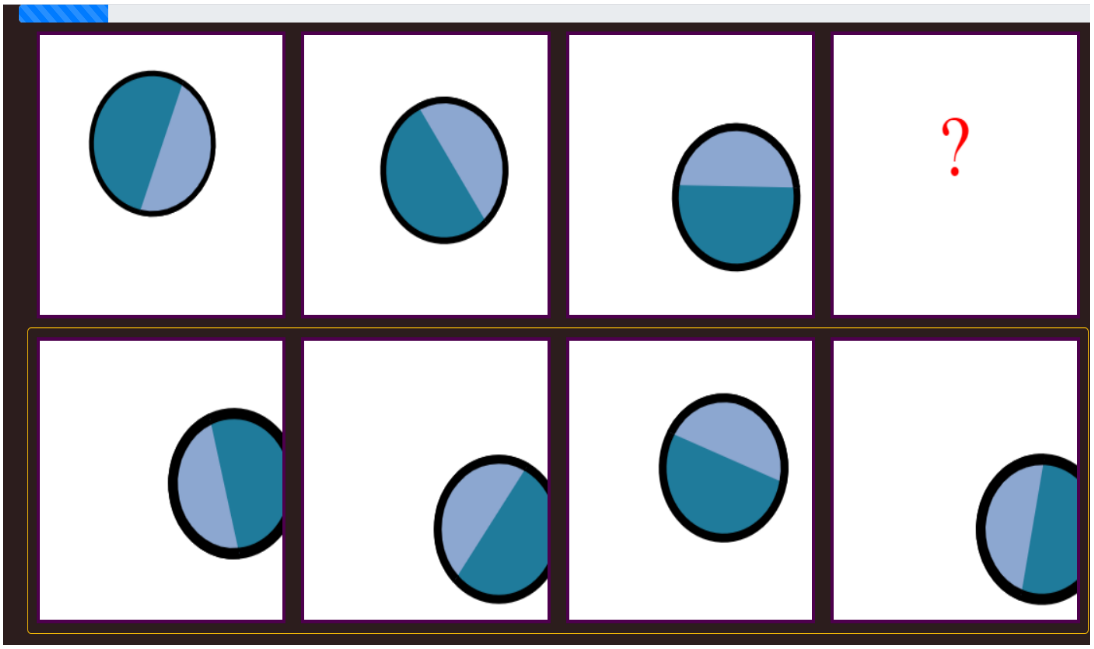
KOTA

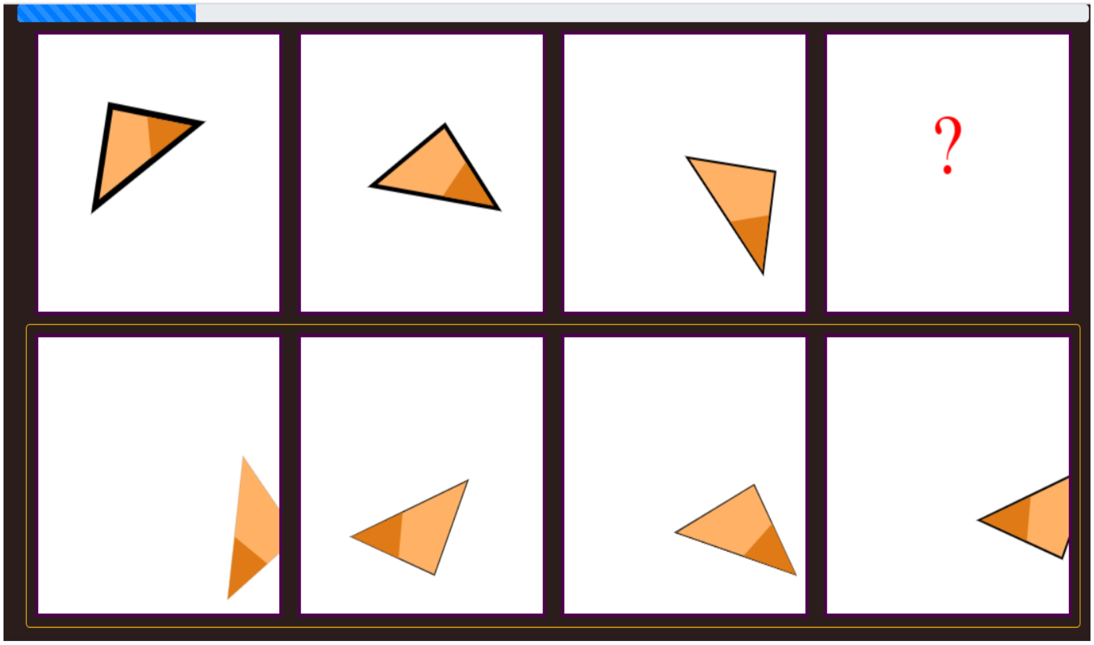
Your KOTA Dashboard:

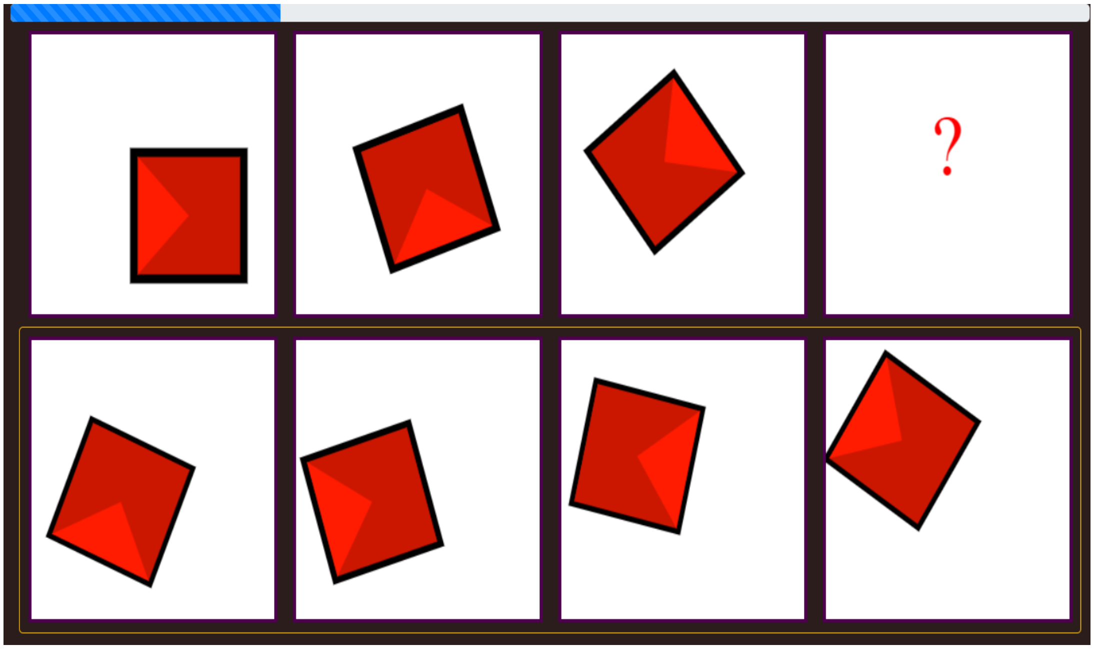
	Name	Description	Link
Take KOTA	KOTA #2	KOTA for SME ratings	http://katyem.com/jsoo431ol
	Completed: 0 Partial: 0	Shared Admin with Tilman	
New KOTA	Add New Project	Enter Description	

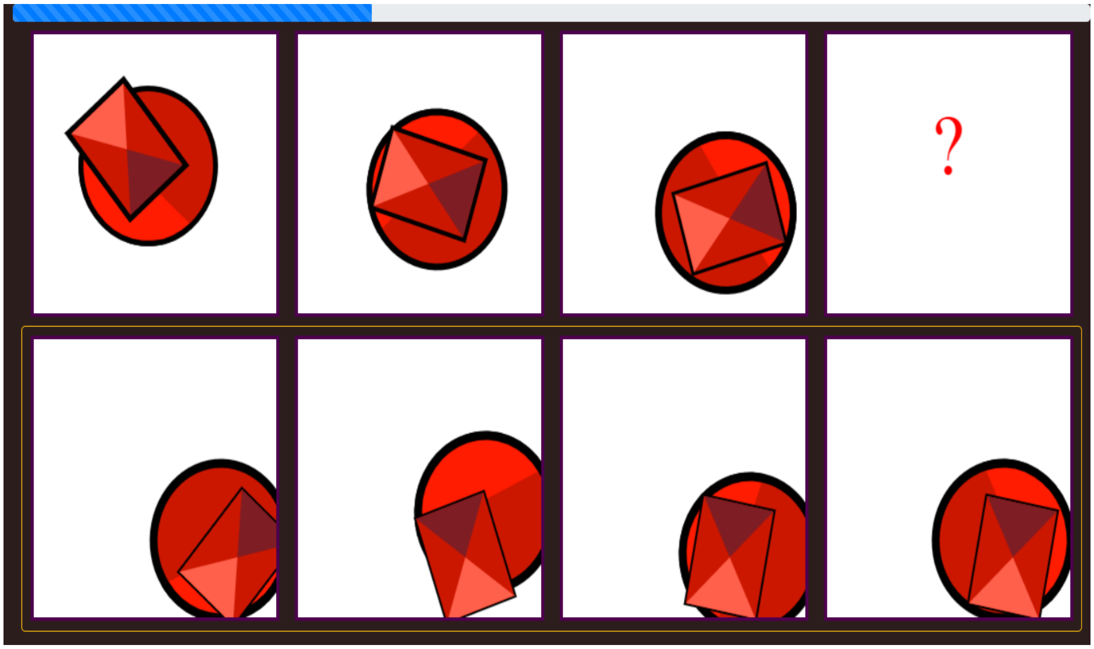


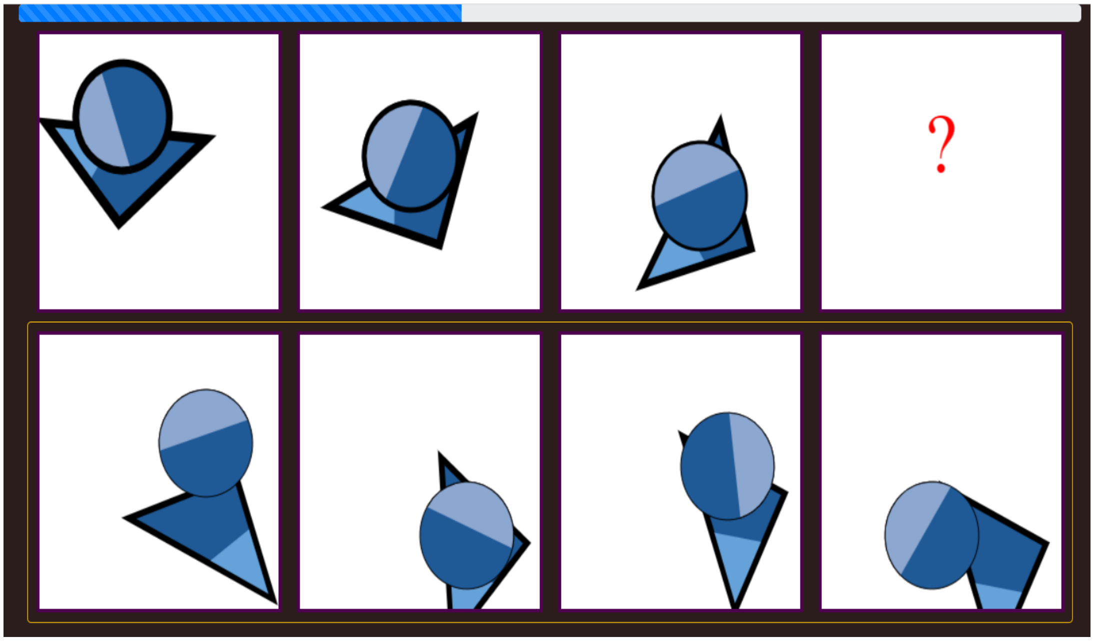
Instructions:
The top row of each item represents shapes that may change according to several patterns: movement, rotation, and border (bigger/smaller).
Your task is to determine the missing square (?) from the four options in the bottom row.
If you select the correct answer, it will turn green and continue to the next item.
If your answer is incorrect, it will turn pink and you have one more chance to answer.
Try answering the item below, press START to begin.

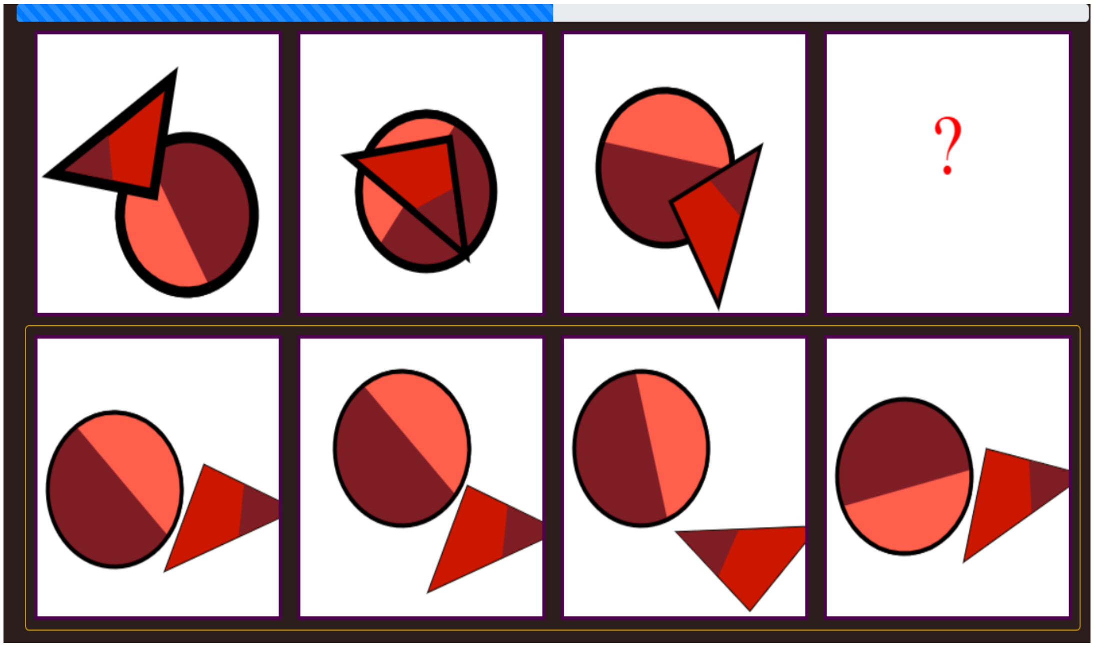


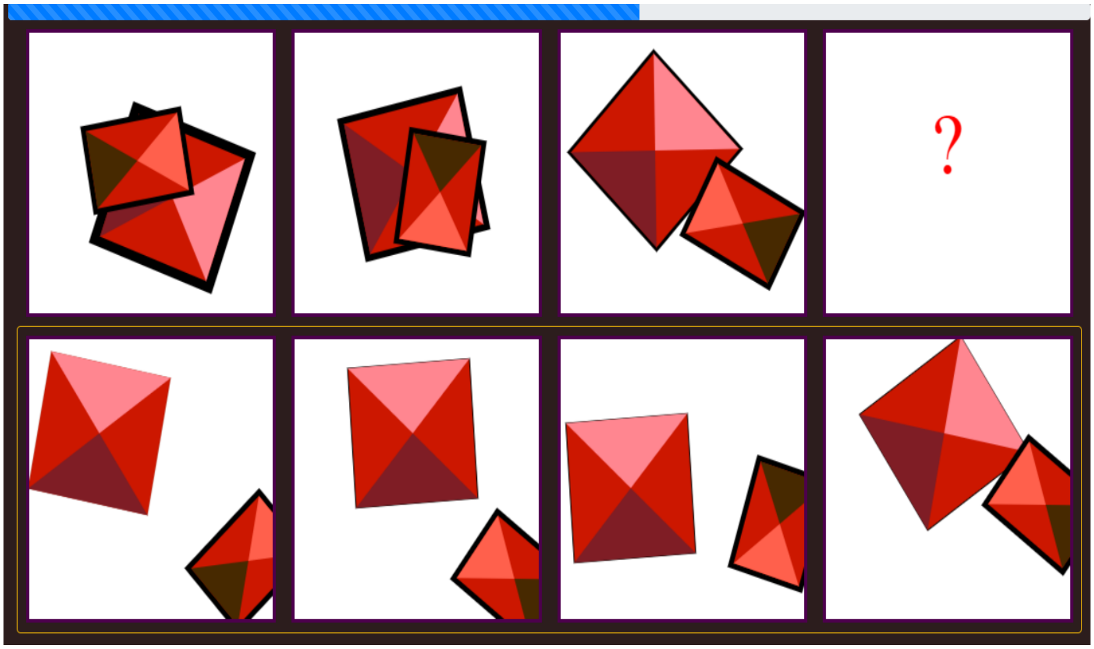


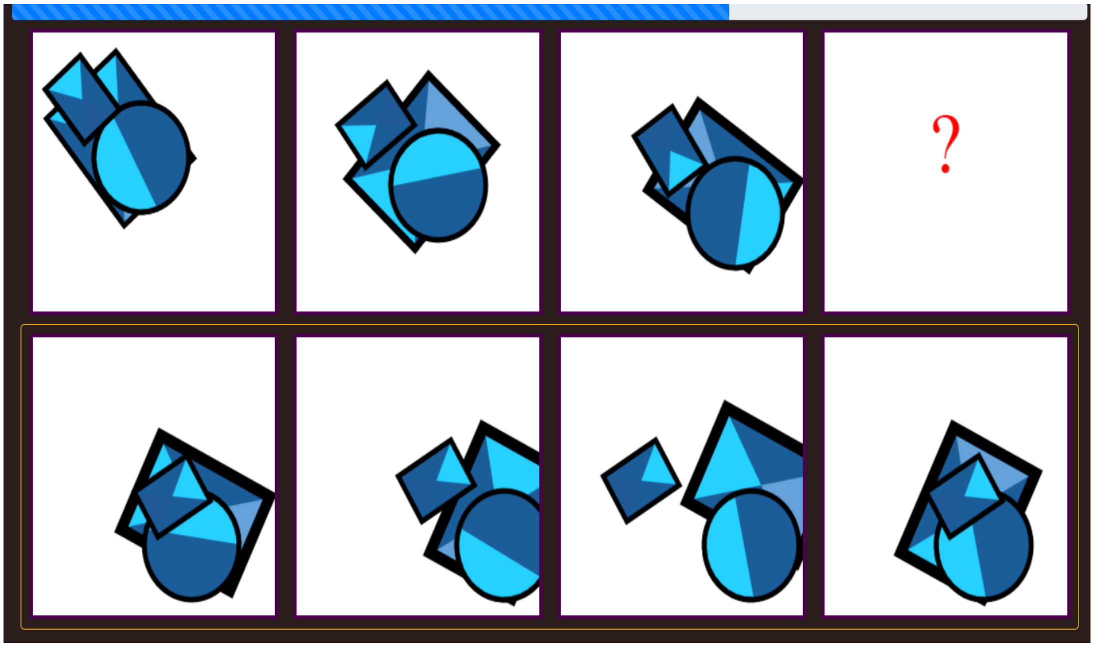


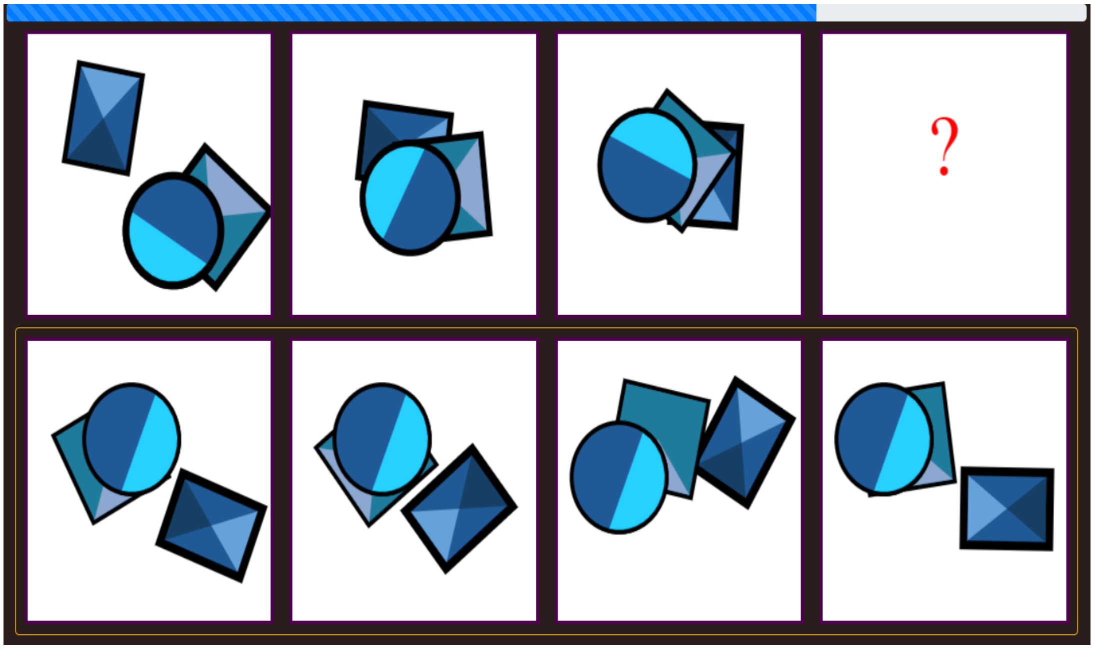


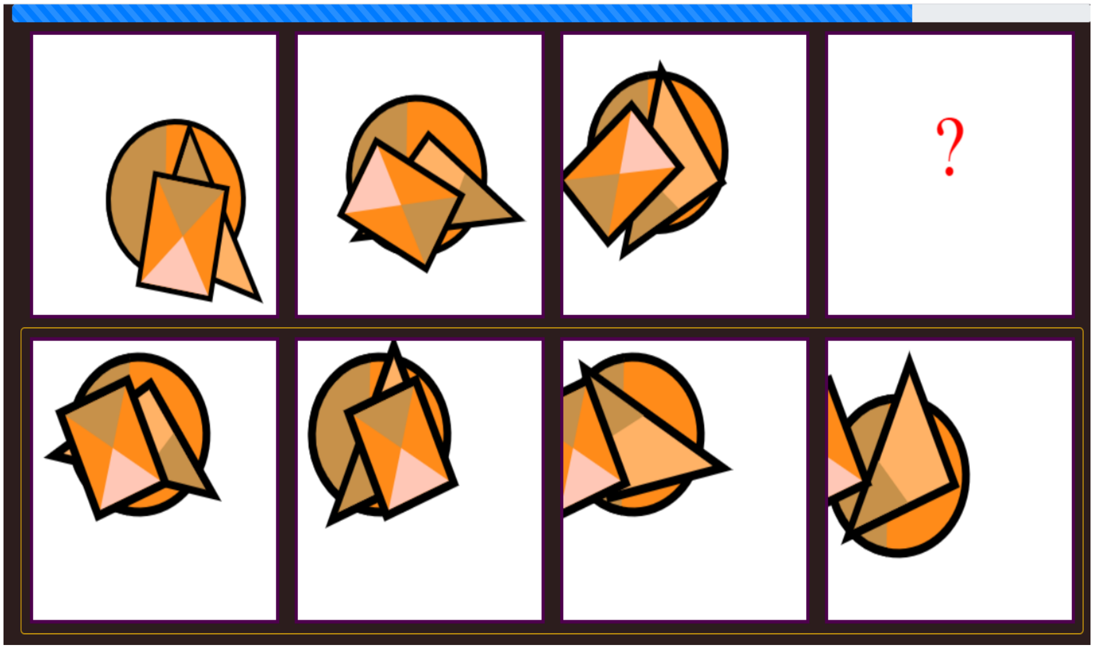


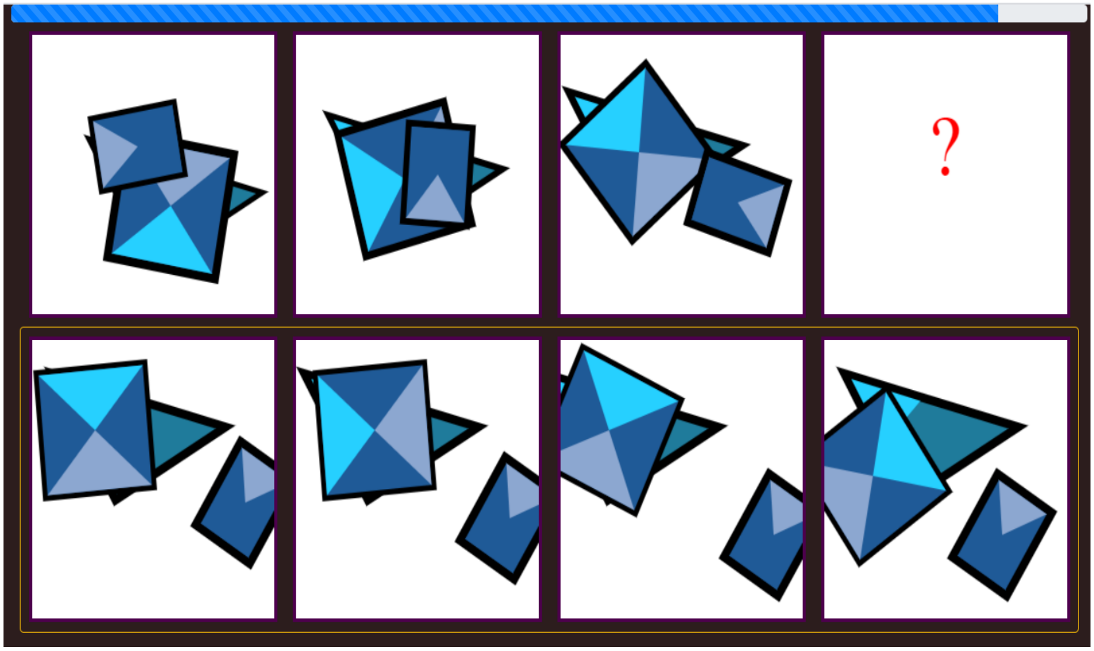


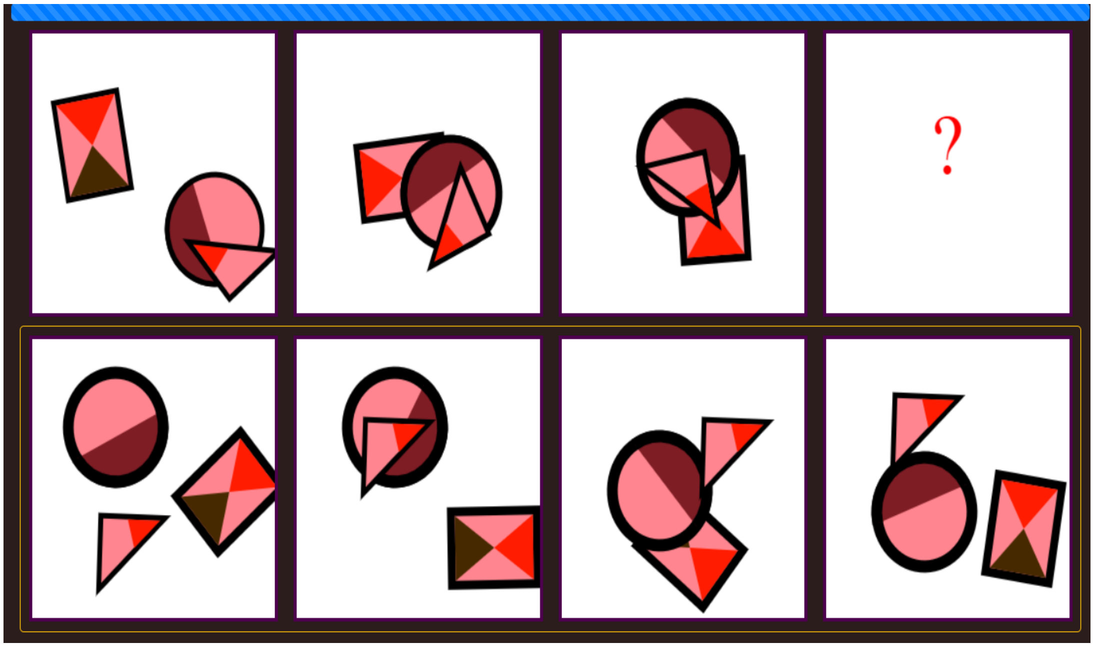


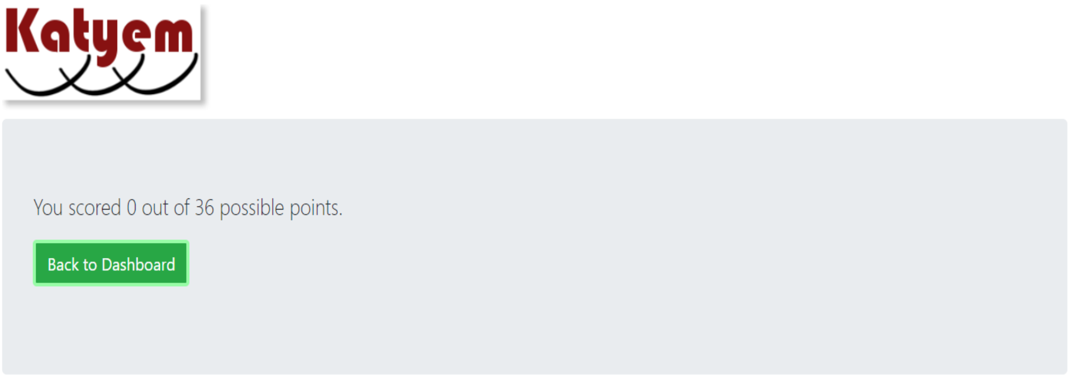








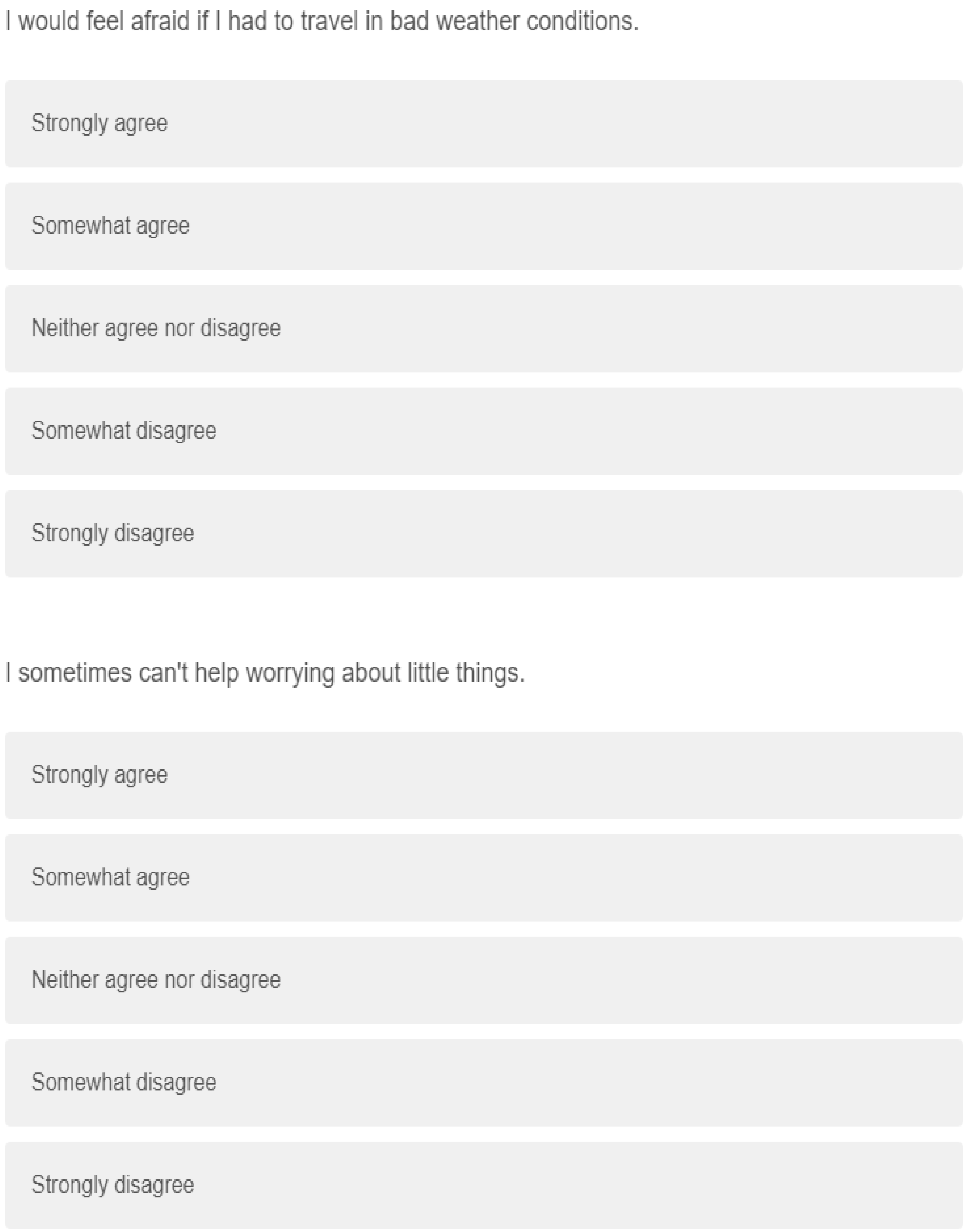




You scored 0 out of 36 possible points.

Appendix D

ITEMS OF THE HEXACO EMOTIONALITY MEASURE



I would feel afraid if I had to travel in bad weather conditions.

Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree

I sometimes can't help worrying about little things.

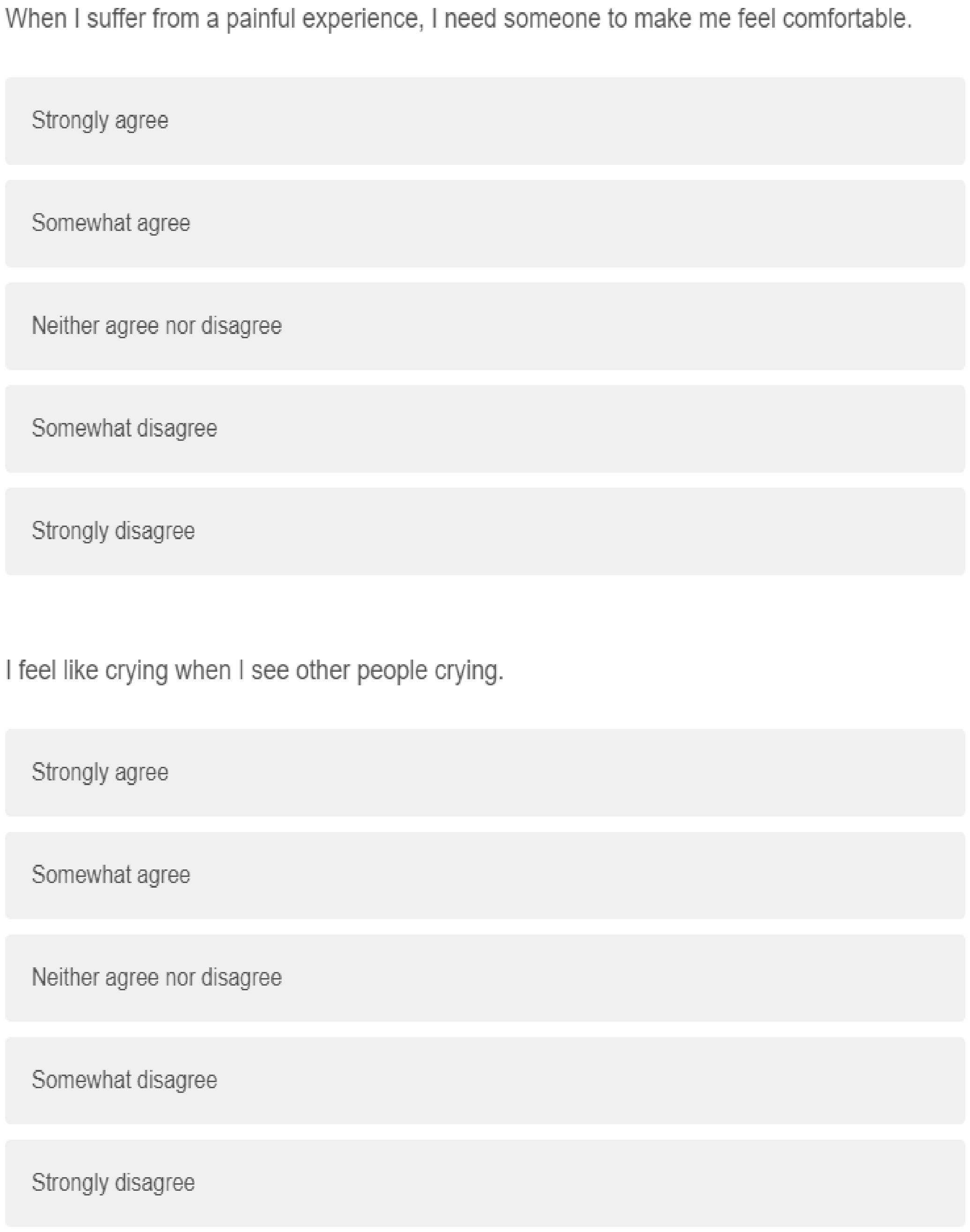
Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree



When I suffer from a painful experience, I need someone to make me feel comfortable.

Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree

I feel like crying when I see other people crying.

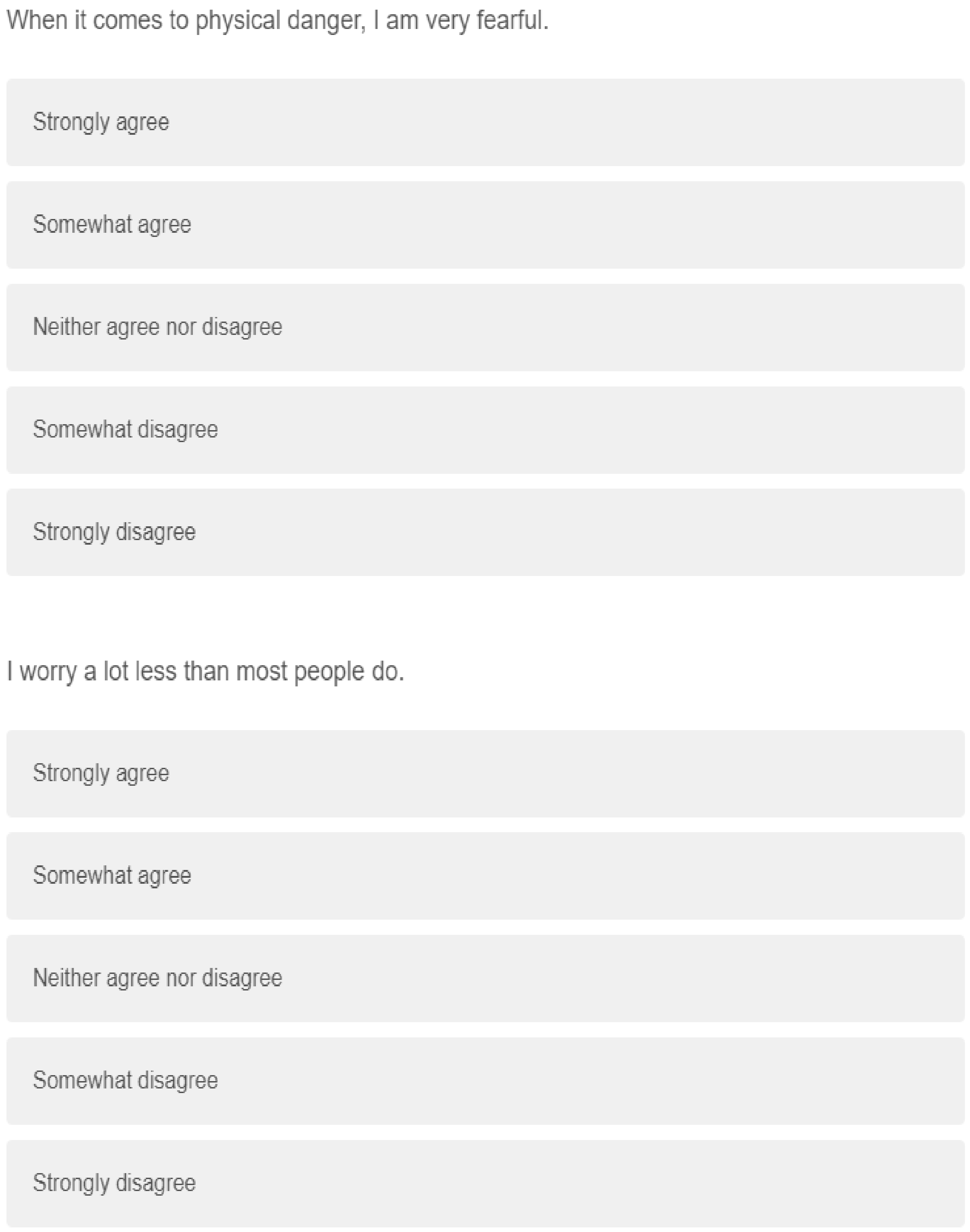
Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree



When it comes to physical danger, I am very fearful.

Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree

I worry a lot less than most people do.

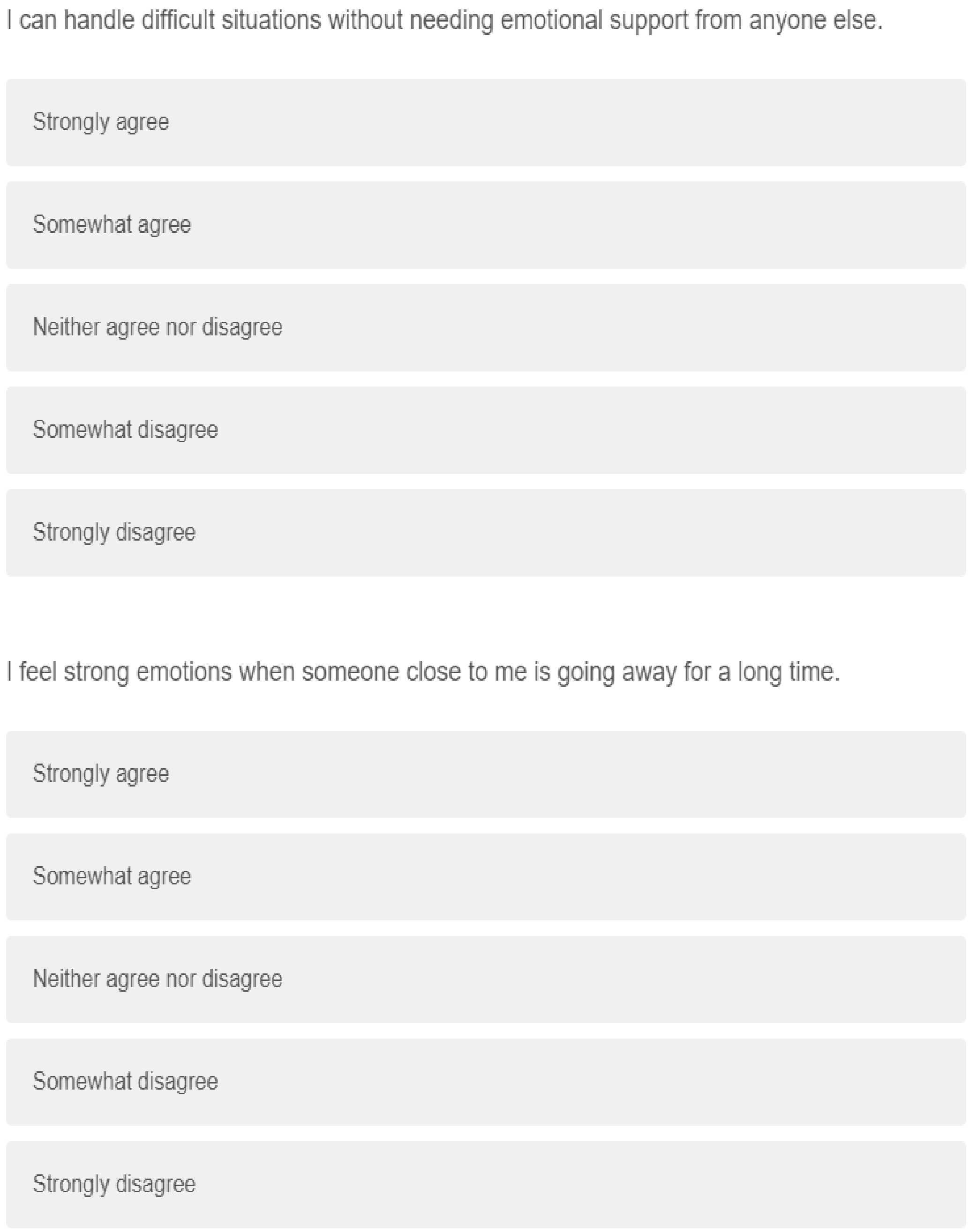
Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree



I can handle difficult situations without needing emotional support from anyone else.

Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree

I feel strong emotions when someone close to me is going away for a long time.

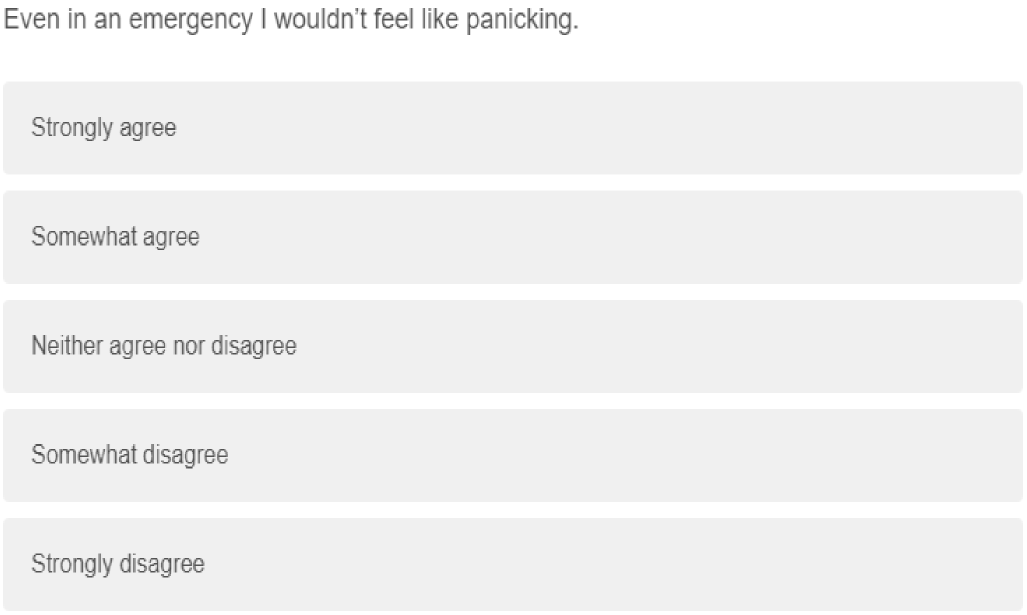
Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree



Even in an emergency I wouldn’t feel like panicking.

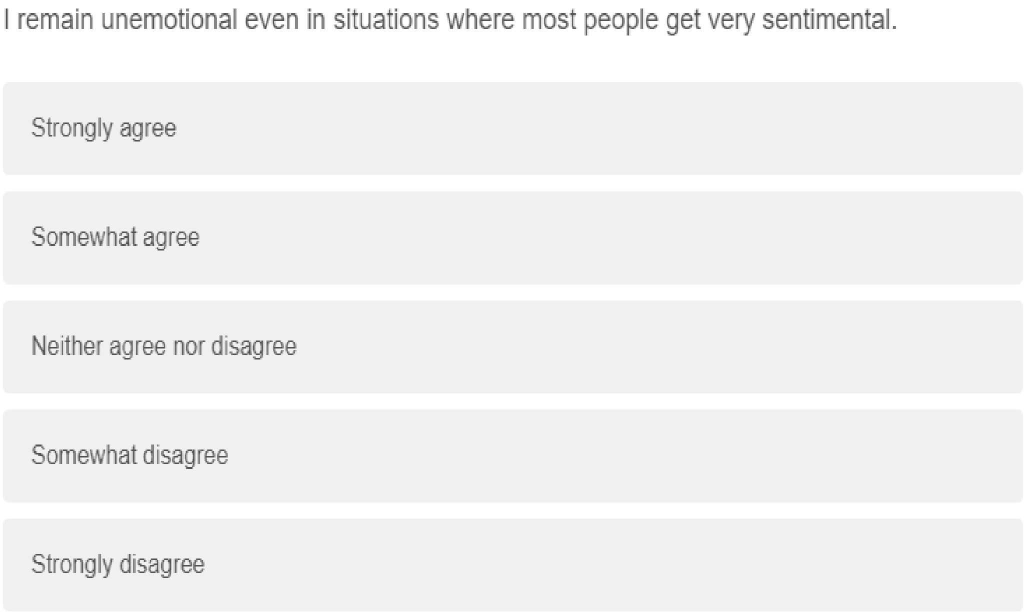
Strongly agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Strongly disagree



I remain unemotional even in situations where most people get very sentimental.

Strongly agree

Somewhat agree

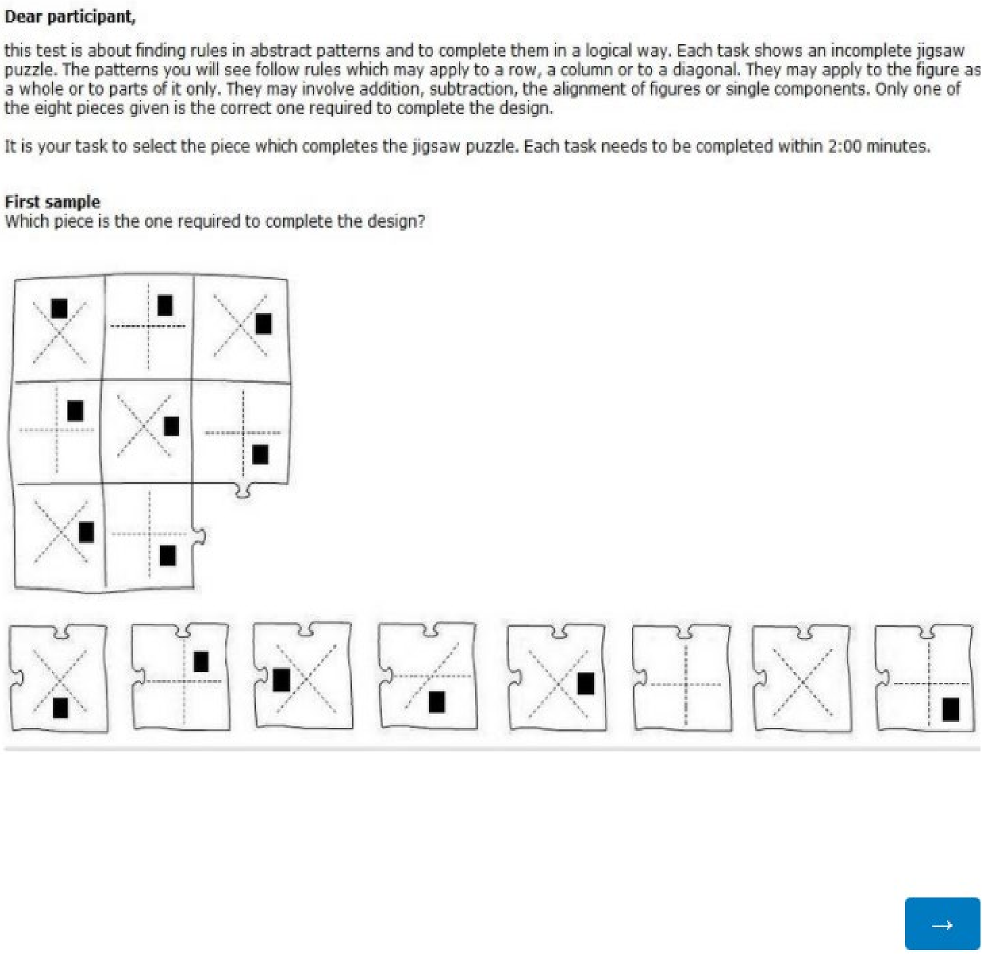
Neither agree nor disagree

Somewhat disagree

Strongly disagree

Appendix E

THE ITEMS IN THE HMT-S

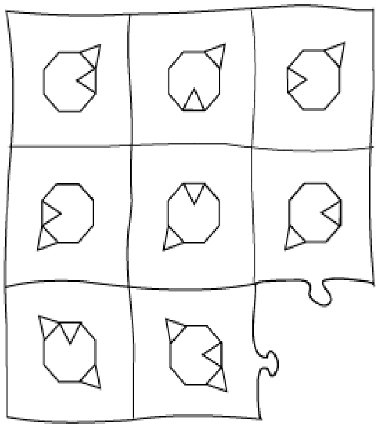


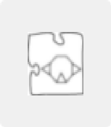
Dear participant,

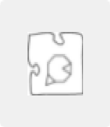
this test is about finding rules in abstract patterns and to complete them in a logical way. Each task shows an incomplete jigsaw puzzle. The patterns you will see follow rules which may apply to a row, a column or to a diagonal. They may apply to the figure as a whole or to parts of it only. They may involve addition, subtraction, the alignment of figures or single components. Only one of the eight pieces given is the correct one required to complete the design.

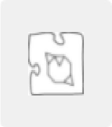
It is your task to select the piece which completes the jigsaw puzzle. Each task needs to be completed within 2:00 minutes.

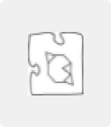
First sample
Which piece is the one required to complete the design?

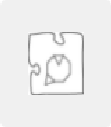


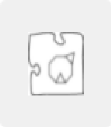


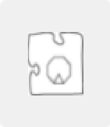


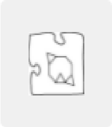




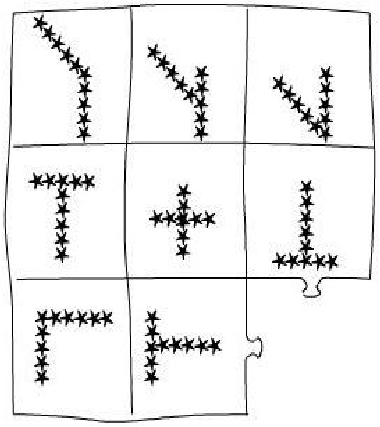


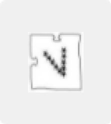


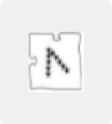


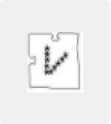


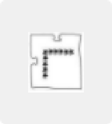
→

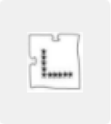


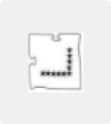


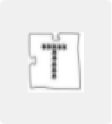


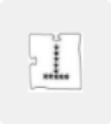




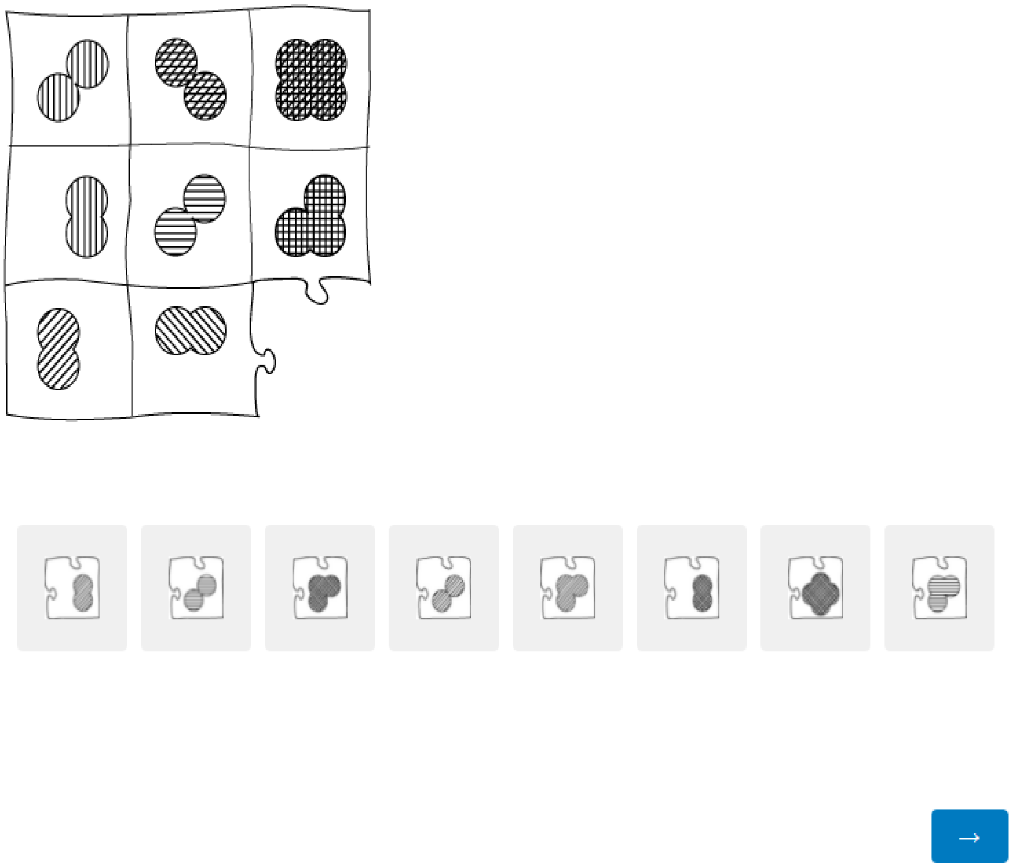


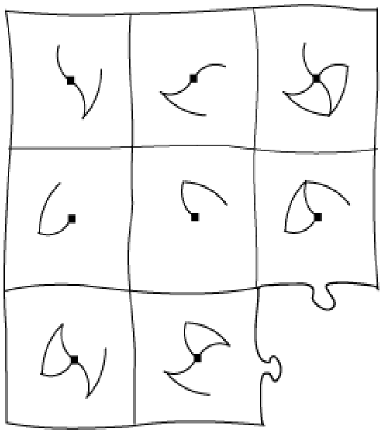


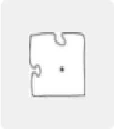


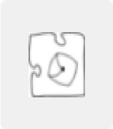


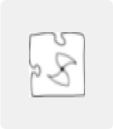
→

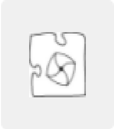


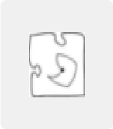


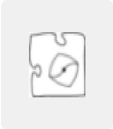


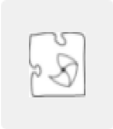


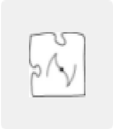




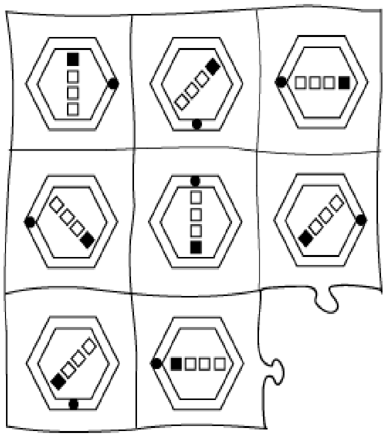


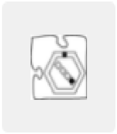


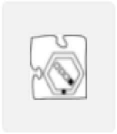


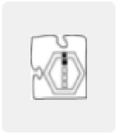


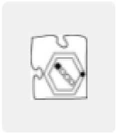
→

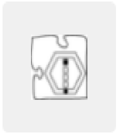


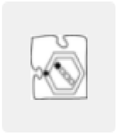


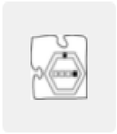


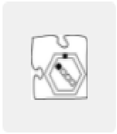




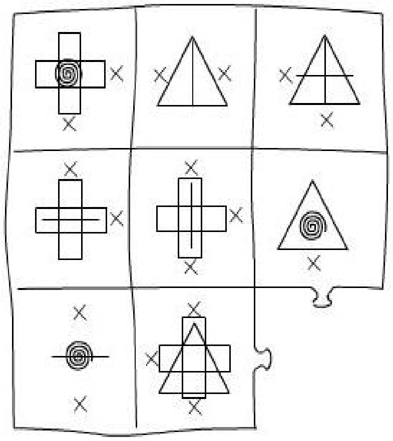


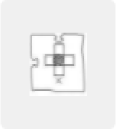


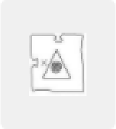


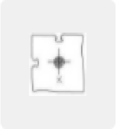


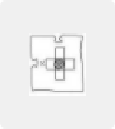
→

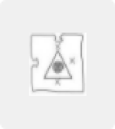


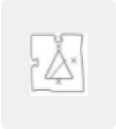


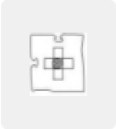


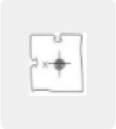








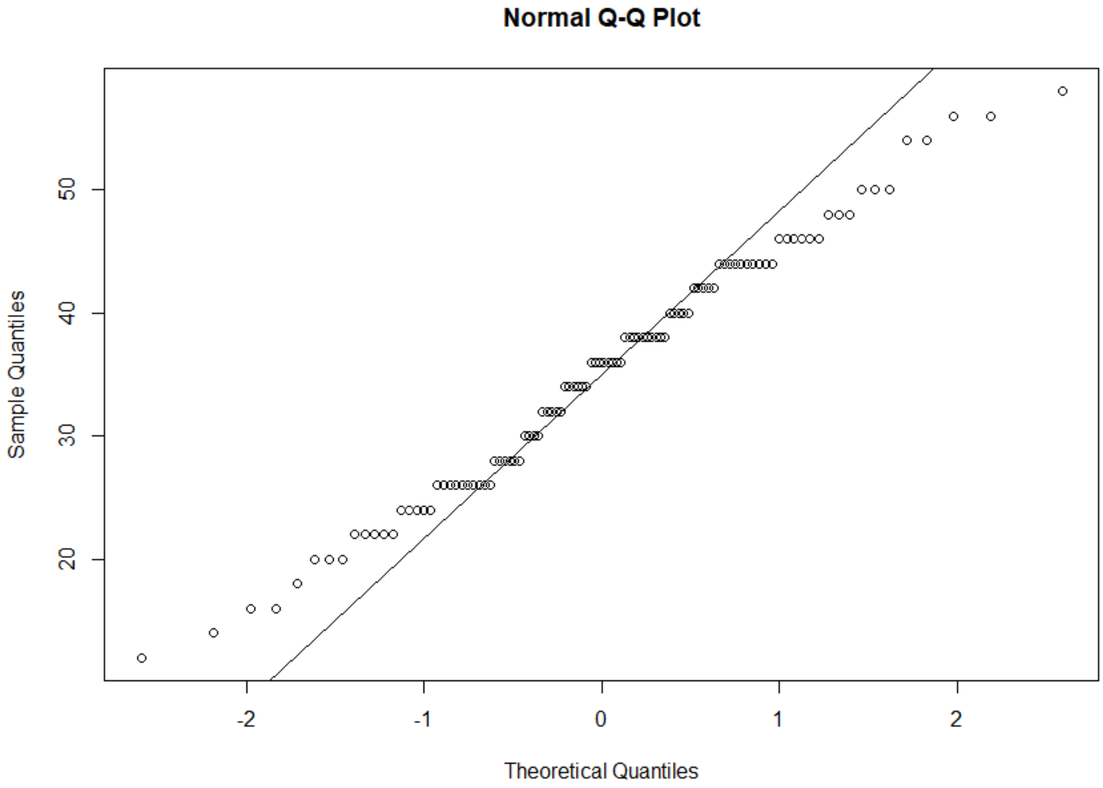




→

Appendix F

NORMAL QQ PLOT FOR SCORES IN STUDY 2


Normal Q-Q Plot
Sample Quantiles
20
30
40
50
-2
-1
0
1
2
Theoretical Quantiles

REFERENCES

Alves, C. B., Gierl, M. J., & Lai, H. (2010). Using automated item generation to promote principled test design and development. American Educational Research Association, Denver, CO, USA.

Anastasi, A. (1934). Practice and variability: A study in psychological method. *Psychological Monographs, 45*(5), i.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*(10), 1086.

Arendasy, M. E., Hergovich, A., & Sommer, M. (2008). Investigating the 'g'-saturation of various stratum-two factors using automatic item generation. *Intelligence*, *36*(6), 574-583.

Arendasy, M., & Sommer, M. (2007). Automatic generation of quantitative reasoning items: A schema-based isomorphic approach. *Learning and Individual Differences, 17*, 366–383. doi: 10.1027/1614-0001.27.1.2

Arendasy, M., & Sommer, M. (2010). Evaluating the contribution of different item features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence, 38*, 574–581. doi: 10.1016/j.intell.2010.06.004

Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences*, *22*(1), 112-117.

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*(4), 340-345.

Ashton, M. C., Lee, K., & Vernon, P. A. (2005). Gc and Gf are both valid intelligence factors: Commentary on Robinson (2005). *Personality and Individual Differences*, *39*(5), 999-1004.

Austin, E. J., Deary, I. J., Whiteman, M. C., Fowkes, F. G. R., Pedersen, N. L., Rabbitt, P., ... & McInnes, L. (2002). Relationships between ability and personality: Does intelligence contribute positively to personal and social adjustment?. *Personality and Individual Differences*, *32*(8), 1391-1411.

Barker, R. G. (1938). V. The Effect of Frustration Upon Cognitive Ability. *Journal of Personality*, *7*(2), 145-150.

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, *11*(1), 118.

Binet, A., & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année Psychologique, 11*(1), 191-244.

Binet, A., & Simon, T. (1916). New methods for the diagnosis of the intellectual level of subnormals.(L'Année Psych., 1905, pp. 191-244).

Blatter, M., Muehlemann, S., Schenker, S., & Wolter, S. C. (2016). Hiring costs for skilled workers and the supply of firm-provided training. *Oxford Economic Papers*, *68*(1), 238-257.

Borman, W. C., Hanson, M. A., Oppler, S. H., Pulakos, E. D., & White, L. A. (1993). Role of early supervisory experience in supervisor performance. *Journal of Applied Psychology*, *78*(3), 443.

Bors, D. A., & Vigneau, F. (2003). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences, 13*(4), 291-312.

Boyte-Eckis, L., Minadeo, D. F., Bailey, S. S., & Bailey, W. C. (2018). Age, gender, and race as predictors of opting for a midterm retest: A statistical analysis of online economics students. *Journal of Business Diversity, 18*(1).

Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: Stereotype threat and the race gap on raven's advanced progressive matrices. *Journal of Applied Psychology, 91*(4), 979.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. Guilford Publications.

Carretta, T. R., Zelenski, W. E., & Ree, M. J. (2000). Basic Attributes Test (BAT) retest performance. *Military Psychology*, *12*(3), 221-232.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge University Press.

Cattell, R. B. (1941). General psychology.

Collins, M. W., & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology*, *93*(2), 463.

DeVellis Robert, F. (2003). Scale development: theory and applications.

Dixon, R. A., Kurzman, D., & Friesen, I. C. (1993). Handwriting performance in younger and older adults: Age, familiarity, and practice effects. *Psychology and Aging, 8*(3), 360.

Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson counts model. *Learning and Individual Differences*, *52*, 121-128.

Dvorak, B. J. (1947). The new USES general aptitude test battery. *Journal of Applied Psychology*, *31*(4), 372.

Embretson, S., & Yang, X. (2006). *Item Response Theory*. Lawrence Erlbaum Associates Publishers.

Equal Employment Opportunity Commission, & Civil Service Commission. (1978). Department of Labor, & Department of Justice.(1978). Uniform guidelines on employee selection procedures. *Federal Register*, *43*(166), 38290-38315.

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence, 50*, 93-99.

Falleti, M. G., Maruff, P., Collie, A., Darby, D. G., & McStephen, M. (2003). Qualitative similarities in cognitive impairment associated with 24 h of sustained wakefulness and a blood alcohol concentration of 0.05%. *Journal of Sleep Research*, 12(4), 265-274.

Flanagan, D. P. (2000). Wechsler-based CHC cross-battery assessment and reading achievement: Strengthening the validity of interpretations drawn from Wechsler test scores. *School Psychology Quarterly*, *15*(3), 295.

Fleishman, E. A., & Fruchter, B. E. N. J. A. M. I. N. (1960). Factor structure and predictability of successive stages of learning Morse code. *Journal of Applied Psychology*, *44*(2), 97.

Fleishman, E. A., & Hempel Jr, W. E. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology*, *49*(5), 301.

Foster, D. F. (2010). Worldwide testing and test security issues: Ethical challenges and solutions. *Ethics & Behavior*, *20*(3-4), 207-228.

Freund, P. A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, *32*(3), 195-210

Geerlings, H., Glas, C. A., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*(2), 337.

Gierl, M. J., Ball, M. M., Vele, V., & Lai, H. (2015, June). A Method for Generating Nonverbal Reasoning Items Using n-Layer Modeling. In *International Computer Assisted Assessment Conference* (pp. 12-21). Springer, Cham.

Gierl, M., Lai, H., & Zhang, X. (2019). Automatic item generation. In *Advanced Methodologies and Technologies in Modern Education Delivery* (pp. 192-203): IGI Global.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757-765.

Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247-261.

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research*, *25*(1), 33-39.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance, 15*(1-2), 25-46.

Gottfredson, L. S. (2004). Intelligence: is it the epidemiologists' elusive" fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology, 86*(1), 174.

Grodsky, E., Warren, J. R., & Felts, E. (2008). Testing and social stratification in American education. *Annu. Rev. Sociol, 34*, 385-404.

Hausdorf, P. A., LeBlanc, M. M., & Chawla, A. (2003). Cognitive ability testing and employment selection: Does test content relate to adverse impact? *Applied H.R.M. Research*, 7(1-2), 41-48.

Hauser, D., Paolacci, G., & Chandler, J. J. (2018). Common Concerns with MTurk as a Participant Pool: Evidence and Solutions. https://doi.org/10.31234/osf.io/uq45c

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 87*(2), 243.

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence, 48*, 1-14.

Heydasch, T. (2014). The Hagen matrices test (HMT).

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104-121.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. *Woodcock-Johnson Technical Manual*, 197-232.

Hornby, G. (2011). Parental involvement in childhood education: Building effective school-family partnerships. Springer Science & Business Media.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*(1-2), 152-194.

Hunter, J., & Hunter, R. (1984). Validity and utility of alternative predictors across studies. *Psychological Bulletin, 96*(1), 72-98.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594.

Irvine, J. (Ed.). (2002). In search of wholeness: African American teachers and their culturally specific classroom practices. Springer.

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality and Individual Differences, 13*(8), 869-881.

Jensen, A. R. (1998). *The g factor: The science of Mental Ability* (Vol. 648): Praeger Westport, CT.

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business and Psychology*, *29*(4), 555-572.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, *47*(7), 635-650.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 453-456

Kock, F. D., & Schlechter, A. (2009). Fluid intelligence and spatial reasoning as predictors of pilot training performance in the South African Air Force (SAAF). *SA Journal of Industrial Psychology*, *35*(1), 31-38.

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A Cost–Benefit Analysis of Automatic Item Generation. *Educational Measurement: Issues and Practice*, *38*(1), 48-53.

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*(2), 435-447.

Lai, H., Alves, C., & Gierl, M. J. (2009, April). Using automatic item generation to address item demands for CAT. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Minneapolis, MN: IACAT.

LeBlanc, P. A. H. M. M., & Chawla, A. (2003). Cognitive ability testing and employment selection: Does test content relate to Adverse Impact?. *Applied HRM Research*, *7*(2), 41-48.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications, 157*, 17.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*(6), 1672.

Link, H. C. (1919). Employment psychology: The application of scientific methods to the selection, training and grading of employees. Macmillan.

Loehlin, J. C., & Lindzey, G. 8: Spuhler, JM (1975). Race differences in intelligence.

Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form—a confrontation of Birnbaum's logistic model. *Psychometrika*, *35*(1), 43-50.

Matthews, T. D., & Lassiter, K. S. (2007). What does the Wonderlic personnel test measure?. *Psychological Reports*, *100*(3), 707-712.

Matton, N., Vautier, S., & Raufaste, E. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence, 37*(4), 412-421.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*(2), 191.

Moutafi, J., Furnham, A., & Paltiel, L. (2004). Why is conscientiousness negatively correlated with intelligence?. *Personality and Individual Differences*, *37*(5), 1013-1022.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: new findings and theoretical developments. *American Psychologist*, *67*(2), 130.

Olenick, J., Bhatia, S., & Ryan, A. M. (2016). Effects of g-Loading and Time Lag on Retesting in Job Selection. *International Journal of Selection and Assessment, 24*(4), 324-336.

Park, H. S., Baker, C., & Lee, D. W. (2008). Need for cognition, task complexity, and job satisfaction. *Journal of Management in Engineering*, *24*(2), 111-117.

Penfield, R., & Camilli, G. (2007). Test fairness and differential item functioning. *Handbook of Statistics*, *26*, 125-167.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153-172.

Poinstingl, H. (2009). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychological Test and Assessment Modeling*, *51*(2), 123.

Reeve, C. L., & Lam, H. (2007). The relation between practice effects, test-taker characteristics and degree of g-saturation. *International Journal of Testing, 7*(2), 225-242.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer III, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*(2), 297-330.

Sackett, P. R., Burris, L. R., & Ryan, A. M. (1989). Coaching and practice effects in personnel selection.

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.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*(3), 432.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*, *66*(2), 166.

Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. *Annual Review of Organizational Psychology and Organizational Behavior*, 1(1), 45–65.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence.

Schuster, M. H., & Miller, C. S. (1981). Evaluating the Older Worker: Use of Employer Appraisal Systems in Age Discrimination Litigation. *Aging and Work: A Journal on Age, Work and Retirement, 4*(4), 229-243.

Society for Industrial, Organizational Psychology (US), & American Psychological Association. Division of Industrial-Organizational Psychology. (2003). *Principles for the Validation and Use of Personnel Selection Procedures*. The Society.

Spearman, C. (1904). Measurement of association, Part II. Correction of 'systematic deviations'. *Am J Psychol, 15*, 88-101.

Spearman, C. E. (1938). Mesure de l'intelligence.

Srikanth, P. B. (2020). The relative contribution of personality, cognitive ability and the density of work experience in predicting human resource competencies. *Personnel Review*.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5), 797.

Stern, W. (1912). The psychological methods of intelligence testing. *G. Whipple, Trans.). Baltimore: Warwick and York.*

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41*(1), 37.

te Nijenhuis, J., van Vianen, A. E., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence, 35*(3), 283-300.

Tett, R. P., & Meyer, J. P. (1993). Job satisfaction, organizational commitment, turnover intention, and turnover: path analyses based on meta-analytic findings. *Personnel Psychology*, *46*(2), 259-293.

Thurstone, L. L. (1938). *Primary mental abilities* (Vol. 119). Chicago: University of Chicago Press.

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin*, *140*(5), 1332–1360. https://doi.org/10.1037/a0037173

Wainer, H. (2002). On the automatic generation of test items: Some whens, whys, and hows. *Item Generation for Test Development*, 287-305.

Waschl, N. A., Nettelbeck, T., Jackson, S. A., & Burns, N. R. (2016). Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability. *Personality and Individual Differences*, *100*, 157-166.

Wilk, S. L., & Sackett, P. R. (1996). Longitudinal analysis of ability-job complexity fit and job change. *Personnel Psychology*, *49*(4), 937-967.

Weschler, D. (1955). Weschler Adult Intelligence Scale. Manual. In: New York, USA, Psychological Corporation.

Wonderlic, E. F. (2007). Wonderlic personnel test-revised: Manual. *Los Angeles, CA: Western Psychological Services*.

Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests*: H. Holt.