The Pennsylvania State University

The Graduate School

College of Education

USING ITEM RESPONSE THEORY TO DEVELOP A RAVEN'S MATRICES SHORT FORM FOR PAKISTANI ADOLESCENTS AND YOUNG ADULTS

A Dissertation in

School Psychology

by

Hongxuan Zhong

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

May 2019

The dissertation of Hongxuan Zhong was reviewed and approved* by the following:

Barbara A. Schaefer
Associate Professor of Education (School Psychology)
Educational Psychology, Counseling, and Special Education
Dissertation Co-Adviser
Co-Chair of Committee

James C. DiPerna
Professor of Education (School Psychology)
Professor in Charge for Graduate Programs in School Psychology
Educational Psychology, Counseling, and Special Education
Dissertation Co-Adviser
Co-Chair of Committee

Pui-Wa Lei
Professor of Education (Educational Psychology)
Educational Psychology, Counseling, and Special Education

Pamela M. Cole
Liberal Arts Professor of Psychology and Human Development and Family Studies
Department of Psychology, College of Liberal Arts

*Signatures are on file in the Graduate School

ABSTRACT

The Raven's Matrices are a group of non-verbal tests designed to measure eductive ability (Raven, Raven, & Court, 1998a). The length of the 72-item Raven's Combined Matrices (RCM) can limit its application in large-scale research studies, as well as potentially cause physical fatigue and/or emotional distress for test takers. Although several research teams have created short forms for different versions of the Raven's Matrices tests (Arthur & Day, 1994; Bilker et al., 2012; Bors & Stokes, 1998; Sefcek, Miller, & Figueredo, 2016; Wytek, Opgenoorth, & Presslich, 1984), few have used modern test theories to do so. In addition, no short forms have been developed for use in Pakistan to date. As such, the purpose of this study was to use Item Response Theory to develop a RCM short form for potential use with adolescents and young adults in Pakistan.

Data were drawn from a longitudinal surveillance follow-up study conducted in Pakistan (N = 1,405). A split-sample approach was used for parameter estimation and validation, along with cross-validation to verify results. Typically developing adolescents and young adults were of interest in this study. As such, items that provided most information around the middle range of the ability continuum were selected to construct the short form. The resulting 10-item RCM short provides similar levels of test information to the RCM long form, maintains the maximum amount of information in the middle of the ability range, demonstrates acceptable reliability for research purposes, and is strongly correlated with the RCM long form. Results, however, also indicated instability in parameter estimation to a certain degree. As such, replication and additional psychometric studies are essential prior to any use of the RCM short form in research or practice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the professional and personal support of a number of individuals. I would like to express my deepest appreciation to my doctoral committee. I am thankful to Dr. Barbara Schaefer for providing me with the opportunities to pursue my research interests over the years and for supporting me in bringing this project to fruition. I am equally grateful to Dr. Jim DiPerna, who fundamentally shaped this dissertation, for his relentless support, guidance, and kindness throughout the duration of this project. Further, I would like to thank Drs. Pui-Wa Lei and Pamela Cole, who contributed invaluable advice and insights.

I must thank Dr. Zeba Rasmussen at the NIH's Fogarty International Center for generously sharing her data with me. My appreciation also goes to Dr. Shirley Woika for her unwavering support and encouragement not only during the dissertation process but also throughout other milestones of my graduate career. I would like to extend my sincere thanks to Drs. Melanie Pellecchia, Keiran Rump, and David Mandell at the University of Pennsylvania for their support and guidance throughout my internship year. In addition, I would like to gratefully acknowledge the administrative support that I received from Ms. Samantha Walker, who is always understanding, kind, and efficient.

I am deeply indebted to my beloved parents, who amazingly have always been able to find the strength to understand, accept, and support all the decisions that I have made over the years, even when we viewed things differently at times. To the best mentor that I ever had, "Papa Tom" (Dr. Thomas D. Oakland): Although you are unable to see the completion of my graduate education, your sense of mission and leadership, your passion to make the world a better place for children, and your wit and wisdom have been perpetual sources of inspiration to me. To my cohort mates and dear friends, Susan Crandall and Molly Kaufman: I cannot imagine making this journey without you, and I will always remember our late-night conversations, the celebrations at Kamrai and Tadashi, and our very famous "Bob Fan Blub" in State College.

I am grateful to the people I have met and for all the experiences over the years, which have reshaped my worldview, expanded my understanding of humanity, and made me realize the power of love, compassion, and kindness. And the journey continues…

Chapter 1

INTRODUCTION

The Raven's Matrices tests frequently have been utilized as nonverbal measures of abstract reasoning in many research and applied settings in the United States (Mills & Tissot, 1995; Raven, 1989; Wiley, Jarosz, Cushen, & Colflesh, 2011). Internationally, the Raven's Matrices tests also have been widely used because they require minimal language to complete and are believed to reduce the impact of cultural factors on test performance (Neisser, 1998; Raven et al., 1998; Raven, 1989; Raven & Raven, 2003). Despite the strengths, the significant amount of time required to complete the full-length Raven's Matrices tests (e.g., 40–60 minutes for the Advanced Progressive Matrices) limits their utility in large-scale research and practice (Arthur & Day, 1994; Bilker et al., 2012; Sefcek et al., 2016). In response to this limitation, several research teams have attempted to develop short forms for the tests (e.g., Arthur & Day, 1994; Bilker et al., 2012; Bors & Stokes, 1998; Caffarra, Vezzadini, Zonato, Copelli, & Venneri, 2003). These efforts have resulted in Raven's short forms ranging from 9–48 items.

Although some of these Raven's short forms have demonstrated promise with regard to their psychometric properties, all have been developed for use with Western populations (e.g., the United States, Canada, and Austria). As such, their validity for use with other cultures is, at best, unknown. Nonetheless, studies (e.g., Murray-Kolb et al., 2014) featuring the full-length Raven's Combined Matrices (RCM) have been completed in developing nations and encountered similar challenges with its use (e.g., a significant amount of recourses required for large-scale research studies, physical fatigue and emotional distress for test takers). To address these challenges, the primary aim of this

study was to develop an RCM short form for use with adolescents and young adults in one developing nation, Pakistan.

Cultural Context

Pakistan, officially known as the Islamic Republic of Pakistan, is a lower-middle-income country (World Bank, 2017). Located in the Eastern Hemisphere on the continent of South Asia, Pakistan shares international land borders with four nations: Afghanistan, China, India, and Iran. The territory of Pakistan covers a total area of 796,095 square kilometers, making it the 36th largest nation in the world (United Nations, 2017). With a population exceeding 207.9 million people, Pakistan is equivalent to 2.77% of the total world population and is the sixth most populous country in the world (U.S. Census Bureau, 2018). The median age in Pakistan is 23.8 years (male: 23.7 years, female: 23.8 years), and sex ratio of the total population is 1.03 male(s)/female. Pakistan ranks as the 40th largest economy in the world, with an estimated gross domestic product (GDP) of approximately 305 billion dollars (World Bank, 2017). Pakistan, as a developing nation, continues to face a multitude of challenges, including but not limited to illiteracy, poverty, and health issues.

Education. Education in Pakistan is operated and regulated by the Federal Ministry of Education and the provincial governments. The constitution of Pakistan requires the state to provide free and compulsory education to children and adolescents ages 5–16 (Malik et al., 2015). Specifically, the public education system in Pakistan is generally divided into six levels: Preschool (3–5 years old), primary (Grades 1–5), middle (Grades 6–8), high (Grades 9–10), intermediate (Grades 11–12), and university programs leading to undergraduate and graduate degrees (Malik et al., 2015; UNESCO, 2018).

Although the Pakistan government has expressed commitment to high-quality education through the development and implementation of national policies, the country continues to suffer from education challenges in practice (Naviwala, 2015). The primary school enrollment rate remains low in Pakistan. Between 2012–2013, the number of out-of-school children was estimated to be 6.7 million (Malik et al., 2015), with roughly 55% being girls. According to a national review in Pakistan (Malik et al., 2015), individuals living in remote and rural areas (60% of Pakistan's total population) have limited access to high-quality education. Girls from extremely low-income families are most disadvantaged, with more than half of them having no access to formal education. Lack of formal schooling can affect an individual's ability to read and write. The literacy rate of adolescents and young adults (15–24 years old) was estimated to be 71.6% (Malik et al., 2015). Literacy rates vary by sex, with a relatively lower rate for females (63.3%) and a higher rate for males (79.4%). Literacy rates also vary by region, with lower rates in rural areas (Malik et al., 2015).

Poverty. Poverty further exacerbates the challenges that hinder improvement and reform of the education system in Pakistan (Malik et al., 2015; Naviwala, 2015). Pakistan's spending on education is low, with 2.4% of the total GDP being allocated to schools (Malik et al., 2015). Schools often struggle with inadequate or severely deficient funding to provide pedagogical training to teachers or to purchase class materials (Khattak, 2012). Consequently, students need to deal with many issues that may interfere with their learning, including limited class materials, poor learning environments, absent or unqualified teachers, and lack of adequate cognitive stimulation in classrooms overall (Khattak, 2012; Murray-Kolb et al., 2014). These issues eventually manifest themselves

in the gap between policy and practice. For example, certain school subjects (e.g., science, art, and arithmetic) in the standard curriculum are often neglected or eliminated in practice due to the shortage of qualified teachers and school supplies (Khattak, 2012).

Moreover, some individuals are completely excluded from educational access and opportunities due to poverty. The incidence of child labor tends to increase in low-income families, especially with a greater number of boys dropping out of school to support their families financially (Malik et al., 2015). Similarly, girls in low-income families are often expected to stay at home to take care of their younger siblings and to complete chores assigned by their parents (Malik et al., 2015).

Health. Pakistan has faced long-standing public health challenges for decades. Individuals are at high risk of being exposed to infectious diseases, malnutrition, contaminated drinking water, unsanitary facilities, and other environmental risks that can potentially impact their development (UNICEF, 2016; United Nations, 2017; WHO, 2010). Despite decades of efforts, the rates of many health issues have remained stubbornly high in Pakistan, with infectious diseases and malnutrition in particular being the most prevalent. Major leading infectious diseases affecting Pakistani populations include hepatitis C, tuberculosis, malaria, and rabies (Sultan & Khan, 2013). These infectious diseases can significantly impact youth development and even threaten livelihoods.

Additionally, between 2008–2012, the prevalence of various conditions associated with malnutrition was high in Pakistan: 11.6% for underweight (low-weight-for-age), 43.7% for stunting (low-height-for-age), 15.1% for wasting (low-weight-for-height), and 6.4% for overweight, according to UNICEF (2013). Malnutrition can lead to irreversible

intellectual and physical damages and remains a major contributor to health impairment in Pakistan (Bhutta et al., 2013; Guerrant, Oriá, Moore, Oriá, & Lima, 2008). Stunting in particular can impact cognition later in life, academic progress, or both (Grantham-McGregor et al., 2007). The interaction of the malnutrition and infectious diseases in Pakistan is particularly alarming given that nutrient deficiencies can leave individuals more debilitated and vulnerable to infectious diseases, and certain infections in turn can exacerbate malnutrition (Goldstein, Katona, & Katona-Apte, 2008).

Monitoring Cognitive Development

As a rapidly expanding nation, Pakistan has invested heavily in addressing a complex set of societal challenges, including education policies (Malik et al., 2015), social and economic inequality (Malik et al., 2015; Naviwala, 2015), and disease (UNICEF, 2016; United Nations, 2017; WHO, 2010), that can affect human development and health. Over the past few decades, considerable efforts have been made to improve human development through education, research, and outreach services in Pakistan (Malik et al., 2015). In order to determine whether or not such efforts are effective, population health studies (i.e., the analysis of population and health using quantitative and qualitative methods) are critical. Among many variables of interest, cognitive development is considered one of the most important outcome domains in population health sciences. Cognitive functioning is a significant predictor of future life outcomes (Martinez, 2010), and cognitive impairment can be long-lasting and irreversible (Grantham-McGregor et al., 2007; Martinez, 2010; Murray-Kolb et al., 2014; Sattler, 2008). Thus, it is important that researchers and healthcare professionals in Pakistan are able to monitor the trajectory of cognitive development across the lifespan.

Understanding the impact of complex societal conditions on the health of populations and evaluating health improvement efforts requires large-scale, longitudinal studies. Such studies not only deal with massive amounts of data with a large number of independent and moderating variables, but also involve longitudinal follow-ups and constant data monitoring. As cognitive testing is often included as part of a comprehensive measurement battery in such studies, measurement efficiency becomes critical (Bilker et al., 2012). For example, a large-scale population health surveillance study was conducted in Pakistan from 1989–1996 (Shah et al., 2015). Follow-up studies were completed from 2012–2014 (Shah et al., 2015), and the RCM was utilized in these studies to measure participants' cognitive functioning and inform health improvement efforts in Pakistan. One significant limitation to using the full-length RCM, however, was that it required significant time and resources to complete (i.e., 60–80 minutes for a single test administration). Similarly, researchers conducting large-scale, longitudinal research studies in Pakistan also need to take into account the response burden associated with the full-length RCM, especially when collecting data with study participants who did not receive formal education or have prior experience with formal testing.

Therefore, the ultimate goal of this study was to develop a RCM short form for use with adolescents and young adults in Pakistan. Such a short form could be particularly useful for large-scale, longitudinal research studies in Pakistan as it would require less time to administer and decrease response burden on test takers.

Short Form Development

Rationale for creating short forms. Measures that are lengthy may need to be streamlined for certain applications, and this process is often referred to as short form

development. In general, short form development is intended to increase measurement efficiency, which can be achieved by removing redundant items from a lengthy measure and preserving items that are psychometrically sufficient to represent the respective full-length measure (i.e., without compromising reliability and validity). Given that short forms reduce the amount of time required for test administration, short form development can increase the feasibility and use of objective measures in applied settings.

Similar to the importance of contextual fit when implementing evidence-based interventions (Coles et al., 2017; May, Johnson, & Finch, 2016; Pfadenhauer et al., 2017), the match between a measure and the specific context in which the measure is to be used is critical. In the context of large-scale and/or longitudinal research (e.g., studies that assess multiple variables over time) or practice (e.g., universal screening that involves a large number of examinees), significant resources (e.g., time, money, human, and training) are often required. For example, despite the benefits of utilizing objective measures to inform decisions, practitioners often encounter barriers (e.g., time required for administration) to incorporating such measures into routine clinical practice for longitudinal progress monitoring (Waldrop & McGuinness, 2017). Short forms, however, can address at least some of these barriers by reducing the amount of resources needed when used in these contexts.

Short form development can also address challenges at the individual level. Measures that are perceived as being lengthy by test-takers have been found to result in lower response rates (Stanton, Sinar, Balzer, & Smith, 2002). The degree of cognitive effort required to complete a measure usually increases as the length of the measure expands. Overly burdensome testing associated with lengthy measures can cause

physical exhaustion and mental frustration for test takers (Harvey, 2012). One way to reduce response burden is to develop short forms by retaining a minimally sufficient number of items to reduce completion time while maintaining the psychometric integrity of scores. Such short forms can provide an adequate amount of information for decision making yet are more manageable for test-takers to complete.

Measurement theories for short form development. Classical Test Theory (CTT) and Item Response Theory (IRT) are widely used measurement frameworks for test construction, including short form development. CTT is also known as true score theory, which conceptualizes that the observed score equals the true score plus the random error of measurement (Gulliksen, 1950; Lord & Novick, 1968). Under the CTT framework, the random error of measurement reflects score fluctuations due to error and is constant among scores in the same population. Test reliability based on CTT is measured by the ratio of the true score variance to the observed score variance, and it usually provides a single value for the test as a whole. Although CTT has been used widely to develop measures in psychology and education, it has been criticized for its shortcomings such as sample dependence, item/test dependence, and lack of accounting for guessing (Hambleton & Jones, 1993).

IRT represents an important innovation in the field of psychometrics. Compared CTT, IRT is characterized as a *modern* approach for test development and validation because it was developed in the most recent three decades in order to address the abovementioned shortcomings of CTT (Hambleton & Jones, 1993). IRT refers to a paradigm in psychometrics in which the probability of correctly answering a test item is described as the function of a latent trait of interest, the θ parameter (Hambleton &

Swaminathan, 1985). The latent trait, theta (θ), can be scaled with a mean of zero and a standard deviation of one. Theoretically, the range of ability is from negative infinity to positive infinity; however, ability levels are usually limited to be within the range between −3 and +3 in practice. Under an IRT framework, items are characterized on three parameters, including item discrimination parameter (*a*; slope parameter), item difficulty (*b*; threshold parameter), and guessing on the probability of a correct response (*g*; guessing parameter).

During the past three decades, IRT has been gradually replacing CTT as the major theoretical framework for test development due to several notable limitations of CTT (Hambleton & Jones, 1993). First, under the CTT framework, the estimates of test and item parameters are sample dependent. As such, results based on CTT cannot be generalized to a different sample. Second, the estimation of an individual's skill level based on CTT depends on the particular sample of test items chosen from the item pool, meaning classical statistics are tied to a certain test form. Third, scoring in classical test theory does not take into account guessing on the probability of a correct response as the amount of measurement error is assumed to be the same for each test taker with CTT. In contrast, results from IRT based statistics are sample- and item-independent when the assumptions of the chosen models are sufficiently satisfied. Moreover, the amount of measurement error is allowed to vary for each test taker under IRT, and guessing can be taken into consideration.

IRT methods have been used by researchers in the most recent decade to develop short forms due to its notable advantages (Anthony, DiPerna, & Lei, 2016; Edelen & Reeve, 2007; Locke et al., 2012). IRT provides researchers with more sophisticated

information and greater flexibility when evaluating individual test items. As a result of these properties, IRT can better inform item selection for the development of short forms. Furthermore, based on the IRT approach, reliability can be estimated across the range of the latent trait (e.g., ability) based on the amount of information provided relative to the theta level (e.g., ability level). After estimating the item parameters of an IRT model, Item Information Function (IIF) curves can also be generated, revealing the amount of information a single item provides at different ability levels on the θ scale. All of the individual IIF curves can be added at each ability level to produce a Test Information Function (TIF) curve, which indicates the amount of information that the entire test contains at different ability levels on the θ scale.

TIF curves allow researchers to examine item information (akin to reliability under CTT) of each individual item at a given ability level of interest. Such a mechanism can be used for selecting a representative set of items to target a certain population of interest with greater precision. For example, if the target population is individuals with typical or average cognitive ability, test items that contribute most information around the middle of the ability continuum would be selected. Alternatively, a study involving individuals with severe cognitive deficits would benefit most from items that provide precision at the lower end of the ability continuum, whereas a study focusing on gifted individuals would prioritize items providing information at the higher end.

In sum, although efficient assessments are needed for assessing the cognitive skills of adolescents and young adults in Pakistan to monitor cognitive growth over time, no short forms have been developed and validated for use with Pakistani populations. Such a short form will decrease the amount of time and resources required to collect data

in large-scale research studies and reduce response burden on test takers. As such, the current study used a modern test theory (i.e., Item Response Theory) to inform the development of a RCM short form. The target population for the RCM short form was typically developing adolescents and young adults.

Chapter 2

LITERATURE REVIEW

Raven's Matrices Tests

The Raven's Matrices tests (Raven et al., 1998a) are a group of nonverbal tests designed to measure an individual's eductive ability of Spearman's *g*. Specifically, eductive ability involves "the ability to make meaning out of confusion; the ability to forge largely non-verbal constructs which make it easy to handle complexity" (Raven et al., 1998, p. G1). It requires test takers to conceptualize spatial design and to reason by analogy at levels ranging from simple to complex. The Raven's Matrices tests can be either administered individually or used in group testing, and the interpretation of test scores is straightforward.

The Raven's Matrices tests are available in three different formats (Raven et al., 1998a): The Standard Progressive Matrices (SPM), the Coloured Progressive Matrices (CPM), and the Advanced Progressive Matrices (APM). The SPM originally was published in 1938 and followed by the CPM and APM in 1947 (Raven et al., 1998a). Each of these forms were developed for use with a different target population (Raven et al., 1998a). Specifically, the SPM was developed for the general population; whereas the APM was developed for the top 20% of the population. The CPM was intended for young children ages 5–11 years, the elderly population, and mentally and physically impaired individuals. The SPM and CPM can be administered together as the Raven Combined Matrices (RCM) to assess general intellectual ability (Raven et al., 1998a, 1998b), providing a broader range of difficulty levels and allowing comparisons across a wide range of ages. Each item set becomes progressively more difficult.

The Raven's SPM and CPM have been most used widely in both research and practice in the United States (Greenfield, 1998; Jensen, 1974; Simpson, Tate, & Weeks, 2005) and abroad (Flynn, 1987, 2009; Murray-Kolb et al., 2014; Soofi et al., 2013) . Initially developed for research purposes, the Raven's Matrices tests have been used in at least 2,500 studies focusing on cognitive development since the tests were published (J C Raven Ltd, n.d.). Particularly, the Raven's Matrices tests have also been frequently used to research how genetic and environmental factors could impact cognitive development (Grantham-McGregor et al., 2007; Murray-Kolb et al., 2014; Raven et al., 1998a; Raven & Raven, 2003). In practice, the Raven's Matrices tests have been widely applied in educational, occupational, and clinical settings throughout the world (J C Raven Ltd, n.d.). As such, researchers from many countries around the world have independently developed national and regional norms for use in practice across different age, education, and/or ethnic groups in their own countries, including Australia, America, Belgium, Brazil, Czechoslovakia, Cuba, China (Mainland), Canada, France, Germany (East), Germany (West), Great Britain, Hong Kong, Ireland, Iran, Iraq, India, Netherlands, Puerto, Peru, Rico, Slovakia, Switzerland, Spain, and Taiwan (Raven et al., 1998b).

The Raven's Matrices tests have been widely used in international contexts primarily due to three reasons. First, the Raven's Matrices tests were developed with reduced cultural concepts that can impact test performance (Neisser, 1998; Raven, 2000). Second, the administration of the Raven's Matrices tests is independent of spoken and written language (Raven et al., 1998a). Third, the Raven's Matrices tests have generally gained popularity due to their relative ease of use in a variety of applied settings especially given they can be administered individually or in group (Raven et al., 1998a).

Although the Raven's Matrices tests require minimal language for testing which can reduce the impact of cultural factors to a large extent (Neisser, 1998), an examinee's response to specific test items or overall test performance can still be influenced by cultural factors (Greenfield, 1998). For example, in some cultures or countries, individuals may respond to the Raven's Matrices tests by following aesthetic principles rather than conceptual patterns based on abstract reasoning as originally intended (Rosselli & Ardila, 2003). Moreover, in countries where formal early education is available, children usually can gain an early understanding of visual organization of rows and columns, as well as the concept of geometric shapes (Greenfield, 1998), which may give the test takers an advantage of understanding the visual patterns better (Lezak, 2012).

Use of the Raven's Matrices tests in Pakistan. Given the low literacy rate in Pakistan (Hussain & Salfi, 2011) and a large number of individuals having limited access to formal education (Malik et al., 2015), the Raven's Matrices tests, as a group of nonverbal intelligence measures, have been frequently used in Pakistan to assess cognitive ability (Amjad & MacLeod, 2014; Behrman, Khan, Ross, & Sabot, 1997; Grantham-McGregor et al., 2007). In addition, the Raven's Matrices tests have particularly been favored by researchers in Pakistan intended to conduct cross-cultural comparisons in large-scale studies (e.g., Murray-Kolb et al., 2014).

Several studies over the last three decades have provided support for use of the Raven's SPM and CPM among Pakistani populations. Shamama-tus-Sabah, Gilani, and Iftikhar (2012) administered the Raven's SPM to 203 elementary school students between ages 8–11 in order to examine the psychometric evidence for using the measure with Pakistani children. The split-half Spearman-Brown coefficient (corrected) was

found to be .80, and the test-retest reliability for a 12-month interval was found to be .77 (Shamama-tus-Sabah et al., 2012). In the Shamama-tus-Sabah et al. (2012) study, girls scored significantly higher than boys did; however, no differences were found among different social classes. Moreover, Ansari (1984) studied the reliability of the Raven's SPM in 432 school aged children in Pakistan, revealing that the overall internal consistency reliability coefficient was high (.95) for both boys and girls. Lastly, the earliest reliability study conducted in Pakistan revealed that the split-half reliability of the Raven's SPM was .72 (Zaki & Beg, 1969). In terms of validity evidence, a research study conducted by Ansari and Iftikhar (1988) showed that the Raven's SPM was a useful measure for assessing *intellectual ability* of Pakistan children in urban areas. In addition, the scores of the Raven's SPM and student achievement were significantly correlated (r = .31) based on a sample of 147 students in Grade 8 (Riaz, 1979).

With regard to the Raven's CPM, Malik, Rehman, and Hanif (2012) included the measure to study how academic interventions would impact developmental skills in a group of students described as slow learners in an urban area in Pakistan. Specifically, the Raven's CPM was used to screen for slow learners (i.e., those who scored between 10th and 25th percentile), indirectly providing some validity evidence for using the CPM as an effective screening tool that can differentiate students with different levels of cognitive ability in Pakistan. Alderman, Behrman, Khan, Ross, and Sabot (1996) used the Raven's CPM to study the regional gap in cognitive skills in rural areas in Pakistan. No significant differences were found in Raven's CPM scores using different regional subsamples; however, older students performed significantly better than younger students (Alderman et al., 1996). A research team incorporated the Raven's CPM for a study

focusing on the relationship between the implementation of a school feeding program for improving nutrition and developmental skills in rural areas in Pakistan (Soofi et al., 2013). A group of children ages 5–12 participated in the study (115 children in the feeding group vs. 39 in the non-feeding group). Children who received the feeding program performed significantly better on the Raven's CPM than those who did not.

Challenges. Although the Raven's Matrices tests have been used in research with Pakistani populations, the amount of time required to complete the different versions of the full-length tests can limit their utility. From a researcher's perspective, the challenges associated with the full-length tests (i.e., the 60-item SPM or 36-item APM) have particularly been evident when used in large-scale research studies (Bilker et al., 2012; Sefcek et al., 2016; Wytek et al., 1984). Such research studies not only deal with a large number of measures of independent and moderating variables, but also involve longitudinal follow-up and constant data monitoring (Bilker et al., 2012). Compared to the Raven's SPM or APM, the full-length RCM can be even more challenging to administer given that it is the longest version of the Raven's Matrices tests (72 items) and requires approximately 60–80 minutes for a single full administration (Raven et al., 1998b). Large-scale research studies are currently being conducted in Pakistan, and the full-length RCM can be particularly problematic when applied in such studies in Pakistan. For example, the Malnutrition and Enteric Disease Study (MAL-ED), a large-scale research network, is being conducted in Pakistan to examine the linkages between malnutrition and intestinal infections and their effects on human development (Murray-Kolb et al., 2014). In order to decrease the financial and time resources required to carry out such research studies in the future, measures included in these studies should be

maximally efficient. This is particularly important in Pakistan, as a lower-middle-income country (World Bank, 2017), where resources are often limited.

Moreover, the full-length Raven's Matrices tests can potentially create an excessive amount of response burden on test-takers at the individual level. This is a critical consideration given that individuals in lower income countries may be unfamiliar with formal testing procedures (Pendergast et al., 2018) and thus may be reluctant to be tested especially when given lengthy measures such as the full-length Raven's Matrices. This especially holds true for certain subpopulations in Pakistan. As mentioned previously in the Introduction chapter, a large number of individuals in Pakistan have limited access to formal education (Malik et al., 2015). The full-length RCM can be perceived as daunting for individuals who did not receive formal education (Murray-Kolb et al., 2014). Specifically, researchers in Pakistan reported that certain participants declined the administration of the full-length RCM in research data collection, and those individuals used lack of schooling as the reason (Murray-Kolb et al., 2014). In addition, due to issues such as poor education quality, poverty, and poor health, a significant number of individuals failed to reach their full cognitive potential in low- and middle-income countries including Pakistan (Grantham-McGregor et al., 2007). It can be more difficult for individuals with lower-than-average cognitive ability to complete the full-length RCM, the longest version of the Raven's, because it is not only time consuming but also can easily cause emotional distress and physical fatigue. Therefore, developing a short form of the RCM is necessary in order to decrease resources required for large-scale research studies and reduce response burden on test takers in Pakistan.

Although several research teams have developed short forms of the Raven's Matrices tests to date, these measures were developed for use in other countries and cultural contexts. In addition, selection of items for many of these short forms was simply based upon practitioner or researcher judgments, with only 3 even drawing upon classical item selection methods. As such, the purpose of the current study was to develop a short form for the full-length RCM for use in Pakistan using modern psychometric methods.

Existing Raven's Short Forms

Clinically derived short forms. As reported in Table 1, in order to address the limitations of the Raven's full-length scales, several teams (Bouma, Mulder, & Lindeboom, 1996; Caffarra et al., 2003; Chiesi, Ciancaleoni, Galli, & Primi, 2012; Elst van der et al., 2013; Smits, Smit, van den Heuvel, & Jonker, 1997) have developed short forms for different versions of the Raven's Matrices tests by using a single or selected subtest(s) based upon practitioner or researcher judgments. Specifically, Chiesi et al. (2012) argued that the first set of the APM consisting of the easiest items would likely work best in a clinical setting. Caffarra et al. (2003) selected the first four sets of the SPM based on the observation that patients with impaired cognitive functioning rarely went beyond Set D to complete Set E in clinical practice. Bouma et al. (1996) decided to omit the easiest and hardest items on Set A and Set E respectively, and administered Sets B, C, and D of the Raven's SPM (36 items) to patients with impaired cognitive functioning. Similarly, Smits et al. (1997) developed a short form of the Raven's CPM by eliminating the whole Set AB and utilizing only items on Set A and Set B (24 items).

The four short forms across these studies were developed for use with clinical (e.g., psychiatric patients with cognitive deficits) or elderly populations (e.g., ages 55–85), and the majority of studies focused on developing norms for the selected subtest(s) with the one exception being a validation study conducted by Chiesi et al. (2012). In addition, the study samples mostly included older adults with only Chiesi et al. (2012) focusing on school-aged children. Although the development of norms for these short forms may have provided clinicians with a useful tool for screening and evaluating treatment outcomes, many of the studies only examined limited types of reliability and validity evidence to justify such use in practice. As noted in Table 1, only Chiesi et al. (2012) reported some reliability and validity evidence relative to the use of Set 1. As the Chiesi et al. (2012) short form contained only Set 1, it includes only the easiest items on the APM and may lack discrimination power due to its limited range of item difficulty. With regard to the length of the short forms, three (Bouma et al., 1996; Caffarra et al., 2003; Elst van der et al., 2013; Smits et al., 1997) still included a fairly large number of items (24–48).

Although clinicians may provide a valuable perspective on which set(s) of items of the Raven's Matrices tests appear to serve well as short forms in practice, the selection of a single or multiple subtest(s) of items based on prior clinical experience has several limitations. First, the method is arbitrary with the selection of items being exclusively informed by clinician's prior clients and experiences. Second, this approach may result in a narrow range of item difficulty level because the items on the Raven's Matrices tests are ordered by increasing difficulty within each set. Similarly, the range of item discrimination also may be limited in clinically derived forms because test items will

have low discrimination if it is too difficult or too easy. In sum, this approach can result in the loss of items that may have better psychometric properties and thus may not be an ideal method. As such, the following section focuses on short forms of the Raven's Matrices tests derived using empirical evidence.

Table 1

Overview of Raven's Clinically Derived Short Forms

Authors	Intended use(s)	Version	Subtests (item total)	Sample	Reliability evidence	Validity evidence (Criterion measure)
Chiesi et al. (2012)	Used for research and clinical practice	APM	Set 1 (12)	N = 1,389; school-aged students (M_{age} = 11.25, SD = 1.82); Italy	Sufficient TIF (i.e., maximum value = 6.6 at Ability Level –0.5; between –1 and +1, the amount of test information was greater than 4.0)	Single-factor structure; r = .46* (Digit Span) r = .35*(Probabilistic Reasoning, PR; primary school); r = .34*(PR; secondary school)
Caffarra et al. (2003)	Early detection of individuals at risk of developing dementia; norms for Italian population	SPM	Sets A–D (48)	N = 248; healthy community members (M_{age} = 52.10, SD = 19.56); Italy	None reported	None reported
Smits et al. (1997)	Norms for the elderly	CPM	Sets A–B (24)	N = 2,815; aging population (adults aged 55–85); Netherlands	None reported	None reported
Bouma et al. (1996) and Elst van der et al. (2013)	Norms for general clinical use	SPM	Sets B–D (36)	N = 453; cognitively healthy adults (24–83 yrs.); Netherlands	2PL information > 10, corresponding with a level of reliability > .90 in the ability range between –2.1 and 0.2	None reported

Note. All values in the "Validity evidence (Criterion measure)" column are Pearson correlations.

Short forms informed by data. As summarized in Tables 2 and 3, five research teams have utilized data-informed item-reduction strategies to develop short forms for the Raven's SPM (Bilker et al., 2012; Wytek et al., 1984) and APM (Arthur & Day, 1994; Bors & Stokes, 1998; Sefcek et al., 2016). The data-informed item-reduction strategies can be further categorized into classical methods relying on CTT-based statistics and modern methods developed in recent decades.

Classical methods. Arthur and Day (1994) developed a 12-item short form for use in research and practice. The 12 items were selected from the APM by dividing the 36-item full-length test into 12 groups of three items each based on the original order of items (Items 1–3 were the first grouping, Items 4–6 were the second, etc.) and then choosing the item within each group that demonstrated the highest item-total correlation. The Arthur and Day (1994) short form required approximately 15 minutes to complete. As shown in Table 2, Arthur and Day administered the full-length APM to a sample of 202 university students from the United States (M_{age} = 21.40, SD = 4.42) to inform item selection for the short form. The Arthur and Day (1994) short form generally maintained the progressive difficulty and single factor structure of the full-length test; however, the internal consistency of the short form (α = .65) was considerably lower than on the full-length test (α = . 86). The short form also was moderately correlated (r = .66) with the full-length test.

The Arthur and Day (1994) 12-item short form has significantly reduced the administration time from 45 (to complete the full-length scale) to 15 minutes for university students. It also has demonstrated some promising, though limited, reliability

and validity evidence for use with university students, and it supposedly can be used with university students in either research or clinical settings in the United States. Nevertheless, a relatively small sample was used to standardize and evaluate the psychometric properties of the short form. Most importantly, the item reduction strategies employed by Arthur and Day (1994) essentially relied on the traditional CTT approach, a relatively less sophisticated psychometric method that is sample-dependent and item-dependent.

Bors and Stokes (1998) raised concerns that Arthur and Day's short form might include redundant items because inter-item correlations were not examined, thus limiting the predictive power of the short form. In addition, they argued that certain retained items were too easy. As such, Bors and Stokes (1998) developed a separate 12-item short form for the Raven's APM (40-minute timed version) using CTT-based methods. Specifically, they rank ordered all items by their item-total correlations and eliminated items with high inter-item correlations. As displayed in Table 2, the Bors and Stokes (1998) short form demonstrated higher internal consistency (i.e., Cronbach's $\alpha = .73$) compared to Arthur and Day (1994)'s short form, and the test-retest reliability was within the acceptable range (i.e., $r = .82$). The total scores of Bors and Stokes (1998)'s short form were moderately correlated with the those of the full-length APM and Arthur and Day (1994)'s short form, as well as other estimates of intelligence (Table 2). The Bors and Stokes (1998) short form has positive reliability and validity evidence to support its use with university students (Table 2). Similar with Arthur and Day (1994)'s research methodology, Bors and Stokes (1998) also used traditional CTT approach.

Aiming to increase the internal consistency of the 12-item short form developed by Arthur and Day (1994), Sefcek et al. (2016) added six items with increasing item difficulty to create a "*medium form*" (p.1) of the Raven's APM. As reported in Table 2, the short form contained 18 items in total and reportedly requires 17–25 minutes to complete. Sefcek et al. (2016) recruited 633 university students in the US (M_{age} = 20.92 years, *SD* = 4.07) for their initial study of the 18-item form. A follow-up validation study revealed that the Sefcek et al. (2016) short form demonstrated higher internal consistency reliability (e.g., Cronbach's α = .79) as compared to that of the Arthur and Day (1994) short form (e.g., Cronbach's α = .73). It also maintained the progressive difficulty of the full-length APM (Sefcek et al., 2016). In addition, the Sefcek et al. (2016) form demonstrated some promising convergent validity as evidenced by the statistically significant correlations between the 18-item form, the Shipley Abstraction scale, and self-reported SAT scores.

The Sefcek et al. (2016) form takes more time for completion (i.e., 17–25 minutes to complete) compared to the 12-item short forms (approximately 15 minutes) developed by Arthur and Day (1994) and Bors and Stokes (1998) when used with university students. Nonetheless, the 50% increase in items from the Arthur and Day (1994) 12-item short form only resulted in a relatively small increase in internal consistency ($\mathbf{\Delta}$ = .05) which may not justify the additional administration time. Although developed primarily for research purposes, initial evidence of psychometric properties provides support for its use with university students. No published evidence is currently available to support its use with other populations, however. Similarly, Sefcek et al. (2016) also

used the CTT approach for selecting items for the short form which, as noted previously, has limitations due to its sample-dependent and item-dependent properties.

Table 2

Raven's Short Form Studies Using Classical Methods

Authors	Intended use(s)	Version	Approach	Final # of items	Sample/target population	Reliability evidence	Validity evidence (Criterion measure)
Arthur and Day (1994)	Research and screening	APM	CTT	12	N = 202; university students; USA	α = .65 α = .86 (full-length) Test-retest = .75	r = .66 (full-length) Unidimensional
Bors and Stokes (1998)	Research	APM (timed)	CTT	12	N = 506; university students; Canada	α = .73 α = .84 (full-length) Test-retest = .82	r = .92 (full-length) r = .88 (Arthur & Day's Short Form) r = .61(Shipley-Abstraction) r = –.42 (Inspection Time Measures)
Sefcek et al. (2016)	Research	APM	CTT	18	N = 633; university students; USA	α = .79 (18-item) α = .74 (embedded 12-item; Arthur & Day, 1994)	r = .22* (Mill-Hill Vocabulary Scale) r = .49* (Shipley-Abstraction) r = .12 (Shipley-Vocabulary) r = .17* (GPA) r = .34* (SAT) r = .35* (Verbal Creativity) r = .29* (Drawing Creativity) r = .44* (ACT) r = .26* (Openness) r = –.16* (Conscientiousness) r = –.03 (Extraversion) r = –.11 (Agreeableness) r = .02 (Neuroticism)

Note. All values in the "Validity evidence (Criterion measure)" column are Pearson correlations.

Modern methods. Compared to the traditional methods discussed, two other research teams (Table 3) have used alternative methods that have been developed in recent decades to create short forms for the Raven's SPM. In order to develop a time-efficient instrument for screening and monitoring treatment outcomes in psychiatric patients, Wytek et al. (1984) developed a 30-item SPM short form based on iterative selection from Rasch model statistics. A sample of 300 psychiatric patients with impaired cognitive functioning from Vienna, Austria comprised the sample for the study. The split-half reliability coefficient of the Wytek et al. (1984) short form was $r = .95$ (Wytek et al., 1984).

The 30-item Wytek et al. (1984) short form is roughly 1.5–2.5 times as long as the short forms described previously (i.e., 12 or 18 items). No estimated completion time was reported; however, based on the reported completion time for the 12- and 18-item short forms, the Wytek et al. (1984) short form would likely require approximately 30 minutes to complete. In addition, the only form of reliability evidence reported was split-half reliability, and no validity evidence was reported. In term of the methodology, Wytek et al. (1984) used iterative selection based on statistics from the Rasch model. However, no analyses appear to have been conducted to determine whether the Rasch model was the best-fitting model for item parameter estimates. In addition, the Rasch model does not take into account guessing and assumes that all items have equivalent discrimination. As a result, the assumptions of the Rasch model are too restrictive compared to IRT models that also integrate guessing and item discrimination. Although the simple Rasch model has some limitations, Wytek et al. (1984) was the first research team that incorporated IRT statistics for constructing short forms for the Raven's SPM,

and the results from their Rasch analysis were generalizable for individuals with psychiatric disorders accompanying cognitive deficits.

Using an alternative approach, Bilker et al. (2012) constructed two 9-item short forms based on the SPM for use in large-scale longitudinal or treatment studies involving psychiatric patients with cognitive deficits. Bilker et al. (2012) applied the Poisson predictive model to identify a set of items that were highly correlated with total scores on the 60-item full-length SPM. As shown in Table 3, a U.S. sample of 180 participants (M_{age} = 33.9, *SD* = 12.6) including healthy volunteers and patients with schizophrenia participated in this study. A split sample approach was used for the short form construction and scale validation, meaning that 90 participants were included to inform item selection for the short form and the remaining 90 participants were analyzed for validation. The two Bilker et al. (2012) short forms reportedly require an average of 3 and 4 minutes to complete, saving approximately 76% and 82% of administration time, respectively. Bilker et al. (2012) used random data splitting for model fitting and an additional validation study, along with cross-validation for results verification. Using the validation dataset and the 9 items identified on each form as predictors, correlations between the total score of the SPM and those on the short forms ranged from .90–.91, indicating good prediction accuracy (Bilker et al., 2012). As expected, the Cronbach's α of the short forms' scores dropped to .80 (Form A) and .83 (Form B) compared to .96 for the full-length test.

Relative to the short forms reviewed previously, the Bilker et al. (2012) short forms produced the most significant administration time savings (reducing the time required from 17 min to 3–4 min). The two substantially shortened forms appear to be

encouraging for researchers who need such a brief assessment in research focusing on individuals with schizophrenia and healthy adults. As reported in Table 3, reliability and validity evidence for Short Forms A and B also appear promising. Methodologically, although the Poisson predictive model does not rely on common CTT or IRT indices, it does offer relative advantages, which are allowing researchers to move beyond item-level characteristics provided by CTT/IRT as well as to identify the best combination of items that are highly predictive of the test scores from the full-length measure as an alternative method of short form construction. However, this approach required significant computational resources and could easily share the same floor and ceiling effects inherent in the original test especially given its heavy reliance on correlations during item selection (Bilker et al., 2012).

Table 3

Raven's Short Form Studies Using Modern Methods

Authors	Intended use(s)	Version	Approach	Final # of items	Sample/target population	Reliability evidence	Validity evidence (Criterion measure)
Wytek et al. (1984)	Clinical	SPM	Iterative selection using Rasch statistics	30	N = 300; psychiatric patients with cognitive deficits; Vienna, Austria;	Split-half: r = .96;	None reported
Bilker et al. (2012)	Clinical	SPM	Poisson predictive model	9	N = 90; healthy adults & psychiatric patients; USA	α = .80 (A) α = .83 (B) α = .96 (full-length)	Validation sample Predicted w/ actual total scores r_A = .91 r_B = .90 Short form w/ full-length test r_A = .98 r_B = .98 Content representation of Forms A and B in abstract reasoning (e.g., gestalt completion) was similar to that of the full-length SPM.

Note. All values in the "Validity evidence (Criterion measure)" column are Pearson correlations.

Conclusions regarding existing short forms. Taken together, the short form development studies up to the present time have primarily based on the Raven's SPM or APM. Although combining the Raven's SPM and CPM can provide a broader range of difficulty levels and allow for comparisons among individuals across a wide range of ages, no prior efforts have attempted to construct short forms for both (i.e., the full-length RCM). In addition, all the existing Raven's short forms have been developed for use with only Western populations (e.g., the United States, Canada, and Austria). As test items may function differently in different cultural contexts under the influences of various educational, cultural, and environmental factors (American Educational Research Association, AERA; American Psychological Association, APA; National Council on Measurement in Education, NCME; 2014), the use of the existing short forms can only be limited to the respective countries.

Methodologically, two broad approaches have been applied in Raven's short form development efforts to date. The first approach simply utilizes a single or several subsets of certain versions of the Raven's Matrices based on practitioners' or researchers' prior experience (Bouma et al., 1996; Caffarra et al., 2003; Chiesi et al., 2012; Elst van der et al., 2013; Smits et al., 1997). Although this method provides a clinical perspective, the item selection process can be somewhat arbitrary. The second approach relies on data-informed item-reduction strategies. The majority of these forms were developed using classical test methods primarily based on item difficulty and item discrimination (Arthur & Day, 1994; Bors & Stokes, 1998; Sefcek et al., 2016). However, the classical approach is limited by its sample dependent item statistics (i.e., item difficulty and item discrimination). In contrast, two studies (Bilker et al., 2012; Wytek et al., 1984) featured

modern item-reduction strategies to develop short forms. As part of the earliest study included in the literature review, Wytek et al. (1984) employed iterative selection based on Rasch model statistics, which was an IRT informed method. However, no analyses were conducted to confirm whether such a model was the best-fitting model for item parameter estimates. Lastly, Bilker et al. (2012) applied a sophisticated Poisson predictive model to construct short forms. Bilker et al. (2012) pointed out two limitations of this approach: the large computational resources required and the influence of floor and ceiling effects inherent in the full-length test due to the sole reliance on correlations during item selection. As demonstrated by this review, IRT techniques have rarely been used to develop short forms for the Raven's Matrices tests in the past in despite of its advantages for short form development (e.g., sample-independent, test-independent).

Rationale and Purpose of the Current Study

The Raven's Matrices tests have been used in research with Pakistani populations; however, the amount of time required to complete the full-length RCM limits its utility. Specifically, potential challenges in the applications of the full-length RCM in large-scale research studies exist, including physical fatigue and/or mental distress – especially among individuals with less formal education or lower cognitive ability. However, no attempts have been made to date to develop a RCM short form for use in Pakistan. Although the Raven's Matrices tests are considered to be culturally reduced measures (Neisser, 1998; Raven et al., 1998a, 1998b; Raven & Court, 1989; Raven & Raven, 2003), cultural and geographical variations have been observed in scores from the tests (Raven, 2000; Rosselli & Ardila, 2003). As such, when the Raven's Matrices tests are used in Pakistan where the culture is different from Western countries and formal education is

not widely accessible, culture may impact how the test items are perceived as well as overall test performance. In accord with the guidelines for test use and test adaptation (AERA et al., 2014; International Test Commission, 2013, 2017), when a test is applied in groups with different cultural backgrounds or geographical regions, the psychometric properties of the test must be evaluated in order to ensure the appropriate test use and valid result interpretation. As such, the existing short forms discussed previously cannot be directly applied in Pakistan without proper validation using Pakistani samples.

Therefore, the purpose of the current study was to develop a short form from the full-length RCM for use in Pakistan. IRT was utilized to inform item selection given its notable advantages. Specifically, the ultimate goal was to create a short form that demonstrates psychometric properties similar to that of the full-length RCM yet can be administered in a much briefer timeframe. Given the target population for the RCM short form was typically developing adolescents and young adults, items providing most information around the middle range of the ability continuum under the IRT framework were selected to construct the short form. A detailed description of the item selection process and criteria is provided in the next section.

Chapter 3

METHOD

Participants

Data were collected with a sample of adolescents and young adults in Pakistan who had participated in an earlier longitudinal study conducted between 1989 and 1996 in a village in Gilgit-Baltistan, Pakistan. Covering approximately 72,496 square kilometers, Gilgit-Baltistan constitutes roughly 9% of Pakistan's territory and has a population over 1.4 million people (Khan, 2017). Initially, 1,857 children under the age of 5 participated in the surveillance study. Over time, some participants were lost due to death (n = 135; 7.3%), while others left the study for unspecified reasons (n = 257; 13.8%). In the follow-up study (2012–2014), a majority of the original participants (n = 1,465) were located and interviewed by trained professionals using standard follow-up questionnaires. During these interviews, 60 participants (3%) declined to complete testing at follow-up. Therefore, the final sample for the current study included 1,405 adolescents and young adults from the surveillance follow-up study. Approximately 49% (n = 686) of these participants were female, and 51% were male (n = 715). Figure 1 shows the flow of participants through phases of the initial and follow-up studies.

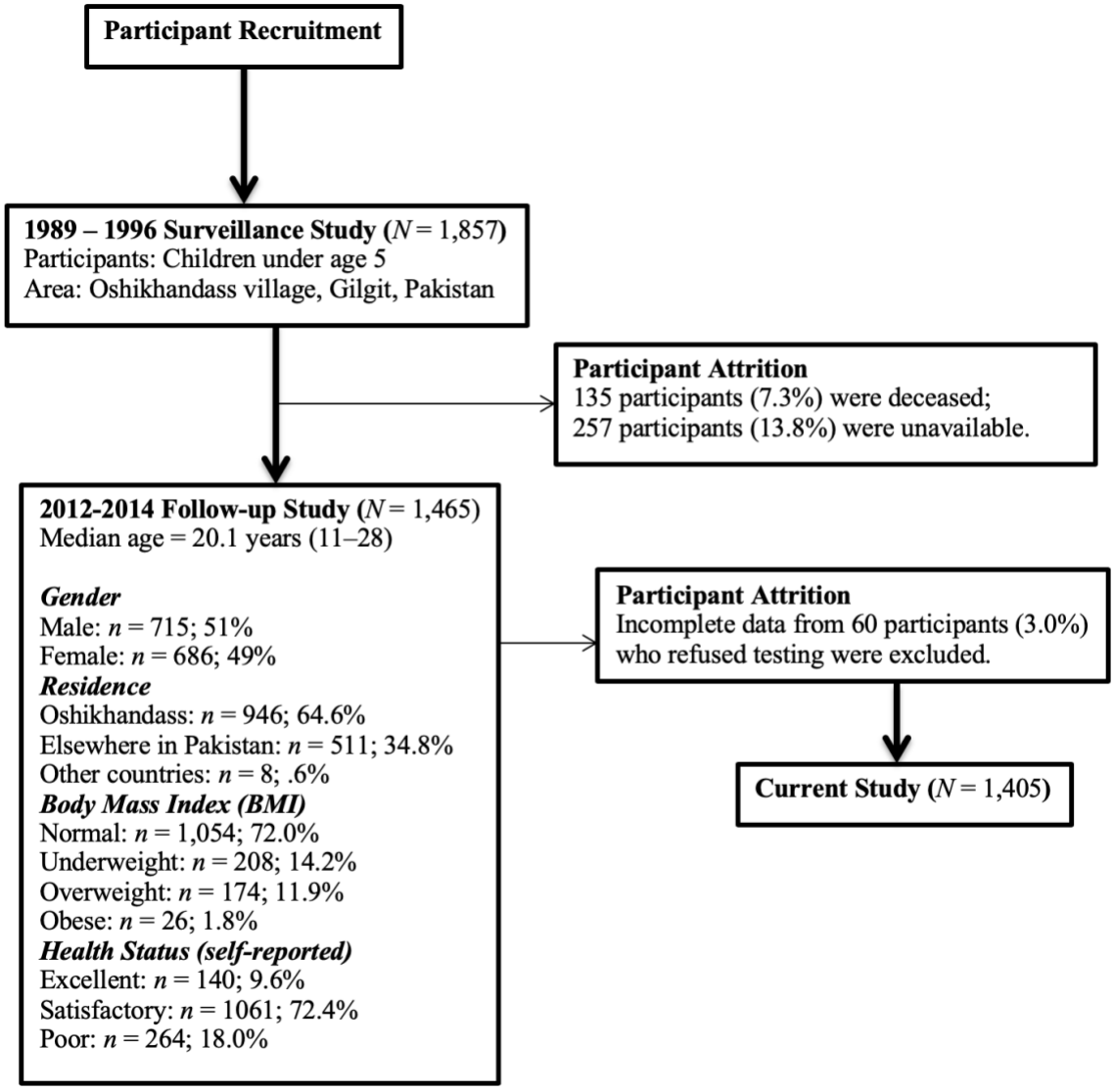

Figure 1. Flow of participants through each stage of data collection.

Measure

As explained in the literature review, the Raven's Progressive Matrices were designed to measure the eductive ability of Spearman's *g*. The Raven's SPM was developed for the general adult population; whereas the CPM includes items with lower levels of difficulty that were developed for use with younger and older individuals. The combined Raven's CPM and SPM Matrices (i.e., RCM) include a broad range of item

difficulty levels, which allows for their use across a wide range of ages and abilities. Specifically, Sets A, AB, and B of the Raven's CPM and Sets C, D, and E of the Raven's SPM were administered as part of the follow-up study. Each Raven's item is comprised of a matrix with a combination of visual geometric elements and uses a multiple-choice format where take takers are supposed to identify the missing element that is consistent with each matrix pattern.

Procedures

The RCM was administered by trained professionals, and all participants began with Set C (i.e., the starting point) of the Raven's SPM. Basal and ceiling rules were utilized throughout the test administration in order to reduce the amount of time required for data collection and to decrease response burden on participants. Specifically, if a participant had less than seven items correct on Section C, they were administered the preceding sections until they had at least seven correct items within any given section. Conversely, if participants provided at least seven correct responses on a given section, they were administered the next section until they failed to meet this criterion. Each item on the RCM was dichotomously scored (correct = 1, incorrect = 0).

Data Analyses

Missing data. As a result of the basal and ceiling rules applied during data collection, actual completion rates for each set of full-length RCM items varied. Consequently, there was a large percentage of missing data – particularly within Sets A and AB of the original dataset (Table 4). Specifically, 467 participants reversed to complete Set B (i.e., completed Sets B and C), and 1265 participants advanced to complete the last section (i.e., completed Sets C, D, and E). Because the Raven's items

are ordered according to progressively increasing difficulty levels, only a small percentage of participants with extremely low cognitive skills would have been administered the items in these two sets (Table 4). Given the goal of this study was to develop a short form for use with the general adolescent and young adult population in Pakistan, only data from Sets B, C, D, and E were used to identify a subset of items for the short form. As used in this study, the term "RCM long form" refers to Sets B, C, D, and E, which serves to differentiate it from the full-length RCM that includes all sets of items.

Table 4

Number and Percent of Participants Completing each RCM Item Set

	Item set					
	A	AB	B	C	D	E
Completion rate (%)	2.28	4.48	33.38	100.00	99.96	90.00
Completed cases (*n*)	32	63	467	1,405	1,400	1,265

Item selection and initial validation. In order to examine if results of the item selection process were stable, the current dataset (N =1,405) was randomized into two subsamples (training and validation) using SPSS. Specifically, the training sample (n = 703) was used for parameter estimation and initial item selection. The resulting short forms then were tested further with the validation sample (n = 702).

Training sample analyses. Several analyses were performed on the training sample, including model comparison, testing of assumptions, as well as parameter estimation and item selection. Specifically, IRT was used to obtain parameters for

evaluating individual items given its notable advantages compared to CTT. In order to estimate the item parameters accurately, several IRT models were compared to determine the best fitting model (model comparison). Two essential assumptions, unidimensionality and local independence, were then examined. Based on the item parameters from the best fitting IRT model, item information was plotted for evaluating the amount of information each item contributed during the item selection process.

IRT model comparison. Given that RCM items are dichotomously scored, dichotomous IRT models were used to analyze the item level data. In order to obtain accurate item parameters, three IRT models (one-parameter model, 1PL; two-parameter model, 2PL; and three-parameter model with beta prior for guessing, 3PL) were compared in order to determine the best fitting model. A range of goodness-of-fit statistics were generated. The 1PL, 2PL, and 3PL models were nested (i.e., the 1PL model was nested in the 2PL model, and the 2PL model was nested in the 3PL model), thus allowing the comparison using the –2 log likelihood. The differences in the –2 log likelihood values between each pair of the nested models and the number of degrees of freedom were calculated. If a *p*-value based on the chi-square test is smaller than .05, the more complex model is regarded as fitting the data better (Embretson & Reise, 2000; Tate, 2003). Penalized-likelihood fit statistics including the Akaike Information Criterion (Akaike, 1974) and the Bayesian Information Criterion (Schwarz, 1978) also were examined. The model with the smallest AIC is generally considered the preferred model (Akaike, 1974); similarly, the model with the smallest BIC value represents the best model fit (Schwarz, 1978). In addition, Maydeu-Olivares and Joe (2006) proposed using M_2, a fit statistic for dichotomous data, to analyze the goodness of fit between two

models. M_2 values along with the degrees of freedom and p values were reported. Based on the M_2 fit statistics, a low RMSEA value smaller than .06 is considered better (Hu & Bentler, 1999). The goodness-of-fit statistics were evaluated simultaneously.

Unidimensionality. The assumption of essential unidimensionality was first assessed by comparing different factor models using NOHARM (Fraser & McDonald, 1988) with guessing parameters obtained from flexMIRT (Cai, 2017). First, the factor loadings were obtained in order to examine if there was a dominant factor (i.e., the majority of the factor loadings are greater than .50; Nunnally & Bernstein, 1994). Results of previous research studies have indicated that, when there is a dominant factor, the impact on the IRT parameter estimation is negligible even with minor dimensions outside the dominant factor (Anderson, Kahn, & Tindal, 2017; Cuesta & Muñiz Fernández, 1999; Harrison, 1986). Second, the Tanaka index of goodness of fit was computed to evaluate the factor models; values greater than .95 indicate a satisfactory fit to the data (Tanaka, 1993). Third, the Root Mean Square Residual (RMSR) values were compared with the values obtained from 4 divided by the square root of the sample size, that is $4/\sqrt{N}$. Specifically, if a RMSR value is smaller or equal to $4/\sqrt{N}$, that indicates a good model fit (Fraser & McDonald, 1988). Lastly, the percent of reduction in RMSR with the addition of a new factor was subsequently calculated; a decrease in RMSR less than 10% indicates that the variance of the additional factor does not substantially contribute to the explanation of the latent trait, supporting a good fit for the original model (Tate, 2003). The goodness-of-fit statistics were evaluated simultaneously.

Local independence. After identifying the best-fitting IRT and factor models, the local independence assumption was further examined to ensure that pair-wise item

responses are not correlated given an ability level. Specifically, the chi-square test of local independence was performed, and standardized χ^2 statistics for each pair of items (Chen & Thissen, 1997) were calculated. Excessively large values (i.e., $\chi^2 > 10$) indicate a violation of the assumption (Chen & Thissen, 1997).

IRT item selection. Item selection using IRT relies on the amount of estimated item/test information, which is analogous to reliability in CTT (Hambleton & Swaminathan, 1985). Concurrent calibration was used to simultaneously estimate item parameters using all data when a common IRT scale was obtained. The completed Sets B, C, D, and E were treated as a single test, and all the participants were considered as a single group. Based on the best fitting IRT model selected from model comparison, the item parameter estimates for the 48 items on Sets B, C, D, and E were generated using the training sample. IRT parameters usually include item difficulty (*b* parameter), item discrimination (*a* parameter), and/or guessing (*g* parameter). Under IRT, the difficulty of an item is a location index that describes where the item functions along the ability scale. Item difficulty can only be properly interpreted on the ability scale, meaning that an item can be easy for one individual but difficult for another. Item discrimination describes how well an item can differentiate between test takers who score high on the test (high-ability test takers) and those who obtain low scores (low-ability test takers).

According to Baker (2001)'s guidelines for interpreting item parameters, item discrimination is classified as "none" (0), "very low" (0.01–0.34), "low" (0.35–0.64), "moderate" (0.65–1.34), "high" (1.35–1.69), "very high" (> 1.69), and "perfect" (infinity). According to IRT, a test taker with higher ability has a higher likelihood of answering a test item correctly. Guessing refers to the probability of getting the item correct by

guessing alone. Theoretically, guessing ranges from 0 to 1 with values smaller than .35 being considered acceptable (Baker, 2001).

The Item Information Function (IIF) was computed based on the item parameters derived from IRT (Hambleton & Swaminathan, 1985). In the IRT item selection process, item information curves were evaluated (Hambleton & Jones, 1993). Given the goal of the current study was to create a short form for use with adolescents and young adults in the general population in Pakistan, items that provided the maximum amount of item information around the middle of the ability range were included. In order to reduce item redundancy, items with excessively large local dependence (LD) χ^2 values (i.e., > 10) were eliminated during the item selection process. Lastly, the TIF of the short form should resemble the RCM long form closely because that increases the likelihood of the short form being psychometrically similar to the RCM.

Initial reliability analysis. Cronbach's alpha coefficients (Cronbach, 1951) were calculated for the RCM long and short forms in order to examine the impact of item reduction on internal consistency. Marginal reliability coefficients for the short forms (based on independent IRT analyses by rerunning the IRT analyses with selected items) and the RCM long form (Green, Bock, Humphreys, Linn, & Reckase, 1984) were additionally computed for comparison. Marginal reliability essentially is an IRT-derived estimate of test score reliability, and individual error variances are averaged (Green et al., 1984; Lord & Novick, 1968). When evaluating reliability evidence for the short forms, coefficients of .70 or higher were considered acceptable. Values ranging from .60 to .70 were considered minimally acceptable, and values below .60 were considered unacceptable (DeVellis, 1991; George & Mallery, 2003; Nunnally, 1978). In addition,

Nunnally (1967) recommended .50 to .60 for early stages of validation research, and George and Mallery (2003) suggested a Cronbach's alpha coefficient of .70 or greater for the purpose of establishing the reliability of a measure for research purposes.

Validation sample analyses. To evaluate whether the results obtained from the item selection analyses with the training sample were stable and would generalize to an independent sample, several analyses were conducted with the validation sample. First, the correlations between the theta scores from the 48 items and those from the selected items were generated. Hinkle, Wiersma, and Jurs (2003) suggested that correlation coefficients between .30–.70 indicate a moderate linear relationship. Moreover, values smaller than .30 indicate a weak linear relationship while values greater than .70 suggest a strong correlation. Ideally, the correlation between the short form and the RCM long form would exceed .90; however, a correlation in the .80 to .90 range is acceptable.

Second, internal consistency coefficients were estimated using the validation sample. Specifically, Cronbach's alpha coefficients (Cronbach, 1951) were first calculated for comparison with those generated from the training sample. In addition, marginal reliability coefficients (Green et al., 1984) based on independent IRT analyses (i.e., re-estimating item parameters using the selected items) were then generated. Marginal reliability coefficients (Green et al., 1984) using cross-validation (i.e., scoring the validation sample using item parameters obtained from the training sample) also were computed. All of these internal consistency coefficients (i.e., Cronbach's alpha, marginal reliability based on independent IRT analyses, and marginal reliability based on cross-validation) also were calculated for the RCM long form for comparison purposes. Third,

the items with large LD χ^2 statistics (i.e., > 10; Chen & Thissen, 1997) identified during the item selection were further examined using the validation sample.

Analytic software. FlexMIRT (Cai, 2017) was utilized for IRT item parameter estimation, marginal reliability coefficients, and exploratory factor analysis. SPSS (IBM, 2011) was used to split the total sample and to calculate the Cronbach's alpha coefficients. RStudio (RStudio Team, 2015) was used to plot the Item/Test Information Functions and calculate the correlation coefficients.

Chapter 4

RESULTS

IRT Model Comparison

Model comparison was conducted utilizing the training sample. The goodness of fit statistics for three IRT models (i.e., 1PL, 2PL, and 3PLwith beta prior for all *g* parameters) are reported in Table 5. As the 1PL, 2PL, and 3PL (with beta prior) models were nested, the differences between the log-likelihood values were considered. In comparing the 1PL and 2PL models, the differences between the –2 log likelihood values indicated a preference for the 2PL model, Δ–2LL = 27161.42 – 26478.20 = 683.22, Δ*df* = 1127 – 1080 = 47, $\chi^2(683.22, 47)$, *p* = .000. When comparing the –2 log likelihood values between the 2PL and 3PL models, the result favored the 3PL model, Δ–2LL = 26478.20 – 26354.77 = 123.43, Δ*df* = 1080 – 1032 = 48, χ^2 $(123.43, 48)$, *p* = .000. Overall, the –2 log likelihood values suggested that the 3PL model was the best fitting model to the data.

Table 5

Comparison of Goodness of Fit Statistics for IRT Models

Fit statistics	IPL	2PL	3PL with beta prior
–2LL	27161.42	26478.20	26354.77
AIC	27259.42	26670.20	26642.77
BIC	27482.63	27107.52	27298.74
M_2	9604.26	4687.60	4373.73
df	1127	1080	1032
p	.000	.000	.000
RMSEA	.10	.07	.07
Marginal reliability	.74	.80	.84

Although AIC values also provided support for the 3PL model as the best fitting model, the BIC values indicated a better model fit for the 2PL model. The difference was not surprising given that BIC penalizes model complexity more heavily; therefore, BIC tends to select simpler models, whereas AIC tends to select more complex models. The values of RMSEA suggested adequate model fit to the data using either the 2PL or 3PL models. Lastly, the marginal reliability value for the 3PL was the highest among the three IRT models.

In sum, although results indicated that the 2PL was adequate, collective evidence indicated that the 3PL model provided the best fit. Consequently, the 3PL model was used primarily for the subsequent IRT data analyses; however, when 3PL failed to converge, the simpler alternative model of 2PL was substituted for the analyses.

Testing of Assumptions

Unidimensionality. Testing of assumptions was conducted based on the training sample. In order to assess unidimensionality, different factor solutions were examined based on exploratory factor analysis. Initially, only the one-factor solution converged properly. The two- and three-factor solutions, based on 2PL and 3PL (with beta prior) IRT model, both failed to converge. This source of the convergence issue was likely due to the correlation matrix being non-positive definite as item analysis revealed that Item E12 appeared to have approximately zero variance. Based on the converged one-factor solution with all test items using 3PL (with beta prior), the majority of the items (71%) demonstrated high factor loadings on the one-factor solution (> .50). The Tanaka Index (.99) and the RMSR value (.02659) smaller than 4 divided by $\sqrt{703}$, that is 0.15, suggested a good model fit for the one-factor solution.

Given Item E12 appeared to have approximately zero variance, it was subsequently removed and the same factor analysis was conducted. Based on the 3PL (with beta prior) IRT model, the one-factor solution showed 74% items loaded on this factor (loadings > .50). Tanaka Index (.99) and RMSR value (.02692) smaller than 4 divided by $\sqrt{703}$, that is 0.15, supported the one-factor solution. When an additional factor was added, 74% of the items continued to load on the first factor, providing evidence for the existence of a dominant factor. Further, when a second factor was added, the reduction in RMSRs from .02692 to .02544 was smaller than 10%, providing additional evidence to support that the one-factor solution was sufficient to explain the latent trait. Based on the collective evidence available, the one-factor solution was determined to sufficiently represent the underlying structure. Thus, the assumption of essential unidimensionality was considered met satisfactorily.

Local independence. Approximately 95% of the standardized LD χ^2 values were smaller than 10, indicating the local independence assumption was considered tolerable. The χ^2 values for items with high LD are reported in Tables 13 and 14 in Appendix B, which were further examined and eliminated during the item selection process.

Item Selection

Due to the percentage of missing values, IRT concurrent calibration was used for item parameter estimation, using the log-normal prior parameters for the slope, the normal prior parameters for the intercept, and the beta prior parameters for guessing. Because the same participants completed Section B, C, D, and E, the data analysis was considered single-group single-test, meaning that the participants were not divided into

subgroups and the test was treated as one full test rather than multiple subtests during data analysis. Based on this IRT model, item parameters (*a* = item discrimination; *b* = item difficulty; *g* = guessing) were estimated and subsequently used to plot the item/test information functions for item selection. Item selection based on an IRT approach relies on the amount of estimated item/test information.

Table 6

IRT Item Parameters for Selected Items (Training)

Item	*a*	*b*	*g*
B1	2.1	–2.6	.2
B2	1.6	–2.8	.2
D1	3.7	–1.2	.2
D2	6.5	–0.6	.1
D3	3.3	–0.5	.1
D4	2.9	0.3	.2
D5	3.3	0.2	.3
D6	3.2	0.4	.2
D7	3.3	0.7	.1
D8	2.2	1.3	.2
D9	2.4	1.1	.1
D10	2.0	1.1	.1
D11	1.5	2.3	.1

Note. *a* = item discrimination, *b* = item difficulty, *g* = guessing; IRT item parameters for all items are available in Appendix A.

During the initial item selection, the amount of test information provided by each item was calculated and compared at each ability level. As a result of this initial evaluation of item information curves, the 13 items (B1, B2, D1, D2, D3, D4, D5, D6, D7, D8, D9, D10, & D11) that provided the maximum amount of overall information across ability levels (especially around the middle range) were included for further examination (see Table 6 for item parameters). Discrimination parameters for these items ranged from 1.5–6.5, with most considered highly discriminating items (> 1.7) according to Baker

(2001)'s guidelines for interpreting item parameter values. The item difficulty parameters ranged between –2.8 and 2.3, with easier items from Section B and increasingly difficult items from Section D accordingly. Lastly, the guessing parameters (*g*) for all 13 items were > 0; however, the observed values (.1–.3) suggested a relatively low level of guessing (< .35; Baker, 2001).

Item information and level of local independence were then further examined for the 13 items initially selected. The test information function (TIF) for the RCM long form and the item information functions (IIFs) for the 13 items are displayed in Figure 2, and the specific amount of item information across all ability levels is reported in Table 7. Standardized LD χ^2 statistics for each pair of items mentioned previously in testing of the assumption of local independence are reported in Tables 13 and 14 in Appendix B.

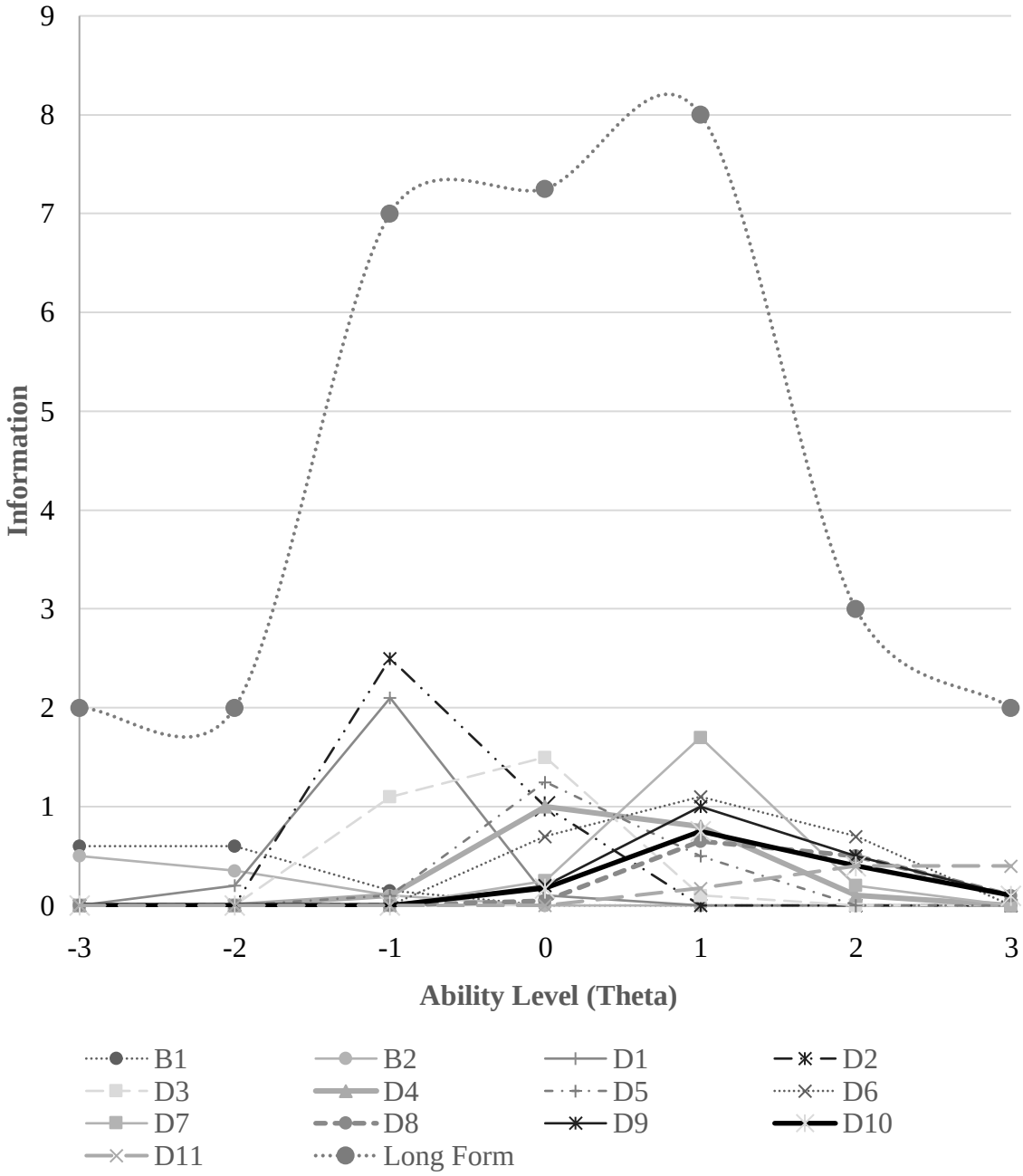

Figure 2. IRT item information functions (Training).

Table 7

IRT Item Information of Selected Items and Test Information of RCM Long Form (Training)

Item/ Form	Ability Level (Theta)						
	–3	–2	–1	0	1	2	3
B1	0.60	0.60	0.15	0.00	0.00	0.00	0.00
B2	0.50	0.35	0.10	0.00	0.00	0.00	0.00
D1	0.00	0.20	2.10	0.10	0.00	0.00	0.00
D2	0.00	0.00	2.50	1.00	0.00	0.00	0.00
D3	0.00	0.00	1.10	1.50	0.10	0.00	0.00
D4	0.00	0.00	0.10	1.00	0.80	0.10	0.00
D5	0.00	0.00	0.10	1.25	0.50	0.00	0.00
D6	0.00	0.00	0.00	0.70	1.10	0.70	0.00
D7	0.00	0.00	0.00	0.25	1.70	0.20	0.00
D8	0.00	0.00	0.00	0.05	0.65	0.50	0.10
D9	0.00	0.00	0.00	0.20	1.00	0.50	0.10
D10	0.00	0.00	0.00	0.18	0.75	0.40	0.10
D11	0.00	0.00	0.00	0.00	0.18	0.40	0.40
RCM Long Form	2.00	2.00	7.00	7.25	8.00	3.00	2.00

Specifically, items B1, D1, D2, D3, D4, D5, D6, D7, D8, D9, and D10 provided most item information across ability levels relative to all other items. Although B2 provided some amount of information at the lower end of the ability level (II_{B2} = 0.60 at Ability Levels –3 and –2), it was locally dependent with several other items that provided higher levels of information (i.e., II_{D1} = 2.10 at Ability Level –1; II_{D2} = 2.50 and 1.00 at Ability Levels –1 and 0; II_{D3} = 1.10 and 1.50 at Ability Levels –1 and 0; and II_{D6} = 1.10 at Ability Level 1). The standardized LD χ^2 statistics indicating the level of local independence were excessively large among those items: B2 and D1 (78.3p), B2 and D2 (26.5n), B3 and D3 (24.1p), B2 and D6 (11.4n). Under locally dependent conditions, only one item of each pair could be retained in order to reduce item redundancy. Given that the inclusion of D1, D2, D3, and/or D6 would save a substantially greater amount of test information, B2 was dropped first. Although D11 provided a similar amount of item information as B2, it was primarily at the upper end of the ability level (0.40 at Ability Levels 2 and 3). Thus, the inclusion of D11 did not appear to contribute to the overall test information substantially, and this was further supported by the slight increase in internal consistency after including D11 into multiple potential short forms (e.g., Cronbach's alpha increased from .72 to .73 after adding D11 to D1, D2, D4, D6, D7, D8, D9, and D10).

In further evaluating items B1, D1, D2, D3, D4, D5, D6, D7, D8, D9, and D10, standardized LD χ^2 statistics were excessively large between B1 and D1 (79.4p). That means either B1 or D1 could be retained on the short form in order to decrease local dependence. D1 was selected given that it provides relatively more item information around the middle range of the ability level (II_{D1} = 2.10 at Ability Level –1) relative to B1 which contributes more

information at the lower end (II_{D1} = 0.60 at Ability Levels –3 and –2), as shown in Table 7 and Figure 2. However, short forms including B1 were still created for the purpose of further examining the overall functioning of the short forms with B1 given that it is the only item to be considered from the CPM. As such, different short forms were created to include either B1 or D1 in order to evaluate the contribution of each item and determine which item should be retained. Detailed information about B1 and D1 (Item Characteristic Curves) are reported in Figure 3 in order to provide additional information about the two items.

Among those items, standardized LD χ^2 statistics for D2 and D3 (10.8p) and D4 and D5 (10.8p) indicated a high level of local dependence (> 10). Given that values near or at the cut-off point are generally less informative or helpful, such values are considered indeterminate. Therefore, two versions of short forms were evaluated, one form with all four items with slightly elevated LD χ^2 statistics and one form including only the item from each pair that provided relatively higher item information (II_{D2} = 2.50 and 1.00 at Ability Levels –1 and 0, II_{D3} = 1.10 and 1.50 at Ability Levels –1 and 0; II_{D4} = 1.00 and 0.80 at Ability Levels 0 and 1, II_{D5} = 1.25 at Ability Level 0).

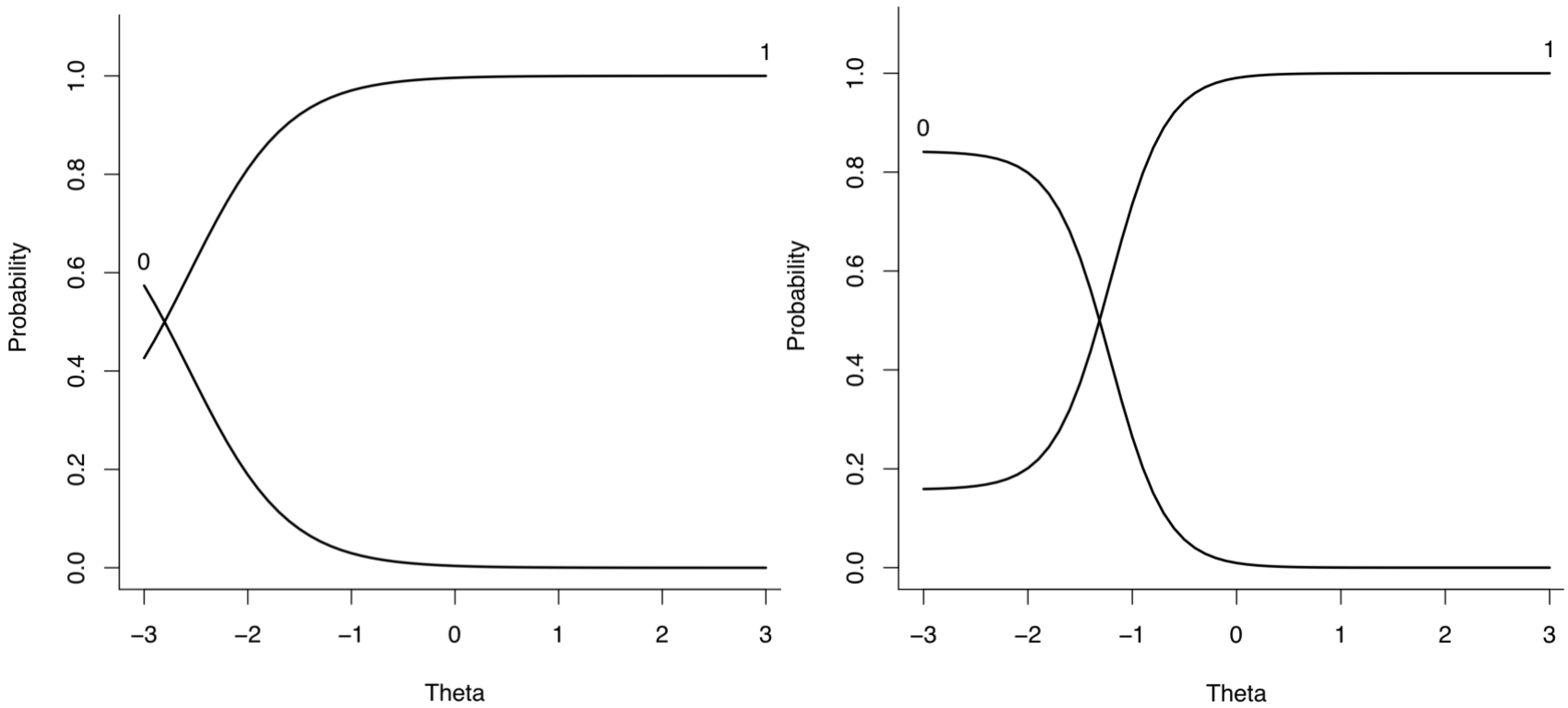

Figure 3. Item Characteristic Curves for Items B1 (left: a = 2.1; b = –2.6; g = .2) and D1 (right: a = 3.7; b = –1.2; g = .2).

In sum, four short forms were constructed and examined. Specifically, Short Form 1 included all items with slightly higher LD χ^2 statistics (D2, D3, D4, & D5) and the only item (B1) from the CPM that provided item information at the lower end of the ability level (–3 and –2). Short Form 2 also incorporated D2, D3, D4, and D5 but replaced B1 with D1 given the latter contributed considerably more item information in the lower to middle ability level (–1). Short Form 3 retained Item B1 but excluded D3 and D5 because they were locally dependent with D2 and D4 and provided less item information. Similarly, Short Form 4 also retained D2 and D4 but included D1 instead of B1. The test information functions for the four short forms are plotted in Figure 4, and the respective item information for the four short forms is reported in Table 8. As displayed in Figure 4, the test information functions of Short Forms 1 and 2 resembled that of the RCM long form; whereas, Short Forms 3 and 4 did so to a lesser extent due to the loss of item information as a result of the exclusion of D3 and D5.

As shown in Table 9, the marginal reliability based on independent IRT analyses with selected items and Cronbach's alpha coefficients for all four short forms were all within the acceptable range ($\alpha > .70$) with Short Forms 1 and 4 being relatively higher. The marginal reliability estimates for all four short forms were lower than the RCM long form ($\alpha = .84$); however, the Cronbach's alpha coefficients for the short forms were higher than the RCM long form ($\alpha = .65$).

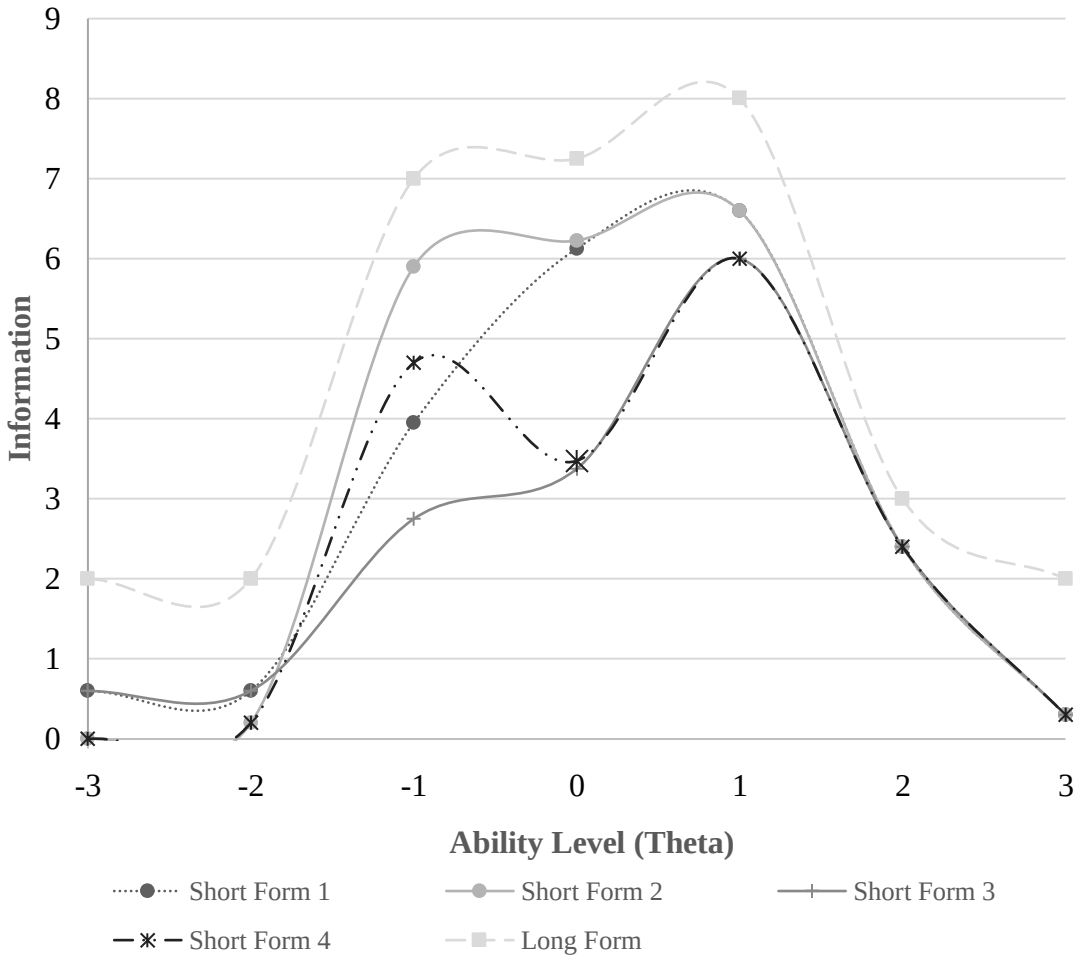

Figure 4. IRT test information functions of short forms (Training).

Table 8

IRT Test Information of Short Forms and Long Form (Training)

	Ability level (Theta)						
	–3	–2	–1	0	1	2	3
Short Form 1	0.60	0.60	3.95	6.13	6.60	2.40	0.30
Short Form 2	0.00	0.20	5.90	6.23	6.60	2.40	0.30
Short Form 3	0.60	0.60	2.75	3.38	6.00	2.40	0.30
Short Form 4	0.00	0.20	4.70	3.48	6.00	2.40	0.30
Long Form	2.00	2.00	7.00	7.25	8.00	3.00	2.00

Table 9

Items, Local Dependence, Reliability, and Test Information for Short Forms (Training)

SF	Items	# of items	LD χ^2	Marginal reliability (Independent IRT analyses)	Cronbach's alpha	Comparison of SF & LF TIFs
1	D2, D3, D4, D5, D6, D7, D8, D9, D10, B1	10	Tolerable D2–D3 (LD χ^2 = 10.8p) & D4 - D5 (LD χ^2 = 10.8p)	.75	.77	Similar
2	D2, D3, D4, D5, D6, D7, D8, D9, D10, D1	10	Tolerable D2–D3 (LD χ^2 = 10.8p) & D4 - D5 (LD χ^2 = 10.8p)	.78	.79	Similar
3	D2, D4, D6, D7, D8, D9, D10, B1	8	Adequate	.71	.69	Different at θ = 0
4	D2, D4, D6, D7, D8, D9, D10, D1	8	Adequate	.70	.73	Different at θ = 0

Note. SF = short form. LF = long form. LD χ^2 = local independence χ^2. TIF = Test Information Function curve.

Initial Testing of Short Forms Using Validation Sample

Based on cross-validation, the correlation coefficients (Table 10) indicated strong positive relationships between the two sets of theta scores. Cronbach's alpha coefficients for Short Forms 1, 2, and 4 were generally acceptable (> .70); whereas, that for Short Form 3 was minimally acceptable. Based on independent IRT analyses, Short Forms 1, 2, and 4 were projected to yield generally acceptable (> .70) marginal reliability; in comparison, the estimate for Short Form 3 was also minimally acceptable. Based on cross-validation, marginal reliability coefficients were all within the minimally acceptable range (> .60), with Short Forms 1 and 2 being higher than Short Forms 3 and 4. However, it is important to note that marginal reliability values obtained from independent IRT analyses were smaller than those from cross-validation. The magnitude of the difference ranged from .10 to .14 for the four potential short forms.

Table 10

Correlations and Reliability Coefficients for Short Forms (Validation)

	r	Cronbach's alpha	Independent Marginal reliability	Cross-validation Marginal reliability	Difference in marginal reliability[a]
SF 1	.85	.70	.77	.66	.11
SF 2	.85	.79	.80	.66	.14
SF 3	.82	.61	.71	.61	.10
SF 4	.83	.73	.72	.61	.11
LF	–	.80	.90	.90	–

Note. SF = short form. LF = long form. *r* = correlation between theta scores from short and long forms. [a]Difference between marginal reliability (independent IRT) and cross-validation estimate.

Additionally, local independence LD χ^2 statistics (based on independent IRT analyses with selected items) were 10.9p (> 10) between Items D2 and D3, as well as 8.1p (< 10) between Items D4 and D5 on Short Form 1. Correspondingly, on Short Form 2, LD χ^2 statistics were 2.5p (< 10) between D2 and D3, as well as 10.5p (> 10) between D4 and D5. This indicates that local independence either remained at the same level or decreased to some extent when utilizing the validation sample.

Chapter 5

DISCUSSION

The goal of this study was to develop a short form for the full-length Raven's Combined Matrices (RCM) for use with adolescents and young adults in Pakistan. To accomplish this goal, IRT analyses were employed to identify prospective items from the RCM long form for inclusion on the short form. The total sample was randomized into two subsamples (training and validation) to examine stability of findings. Using the training sample, items were selected to maximize the amount of item information and minimize local dependence in order to create a short form that would approximate the full-length RCM. Using these criteria, four potential short forms were constructed based on IRT results from the training sample. These forms were then tested using the validation sample.

Interpretation of Evidence for Potential Short Forms

Short Form 1. Short Form 1 includes 10 items (B1, D2, D3, D4, D5, D6, D7, D8, D9, D10). Based on the training sample, Short Form 1 provides most information for individuals between Ability Levels –1 and 2, and it provides a small amount of information for Ability Levels –3 and –2. It is psychometrically similar to the RCM long form as demonstrated by similar TIFs (Figure 4). Short Form 1 demonstrates a level of redundancy (between Items D2 and D3 and Items D4 and D5) that is considered tolerable (Chen & Thissen, 1997). Based on cross-validation, Short Form 1 appears to be strongly correlated with the RCM long form (r = .85).

Across both the training and validation samples, all internal consistency coefficients for Short Form 1 are within the acceptable range (.70–.77) for research purposes (DeVellis, 1991; George & Mallery, 2003; Nunnally, 1967, 1978). However, the Cronbach's alpha coefficients dropped substantially with the validation sample. Using cross-validation, marginal reliability of Short Form 1 is within the minimally acceptable range (.66); however, it is still considered acceptable during early stages of test validation research (Nunnally, 1967).

Short Form 2. Short Form 2 also is comprised of 10 items (D1, D2, D3, D4, D5, D6, D7, D8, D9, D10). Using the training sample, Short Form 2 yields most information for individuals with Ability Levels between –1 and 2, but it provides essentially no information for the extremely high or low ability levels. It is psychometrically *most* similar to the RCM long form as demonstrated by a TIF curve most comparable to that of the RCM long form (Figure 4). Compared to Short Form 1, Short Form 2 provides more information at Ability Level –1 due to the inclusion of Item D1 rather than B1. Similar to Short Form 1, Short Form 2 also demonstrates some item redundancy (between Items D2 and D3 and Items D4 and D5) that is considered tolerable (Chen & Thissen, 1997). Based on cross-validation, Short Form 2 is also strongly correlated with the RCM long form (r = .85).

All internal consistency coefficients for Short Form 2 fall solidly within the upper end of the acceptable range (.78–.80) for research purposes (DeVellis, 1991; George & Mallery, 2003; Nunnally, 1967, 1978), and the results are consistent across both samples. Using cross-validation, marginal reliability of Short Form 2 is within the minimally

acceptable range (.66) and is also considered acceptable while validating measures during initial stages (Nunnally, 1967).

Short Form 3. Short Form 3 consists of 8 items (B1, D2, D4, D6, D7, D8, D9, D10). When analyzed using the training sample, Short Form 3 maintains most information for individuals with Ability Level –1 and Ability Level 1, and it also contains a small amount of information for Ability Levels –3 and –2. Psychometrically, Short Form 3 appears less similar to the RCM long form, particularly at Ability Level 0 (Figure 4). Redundancy on Short Form 3 was reduced to an adequate level by the removal of items with high local dependence (i.e., D3 and D5; Chen & Thissen, 1997). Using cross-validation, Short Form 3 is still strongly correlated with the RCM long form (r = .82); however, the level of association is slightly lower in comparison to Short Forms 1 and 2

Using both the training and validation samples, all internal consistency coefficients for Short Form 3 are within the lower end of the acceptable range (.61–.71) for research purposes; however, the Cronbach's alpha coefficients dropped considerably with the validation sample. Using cross-validation, marginal reliability of Short Form 3 is within the lower end of the minimally acceptable range (.61).

Short Form 4. Short Form 4 also includes 8 items (D1, D2, D4, D6, D7, D8, D9, D10). Using the training sample, Short Form 4 retains most information for individuals with Ability Level –1 and Ability Level 1, but it provides essentially no information for individuals at the extremely high or low ability levels. Psychometrically, Short Form 4 is not as similar to the RCM long form as it provides a reduced level of information at Ability Level 0 (Figure 4). Short Form 4 offers more information at Ability Level –1

than Short Form 3 due to the inclusion of D1 rather than B1. Similarly, item redundancy on Short Form 4 was reduced to an adequate level by the removal of items with high local dependence (i.e., D3 and D5; Chen & Thissen, 1997). Short Form 4 is still strongly correlated with the RCM long form (r = .83), although it is slightly lower than those for Short Forms 1 and 2.

Internal consistency coefficients for Short Form 4 consistently fall within the lower end of the acceptable range (.70–.73) for research purposes across the training and validation samples. Using cross-validation, marginal reliability of Short Form 4 is within the lower end of the minimally acceptable range (.61).

Selection of best short form. Collective evidence provides support for Short Form 2 as the best short form for three reasons. First, Short Form 2 provides similar levels of test information to the RCM long form across the broadest range of ability levels. In comparison, the test information of Short Forms 1 and 4 dropped substantially at Ability Level –1 and 0 respectively, and the amount of test information of Short Form 3 decreased to the greatest extent due to the loss of information at *both* of these ability levels. Moreover, although the inclusion of B1 rather than D1 on Short Forms 1 and 3 provide some information for Ability Levels –3 and –2, the amount of information (0.60 for both ability levels) is actually considered negligible (Petrillo, Cano, McLeod, & Coon, 2015). As a result, none of these three short forms resemble the TIF of the RCM long form closely. In contrast, Short Form 2 essentially preserves the shape of the original TIF of the RCM long form, despite the loss of a small amount of information at Ability Levels –3 and –2.

Second, Short Form 2 provides the maximum amount of information around the middle range of the ability level (between Ability Levels –1 and 2). This is preferred given the primary goal of this study was to develop a short form for use with the general population of adolescents and young adults in Pakistan. As mentioned previously, the test information provided by Short Form 3 and Short Form 4 was substantially reduced at the middle range of the ability level. Although Short Forms 1 and 3 provide a small amount of information for the lower end of the ability level (Ability Levels –3 and –2), the top priority for this project was maximizing information in the middle of the ability range given the primary goal of this study.

Third, Short Form 2 consistently demonstrates the highest level of internal consistency (within the upper end of the acceptable range) compared to other three short forms across the training and validation samples. Similarly, based on cross-validation, the marginal reliability for Short Form 2 remains among the highest. In addition, Short Form 2 is strongly correlated with the RCM long form. Thus, Short Form 2 appears to provide relatively the most reliable scores compared to other short forms.

Nevertheless, while it was maintained at a tolerable level, Short Form 2 contained two items (D 3 and D 5) with LD χ^2 values slightly above the cutoff score (> 10). However, further examination using the validation sample revealed a decreased level of local dependence between one pair of items, D2 & D3 (LD χ^2 = 2.5p) while the LD χ^2 between D4 & D5 remained at the same level (i.e., 10.5p). This indicates that the local dependence between D2 and D3 may be attributed to random sampling errors. Overall, Short Form 2 was determined to be the best short form based on collective evidence.

Final RCM Short Form in Context of Prior Short Forms

As detailed in the literature review (Chapter 2), five other Raven's short forms have been developed previously using a variety of item-reduction strategies for a number of assessment purposes and populations. Table 11 provides a summary of key aspects of these measures along with the final form resulting from the current study. Given the lack of validity data for those measures, the clinically derived short forms (Bouma et al., 1996; Caffarra et al., 2003; Chiesi et al., 2012; Elst van der et al., 2013; Smits et al., 1997) are not included. In addition, it is important to note that the previous short forms were constructed from different versions of the Raven's Matrices tests and in different cultural contexts. Specifically, all existing short forms were developed for use with Western populations; whereas, the current study was the first attempt to develop a Raven's short form for use in Pakistan. This study also represents an initial attempt to use IRT test information to increase measurement efficiency in the Raven's Matrices tests. This comparison is intended to situate the RCM short form relative to short forms developed using different item-reduction strategies and for use in different cultural contexts.

In comparison to the majority of the short forms, the length of the 10-item RCM short form is briefer and it has resulted in the greatest percentage of reduction in the number of items, with the exception of the Bilker et al. (2012) forms. Specifically, the reduction from 48 to 10 items on the RCM short form represents a 79% decrease in the number of items. Cronbach's alpha coefficients for the RCM short form, the Sefcek et al. (2016) form, and the Bilker et al. (2012) forms are comparable, with all values falling in the upper end of the acceptable range; however,, the alpha coefficients for the two 12-

item short forms (Arthur & Day, 1994; Bors & Stokes, 1998) are substantially lower. Although the RCM short form has fewer items compared to the 12-item short forms (Arthur & Day, 1994; Bors & Stokes, 1998), the alpha coefficient for the RCM short form is comparatively higher. Compared to the RCM long form, the alpha coefficient for the RCM short form dropped from .80 to .79. However, this amount of decrease (< .10) after item removal was considered small (Lei, Wu, DiPerna, & Morgan, 2009).

The RCM short form is strongly correlated with the RCM long form. However, the level of association is lower than the Bors and Stokes (1998) and Bilker et al. (2012) forms. The correlations of the Bilker et al. (2012) forms were the highest, which is perhaps not a surprise given that item selection in their study heavily relied on the correlations between observed scores and predicted scores to select items.

Table 11

Summary of Statistical Indices for All Short Forms

Short form	Sample/ target population	Approach	# of items[a]	% Item reduction[c]	Short form reliability (α)	Reduction in reliability[c] (%)	Short/full correlation
Wytek et al. (1984)	Psychiatric patients with cognitive deficits; Vienna, Austria	Iterative selection using Rasch statistics	30/ 60	50%	Not reported	Not reported	Not reported
Arthur and Day (1994)	University students; USA	CTT	12/ 36	67%	.65	.21 (24.42%)	.66
Bors and Stokes (1998)	University students; Canada	CTT	12/ 36	67%	.73	.11 (13.09%)	.92
Bilker et al. (2012)	Healthy adults & psychiatric patients; USA	Poisson predictive model	9/ 60	85%	.80 (A) .83 (B)	.16 (16.67%) .13 (13.54%)	.98 (A) .98 (B)
Sefcek et al. (2016)	University students; USA	CTT	18/ 36	50%	.79	Not reported	Not reported
Current Study (Zhong, 2019)	Adolescents and young adults; Pakistan	IRT Test Information	10/ 48[b]	79%	.79 (training & validation samples)	.01 (1.25%)	.85 (theta scores using cross-validation)

Note. [a]Number of items (short form/ full-length measure). [b]RCM short form/ RCM long form. [c]Relative to full-length/long form from which the short form was derived.

Limitations and Directions for Future Research

Test validation is an ongoing process of accumulating different sources of validity and reliability evidence rather than an activity that occurs once a measure is developed, and it begins with test design and continues throughout test development and implementation (AERA et al., 2014; Cook & Hatala, 2016). As such, several important limitations to the current study point to directions for future research.

Missing data. Basal and ceiling rules were applied during test administration in order to decrease the amount of time and resources needed for data collection. As such, less than 5% of the participants actually completed Sets A and AB of the full-length RCM, and these two sets of items were excluded from the current data analyses. Although the loss of data was unfortunate, the use of only Sets B, C, D, and E served to increase confidence in the item selection and the final short form resulting from this study. Future studies could replicate the current study using IRT with an independent sample from Pakistan without applying the basal and ceiling rules.

Questionable stability in parameter estimation. Initial testing of the four potential RCM short forms using the validation sample revealed that marginal reliability values from the independent IRT analyses dropped to an appreciable extent in cross-validation. This discrepancy in reliability estimates can be indicative of instability in the IRT estimation procedure, which consequently may reduce validity of the IRT item parameter estimates in this study. As such, the purported benefits of IRT may not have been fully realized in this study. This may be explained by the amount of missing data on Set B of the RCM. Thus, a replication research study using a larger sample is recommended.

Initial reliability and validity evidence. In addition, the current study was an initial effort to develop a short form for use in Pakistan. During this early stage of test development, only limited reliability and validation evidence was examined using the dataset currently available. However, given that the Raven's Matrices tests were developed originally for use with Western populations, it is important to evaluate how the culture of Pakistan could potentially impact the psychometric properties of the RCM short form. Given the aforementioned limitations, three important lines of future research with Pakistani samples are recommended in order to validate the RCM short form resulting from the current study. Specifically, these lines focus on validity, reliability, and test fairness of the RCM short form (AERA et al., 2014).

Validity. Additional sources of validity evidence (e.g., content, response processes, internal structure, relations with other variables) are necessary to determine the appropriateness of using scores from the RCM Short Form within the cultural context of Pakistan.

Content evidence. Content evidence refers to the relationship between the content of the test and the construct it is intended to measure (AERA et al., 2014). When the RCM short form is applied in Pakistan, test items can potentially be perceived differently. As such, whether the test items can accurately reflect the same underlying construct in the cultural context of Pakistan should be evaluated. Content evidence can be examined by an expert panel with knowledge of cognitive development and education in Pakistan. Emphasis should be placed on whether the visual patterns and underlying numerical relationships indeed measure abstract reasoning in the cultural context of Pakistan. In

addition, the relative importance of aspects of the content (e.g., visual patterns) and sensitivity can also be examined by the expert panel.

Response processes. Validity evidence based on the response processes during testing is the fit between the construct and the test taker's response to test items during the problem-solving process (AERA et al., 2014). This is important to determine whether there exists a mismatch between intended and actual cognitive processes that the Raven's items evoke when used in Pakistan. Given the aforementioned cultural and educational influences, individuals in Pakistan may analyze the visual patterns in alternative ways by engaging in thinking processes or actions that are not abstract reasoning. This source of validity evidence can be obtained by interviewing test takers in Pakistan about their problem-solving processes during testing. Specifically, test takers should demonstrate the use of reasoning by identifying the conceptual principles that govern the visual patterns in each row and/or column in order to locate the missing piece rather than other irrelevant rules such as aesthetic principles (e.g., novelty, use of space to convey values, or use of a shape to capture emotion).

Internal structure. Evidence for the internal structure addresses the relationship between the test items and the underlying construct (AERA et al., 2014). In order to validate the RCM short form, how the 10 test items relate to the overarching abstract reasoning ability (Raven & Court, 1989) should be investigated in order to ensure item homogeneity. The majority of previous factor analytic studies of the full-length Raven's Matrices tests have supported a single *g* factor solution, an indicator of general intelligence based on the Spearman's theory of cognitive ability (Burke, 1972; Jensen,

1974; Raven et al., 1998a). Using confirmatory analyses, Arthur and Day (1994) found that a single-factor model adequately represents the underlying structure of their 12-item short form. Similarly, dimensionality studies focusing on Pakistani samples (e.g., using exploratory factor analysis based on tetrachoric correlations or confirmatory factor analysis for testing multi-factor models) would provide insight into the factor structure of the RCM short form.

Relationships with other variables. This source of evidence refers to the degree to which scores from the target measure are related to those from other external measures (AERA et al., 2014). External measures can be grouped into tests that have been established to assess similar constructs as the target measure (i.e., convergent) and those designed to measure different constructs (i.e., discriminate). Convergent evidence for the RCM short form should be examined by analyzing the relationships of the short form scores to other tests measuring the same constructs (e.g., intelligence and achievement). Locally accessible and validated measures in Pakistan are necessary to examine this line of validity evidence. For example, the Test of Non Verbal Intelligence for Youth (Chaudhry, Khalid, & Mohsin, 2018); the Standardized Achievement Test for assessing knowledge of Language (Sindhi, Urdu, and English), Math, and Science (Chang & Jilani, 2015); or criterion-referenced tests in schools in Pakistan could be considered as potential criterion measures for examining convergent evidence. When the RCM short form is validated with young adults who have entered the labor market, researchers may consider measures that assess behaviors, knowledge, and skills necessary to perform a job in employment settings as a form of predictive validity evidence. Regarding discriminant

evidence, scores from the RCM short form should not be highly correlated with scores from measures designed to assess different constructs such as measures of personality. For example, the Big Five Personality Test (e.g., extraversion, neuroticism, and agreeableness) has been used and validated in Pakistan (Ahmad, 2010), which can be used as potential criterion measures in Pakistan to investigate discriminant evidence.

Reliability/precision. Reliability/precision refers to the consistency of the test scores across instances of the testing procedure (AERA et al., 2014). Using a modern framework, IRT can serve as a powerful tool to address the reliability/precision of the RCM short form based on test information functions (AERA et al., 2014). The amount of item information contributed by each item and the respective location can be examined. In addition, test information obtained from IRT parameter estimates can be manually converted into a standard estimate of reliability (i.e., reliability = 1 – [1/information]). In doing so, common rules of thumb for interpreting reliability (DeVellis, 1991; Nunnally, 1978) can be applied in evaluating the amount of the item information (Petrillo et al., 2015).

Test fairness. Test fairness refers to the extent to which the test score interpretations for intended uses are valid for different subgroups of test-takers, such as race, ethnicity, religion, gender, language, culture, and socioeconomic status (AERA et al., 2014). As previously discussed in the literature review, poverty and quality of education (especially educational inequalities among girls and individuals living in remote villages) are major issues faced by Pakistan (Malik et al., 2015). These factors can potentially affect an individual's cognitive development, and thus may impact test

performance on the RCM short form either at the item level or test level. At the item level, efforts should be made to evaluate whether each of the items functions differently for different subgroups in order to detect the source of differences. Differential item functioning (DIF) analysis can be conducted to address item equivalence (AERA et al., 2014), especially to identify potential sources of differences as related to the specific cultural context (International Test Commission, 2017). Analyses of DIF (e.g., IRT likelihood ratio test) across gender, SES, and geographic area groups are important directions for future research. At the test level, the first step would be to examine if the means of test scores of the RCM short form are different between subgroups on variables of interest (e.g., gender) when using Pakistani samples, and if so, whether the differences are statistically significant. The detection of differences may not always indicate biases in an item; however, such items should be flagged and evaluated for sources of validity threats in the cultural context of Pakistan (e.g., issues previously addressed in *Content evidence* and *Reponses processes*).

Potential Implications

Based on the results of the current study, there are potential implications within two broad domains. First, results of this study illustrate that the IRT approach is a promising methodology for developing short forms for non-verbal intelligence tests. Second, with appropriate validation as outlined previously, the 10-item RCM short form may be useful for research purposes in Pakistan. Any clinical applications, however, would require additional validation with appropriate target clinical populations.

Methodology. The results of this study provide further evidence that IRT is a promising approach for short form development for non-verbal intelligence tests. This approach differs from the commonly used classical approaches for developing short forms for the Raven's Matrices tests (and other measures) since the 1980s. In this study, the decrease in Cronbach's alpha was minimal (1.25%) after reducing the number of items by nearly 80% based on the validation sample, and the internal consistency coefficients consistently remain in the upper end of the acceptable range. This demonstrates that use of IRT techniques can facilitate construction of short forms that will result in substantial time savings while preserving a sufficient amount of test information (reliability). Furthermore, testing in this study indicated that the local independence was tolerable and essential unidimensionality was not violated, further supporting the feasibility of using IRT and improved the accuracy in IRT parameter estimation compared to the Wytek et al. (1984) study.

The results of this study also revealed that it is feasible to use a more complex IRT model than the Rasch model used by Wytek et al. (1984) to analyze the *test scores* of the RCM long form. Wytek et al. (1984) was the first research team to incorporate IRT statistics when constructing a Raven's short form. However, Wytek et al. (1984) used the Rasch model without conducting model comparison, and the Rasch model was essentially considered a 1PL IRT model. The Rasch model was overly restrictive because it assumed that all items had equal discrimination and there was no provision for guessing in the model. In comparison, testing in this study suggested that the 3PL IRT model could be used for analyzing the test scores of the RCM long form. The 3PL model

allowed item discrimination and guessing to be taken into account, thus providing more sophisticated information during the item selection process.

Potential uses in Pakistan. Compared to the full-scale RCM, the final RCM short form resulting from the current study substantially reduces the number of items while maintaining a similar level of information as the RCM long form. As such, the 10-item RCM short form is considerably easier and faster to administer. Although additional validation studies with Pakistani samples (as outlined in the previous section) are essential before the RCM short form could be utilized for research purposes, it holds potential as a brief estimate of cognitive ability. From a researcher's perspective, the RCM short form could be particularly useful in large-scale research studies in Pakistan, as a lower-middle-income country (World Bank, 2017), where resources are often limited. In large-scale studies, resources such as time, money, human, and training are important considerations in analyzing cost-effectiveness and efficiency during data collection. The 10-item RCM short form could serve as an efficient tool for a quick, accurate estimate of general intelligence, minimizing associated resources and costs needed and eventually promoting the feasibility of using this measure in large-scale research in Pakistan.

At the individual level, the 10-item RCM short form would significantly reduce the response burden of testing for test takers, thus decreasing potential emotional distress and physical fatigue during test administration. Though helpful for the general population, this reduction is particularly salient for certain subpopulations in Pakistan. For example, individuals in Pakistan are at higher risk of failing to reach their full cognitive potential due to multiple factors such as poverty, poor health, and unstimulating

environments (Grantham-McGregor et al., 2007). In addition, the RCM short form may be less intimidating to individuals who did not receive formal education and may find cognitive testing a daunting experience, such as individuals from low-income families in Pakistan who may have dropped out of school early (Malik et al., 2015). Overall, compared to the 72-item full-length RCM, the 10-item RCM short form should induce less stress at the individual level and thus be easier for those individuals to complete. As a direct result of the reduced number of items, researchers may be able to obtain higher response rates with the RCM short form rather than the full-length RCM.

Conclusion

The purpose of the current study was to develop a RCM short form for potential use with typically developing adolescents and young adults in Pakistan. In order to achieve this goal, item-level data were analyzed using item response theory (IRT) models. Using the results of these analyses, items with more information across all ability levels were selected, and items with high local dependence were eliminated. The final RCM short form provides similar levels of test information to the RCM long form overall (as demonstrated by the similar information curves), maintains the maximum amount of information around the middle range of the ability level of the population of inference (i.e., adolescents and young adults in Pakistan), and consistently demonstrates acceptable reliability for research purposes. However, it should be emphasized that cross-validation also suggested instability in parameter estimation to a certain degree. Additional replication studies and psychometric studies are critical to examine whether

the current results can be reproduced and to provide additional insight regarding the properties and potential utility of the RCM short form.

This study represents an initial attempt to use the IRT approach to increase measurement efficiency in the full-length RCM. The results illustrate that the IRT approach is a promising methodology to develop short forms for non-verbal intelligence tests. Among all Raven's short forms development efforts to date, this was the first attempt to develop a Raven's short form using a Pakistani sample. Provided the results of future validation studies are positive, the RCM short form may serve as a quick and accurate estimate of general intelligence for use with *adolescents and young adults* in Pakistan in order to monitor cognitive development in response to Pakistan's improvement efforts.

REFERENCES

Ahmad, I. (2010). The Big Five Personality Inventory: Performance of students and community in Pakistan. *Journal of Behavioural Sciences, 20*, 63–79.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. doi:10.1109/TAC.1974.1100705

Alderman, H., Behrman, J. R., Khan, S., Ross, D. R., & Sabot, R. (1996). Decomposing the regional gap in cognitive skills in rural Pakistan. *Journal of Asian Economics, 7*, 49–76. doi:10.1016/S1049-0078(96)90034-2

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Amjad, R., & MacLeod, G. (2014). Academic effectiveness of private, public and private – public partnership schools in Pakistan. *International Journal of Educational Development, 37*, 22–31. doi:10.1016/j.ijedudev.2014.02.005

Anderson, D., Kahn, J. D., & Tindal, G. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Applied Measurement in Education, 30*, 163–177. doi:10.1080/08957347.2017.1316277

Ansari, Z. A. (1984). *Validity of Raven's Standard Progressive Matrices for urban and rural school children in Pakistan*. Islamabad, Pakistan: National Institute of Psychology, Centre of Excellence, Quaid-i-Azam University.

Ansari, Z. A., & Iftikhar, M. (1988). Validity of Raven's Standard Progressive Matrices for urban and rural school children in Pakistan (Part-1: Basic facts). *Psychology Quarterly, 19*, 14–27.

Anthony, C. J., DiPerna, J. C., & Lei, P.-W. (2016). Maximizing measurement efficiency of behavior rating scales using Item Response Theory: An example with the Social Skills Improvement System - Teacher Rating Scale. *Journal of School Psychology, 55*, 57–69. doi:10.1016/j.jsp.2015.12.005

Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement, 54*, 394–403. doi:10.1177/0013164494054002013

Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Behrman, J. R., Khan, S., Ross, D., & Sabot, R. (1997). School quality and cognitive achievement production: A case study for rural Pakistan. *Economics of Education Review, 16*, 127–142. doi:10.1016/S0272-7757(96)00045-3

Bhutta, Z. A., Hafeez, A., Rizvi, A., Ali, N., Khan, A., Ahmad, F., . . . Jafarey, S. N. (2013). Reproductive, maternal, newborn, and child health in Pakistan: Challenges and opportunities. *Lancet, 381*, 2207–2218. doi:10.1016/S0140-6736(12)61999-0

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's Standard Progressive Matrices test. *Assessment, 19*, 354–369. doi:10.1177/1073191112446655

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*, 382–398. doi:10.1177/0013164498058003002

Bouma, A., Mulder, J., & Lindeboom, J. (1996). *Neuropsychologische diagnostiek: Handboek [Neuropsychological assessment: Manual]*. Lisse, Netherlands: Swets & Zeitlinger.

Burke, H. R. (1972). Raven's Progressive Matrices: Validity, reliability, and norms. *The Journal of Psychology, 82*, 253–257. doi:10.1080/00223980.1972.9923815

Caffarra, P., Vezzadini, G., Zonato, F., Copelli, S., & Venneri, A. (2003). A normative study of a shorter version of Raven's Progressive Matrices 1938. *Neurological Sciences, 24*, 336–339. doi:10.1007/s10072-003-0185-0

Cai, L. (2017). *flexMIR version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]*. Chapel Hill, NC: Vector Psychometric Group.

Chang, F. H., & Jilani, D. B. S. (2015). *Standardized Achievement Test (SAT) III: Sindh government schools achievement class V & VIII*. Retrieved from Sindh, Pakistan: https://www.researchgate.net/publication/319750369_Standardized_Achievement_Test_SAT_III_Sindh_Government_Schools_Achievement_Class_V_VIII_-_Subjects_Mathematics_Science_and_Languages

Chaudhry, M. I., Khalid, S., & Mohsin, M. N. (2018). Validation of Test of Nonverbal Intelligence for Pakistani youth. *Pakistan Journal of Education, 35*, 223–237.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289. doi:10.2307/1165285

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012). Using the Advanced Progressive Matrices (Set I) to assess fluid ability in a short time frame: An item response theory-based analysis. *Psychological Assessment, 24*, 892–900. doi:10.1037/a0027830

Coles, E., Wells, M., Maxwell, M., Harris, F. M., Anderson, J., Gray, N. M., . . . MacGillivray, S. (2017). The influence of contextual factors on healthcare quality improvement initiatives: What works, for whom and in what setting? Protocol for a realist review. *Systematic Reviews, 6*, 1–10. doi:10.1186/s13643-017-0566-8

Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: A primer for simulation and beyond. *Advances in Simulation, 1*, 1–12. doi:10.1186/s41077-016-0033-y

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/bf02310555

Cuesta, M., & Muñiz Fernández, J. (1999). Robustness of item response logistic models to violations of the unidimensionality assumption. *Psicothema, 11*, 175–182.

DeVellis, R. F. (1991). *Scale development*. Newbury Park, NJ: Sage.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5-18. doi:10.1007/s11136-007-9198-0

Elst van der, W., Ouwehand, C., Boxtel, v. M., Rijn, v. P., Lee, N., & Jolles, J. (2013). The shortened Raven Standard Progressive Matrices: Item Response Theory-based psychometric analyses and normative data. *Assessment, 20*, 48–59.

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

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191. doi:10.1037/0033-2909.101.2.171

Flynn, J. R. (2009). Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938–2008. *Economics & Human Biology, 7*, 18–27. doi:org/10.1016/j.ehb.2009.01.009

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 263–265. doi:10.1207/s15327906mbr2302_8

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference 11.0 update*. Boston, MA: Allyn & Bacon.

Goldstein, E. J. C., Katona, P., & Katona-Apte, J. (2008). The interaction between nutrition and infection. *Clinical Infectious Diseases, 46*, 1582–1588. doi:10.1086/587658

Grantham-McGregor, S., Cheung, Y. B., Cueto, S., Glewwe, P., Richter, L., Strupp, B., & International Child Development Steering Group. (2007). Developmental potential in the first 5 years for children in developing countries. *Lancet, 369*, 60–70. doi:10.1016/S0140-6736(07)60032-4

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347–360. doi:10.1111/j.1745-3984.1984.tb01039.x

Greenfield, P. M. (1998). The cultural evolution of IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 81–123). Washington, DC: American Psychological Association.

Guerrant, R. L., Oriá, R. B., Moore, S. R., Oriá, M. O. B., & Lima, A. A. M. (2008). Malnutrition as an enteric infectious disease with long-term effects on child development. *Nutrition Reviews, 66,* 487–505. doi:10.1111/j.1753-4887.2008.00082.x

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice, 12,* 38–47. doi:10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*, 91–115. doi:10.2307/1164972

Harvey, P. D. (2012). Clinical applications of neuropsychological assessment. *Dialogues in Clinical Neuroscience, 14*, 91–99.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. doi:10.1080/10705519909540118

Hussain, A., & Salfi, N. A. (2011). Causes of low literacy rate in Pakistan: A survey based study. *International Journal of the Book, 8*, 151–164.

IBM. (2011). *IBM SPSS statistics for Windows, Version 20.0*. International Business Machines Corporation.

International Test Commission. (2013). The International Test Commission guidelines (ITC guidelines). *The ITC guidelines on test use.* Retrieved from https://www.intestcom.org/page/17

International Test Commission. (2017). The International Test Commission guidelines (ITC guidelines). *The ITC guidelines for translating and adapting Tests (2nd edition).* Retrieved from https://www.intestcom.org/page/16

J C Raven Ltd. (n.d.). Raven's applications. Retrieved from http://www.jcravenltd.com/applications.htm

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 90,* 185–244.

Khan, E. M. (2017). Constitutional status of Gilgit-Baltistan: An issue of human security. Retrieved from https://www.ndu.edu.pk/issra/issra_pub/articles/margalla-paper/Margalla-Paper-2017/7-Constitutional-Status-Dr-Ehsan-Mehmood-Khan.pdf

Khattak, S. G. (2012). Assessment in schools in Pakistan. *SA-eDUC Journal, 9*, 1–13.

Lei, P.-W., Wu, Q., DiPerna, J. C., & Morgan, P. L. (2009). Developing short forms of the EARLI numeracy measures: Comparison of item selection methods. *Educational and Psychological Measurement, 69*, 825–842. doi:10.1177/0013164409332215

Lezak, M. D. (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford University Press.

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P.-W., Hayes, J. A., Castonguay, L. G., . . . Lin, Y.-C. (2012). Development and initial validation of the Counseling Center Assessment of Psychological Symptoms–34. *Measurement and Evaluation in Counseling and Development, 45*, 151–169. doi:10.1177/0748175611432642

Lord, F. N., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Malik, A. B., Amin, N., Ahmad, K., Mukhtar, E. M., Saleem, M., & Kakli, M. B. (2015). *Education For All national review*. Retrieved from http://unesdoc.unesco.org/images/0022/002297/229718E.pdf

Malik, N. I., Rehman, G., & Hanif, R. (2012). Effect of academic interventions on the developmental skills of slow learners. *Pakistan Journal of Psychological Research, 27*, 135–151.

Martinez, M. E. (2010). Intelligence (Chapter 10). In M. E. Martinez (Ed.), *Learning and cognition: The design of the mind* (pp. 315–353). Upper Saddle River, NJ: Pearson Education.

May, C. R., Johnson, M., & Finch, T. (2016). Implementation, context and complexity. *Implementation Science, 11*, 1–12. doi:10.1186/s13012-016-0506-3

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732. doi:10.1007/s11336-005-1295-9

Mills, C. J., & Tissot, S. L. (1995). Identifying academic potential in students from under-represented populations: Is using the Ravens Progressive Matrices a good idea? *Gifted Child Quarterly, 39*, 209–217. doi:10.1177/001698629503900404

Murray-Kolb, L. E., Rasmussen, Z. A., Scharf, R. J., Rasheed, M. A., Svensen, E., Seidman, J. C., . . . MAL-ED Network Investigators. (2014). The MAL-ED cohort study: Methods and lessons learned when assessing early child development and caregiving mediators in infants and young children in 8 low- and middle-income countries. *Clinical Infectious Diseases, 59,* S261–S272. doi:10.1093/cid/ciu437

Naviwala, N. (2015). *Pakistan's education crisis: The real story*. Washington, DC: Wilson Center Asia Program.

Neisser, U. (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, D.C.: American Psychological Association.

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory (2nd ed.)*. New York, NY: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). The assessment of reliability. *Psychometric Theory, 3*, 248–292.

Pendergast, L. L., Schaefer, B. A., Murray-Kolb, L. E., Svensen, E., Shrestha, R., Rasheed, M. A., . . . MAL-ED Network Investigators. (2018). Assessing development across cultures: Invariance of the Bayley-III Scales Across Seven International MAL-ED sites. *School Psychology Quarterly, 33*, 604–614. doi:10.1037/spq0000264

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health, 18*, 25–34. doi:10.1016/j.jval.2014.10.005

Pfadenhauer, L. M., Gerhardus, A., Mozygemba, K., Lysdahl, K. B., Booth, A., Hofmann, B., . . . Rehfuess, E. (2017). Making sense of complexity in context and implementation: the Context and Implementation of Complex Interventions (CICI) framework. *Implementation Science, 12*, 1–13. doi:10.1186/s13012-017-0552-5

Raven, J., Raven, J. C., & Court, J. H. (1998a). *Raven manual: Section 1 - General overview*. Oxford, England: Oxford Psychologists Press.

Raven, J., Raven, J. C., & Court, J. H. (1998b). *Raven manual: Section 2 - Coloured Progressive Matrices*. Oxford, England: Oxford Psychologists Press.

Raven, J. C. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement, 26*, 1–16. doi:10.1111/j.1745-3984.1989.tb00314.x

Raven, J. C. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1–48. doi:10.1006/cogp.1999.0735

Raven, J. C., & Court, J. H. (1989). Normative, reliabiilty, and validity studies - References. In *Raven manual: Research supplement 4 (Updated)*. Oxford, England: Oxford Psychologists Press.

Raven, J. C., & Raven, J. (2003). Raven Progressive Matrices. In R. S. McCallum (Ed.), *Handbook of Nonverbal Assessment* (pp. 223–237). New York, NY: Springer US.

Riaz, M. N. (1979). A study of intelligence creativity distinction and their relationship with academic achievement. *Pakistan Psychological Studies, 3*, 58–70.

Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition, 52*, 326–333. doi:org/10.1016/S0278-2626(03)00170-2

RStudio Team. (2015). *RStudio: Integrated development for R*. Boston, MA: RStudio.

Sattler, J. M. (2008). *Resource guide to accompany assessment of children: Cognitive foundations (5th ed.)*. San Diego, CA: Author.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. doi:10.1214/aos/1176344136

Sefcek, J. A., Miller, G. F., & Figueredo, A. J. (2016). Development and validation of an 18-Item medium form of the Ravens Advanced Progressive Matrices. *SAGE Open, 6*, 1–7. doi:10.1177/2158244016651915

Shah, W. H., Bader, F., Hussain, E., Jahan, A., Sikandar, F., Wasim, S., . . . Rasmussen, Z. A. (2015). *Description of a cohort of adolescents and young adults from Oshikhandass village, Gilgit, Pakistan: Health status, socio-economic status, and educational attainment*, Poster session presented at the Aga Khan University 9th Health Sciences Research Assembly, Karachi City, Pakistan.

Shamama-tus-Sabah, S., Gilani, N., & Iftikhar, R. (2012). Ravens Progressive Matrices: Psychometric evidence, gender and social class differences in middle childhood. *Journal of Behavioural Sciences, 22*, 120–131.

Simpson, B. B., Tate, J. A., & Weeks, A. (2005). The biogeography of Hoffmannseggia (Leguminosae, Caesalpinioideae, Caesalpinieae): A tale of many travels. *Journal of Biogeography, 32*, 15–27. doi:10.1111/j.1365-2699.2004.01161.x

Smits, C. H., Smit, J. H., van den Heuvel, N., & Jonker, C. (1997). Norms for an abbreviated Raven's Coloured Progressive Matrices in an older sample. *Journal of Clinical Psychology, 53*, 687–697. doi:10.1002/(sici)1097-4679(199711)53:7<687::aid-jclp6>3.0.co;2-f

Soofi, S. B., Hussain, I., Mehboob, N., Hussain, M., Bhatti, Z., Khan, S., . . . Bhutta, Z. A. (2013). Impoverished rural districts of Pakistan: An independent evaluation of impact on educational and cognitive outcomes in Sindh Province, Pakistan. *IDS Bulletin, 44*, 48–56. doi:10.1111/1759-5436.12030

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*, 167–194. doi:10.1111/j.1744-6570.2002.tb00108.x

Sultan, F., & Khan, A. (2013). Infectious diseases in Pakistan: A clear and present danger. *Lancet, 381*, 2138–2140. doi:10.1016/S0140-6736(13)60248-2

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen; & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159–203. doi:10.1177/0146621603027003001

U.S. Census Bureau. (2018). Pakistan demographic data. Retrieved from https://www.census.gov/popclock/world/pk

UNESCO. (2018). Pakistan. *Institute of Statistics*. Retrieved from http://uis.unesco.org/en/country/PK

UNICEF. (2013). Pakistan statistics. Retrieved from https://www.unicef.org/infobycountry/pakistan_pakistan_statistics.html

UNICEF. (2016). UNICEF Data: Monitoring the situation of children and women. Retrieved from https://data.unicef.org/country/pak/

United Nations. (2017). UNFPA Pakistan. Retrieved from https://www.unfpa.org/data/transparency-portal/unfpa-pakistan

Waldrop, J., & McGuinness, T. M. (2017). Measurement-based care in psychiatry. *Journal of Psychosocial Nursing & Mental Health Services, 55*, 30–35. doi:10.3928/02793695-20170818-01

WHO. (2010). Communicable diseases in the South-East Asia Region of the World Health Organization: Towards a more effective response. Retrieved from August 2, 2018 http://www.who.int/bulletin/volumes/88/3/09-065540/en/

Wiley, J., Jarosz, A. F., Cushen, P. J., & Colflesh, G. J. H. (2011). New rule use drives the relation between working memory capacity and Raven's Advanced Progressive Matrices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 256–263. doi:10.1037/a0021613

World Bank. (2017). Pakistan data. Retrieved from https://data.worldbank.org/country/pakistan?view=chart

Wytek, R., Opgenoorth, E., & Presslich, O. (1984). Development of a new shortened version of Raven's Matrices Test for application and rough assessment of present intellectual capacity within psychopathological investigation. *Psychopathology, 17*, 49–58.

Zaki, S., & Beg, M. A. (1969). A comparative study of the validity of Raven's Standard Progressive Matrices (1960), Chicago Non-verbal Examination and IER Scholastic Aptitude Test for Pakistani Children. *Journal of Scientific Research, 4*, 35–43.

Appendix A: Item Parameters

Table 12

IRT Item Parameters for All Raven's Items Based on Training Sample

Item	*a*	*b*	*g*
B1	2.1	-2.6	0.2
B2	1.6	-2.8	0.2
B3	1.0	-3.6	0.2
B4	0.8	-3.8	0.2
B5	0.5	-4.0	0.2
B6	0.1	-0.9	0.2
B7	0.2	8.3	0.2
B8	0.1	22.0	0.1
B9	0.2	7.4	0.2
B10	0.1	15.9	0.2
B11	0.2	10.1	0.1
B12	0.4	6.4	0.1
C1	0.4	-12.3	0.2
C2	0.1	-41.0	0.2
C3	0.2	-7.1	0.2
C4	0.2	-3.6	0.2
C5	0.1	-7.1	0.2
C6	0.3	1.8	0.2
C7	0.1	6.6	0.2
C8	0.5	3.5	0.2
C9	0.2	4.6	0.2
C10	0.3	6.3	0.2
C11	0.5	5.6	0.1
C12	0.2	23.9	0.1
D1	3.7	-1.2	0.2
D2	6.5	-0.6	0.1
D3	3.3	-0.5	0.1
D4	2.9	0.3	0.2
D5	3.3	0.2	0.3
D6	3.2	0.4	0.2
D7	3.3	0.7	0.1
D8	2.2	1.3	0.2
D9	2.4	1.1	0.1
D10	2.0	1.1	0.1

Note. *a* = item discrimination, *b* = item difficulty, *g* = guessing

(Continued)

Table 12

IRT Item Parameters for All Raven's Items Based on Training Sample (continued)

Item	*a*	*b*	*g*
D11	1.5	2.3	0.1
D12	1.0	3.8	0.1
E1	0.1	1.5	0.2
E1	0.1	1.5	0.2
E1	0.1	1.5	0.2
E2	0.0	24.9	0.2
E3	0.1	19.8	0.2
E4	0.1	19.8	0.2
E5	0.1	28.2	0.2
E6	0.1	26.8	0.2
E7	0.1	21.8	0.2
E8	0.1	23.8	0.1
E9	0.1	28.0	0.1
E10	0.2	18.0	0.1
E11	0.1	45.0	0.0
E12	0.1	33.5	0.1

Note. *a* = item discrimination, *b* = item difficulty, *g* = guessing

Appendix B: Standardized LD χ^2 Statistics

Table 13

Standardized LD χ^2 Statistics (B1–B12)

Item	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
B2	34.4p											
B3	12.9p	33.9p										
B4	17.9p	18.3p	17.1p									
B5	11.3p	10.6p		87.7p								
B9	10.0n							28.8p				
B10								16.6p	32.4p			
B11								19.6p	32.8p	22.5p		
B12								11.0p			16.0p	
C3				16.3p								
C4				13.0p								
C5				12.5p								
D1	79.4p	78.3p	77.6p	77.1p	77.1p	77.7n	77.0n	77.2n	77.1n	77.1n	77.1p	77.8p
D2		26.5n	25.0n	28.5n	27.1n	25.2n	24.9p	24.8n	25.2p	24.8p	24.7n	25.1p
D3		24.1p	22.9n	23.7n	23.0n	23.1n	22.7n	22.7n	22.5p	22.9n	22.9p	23.9p
D5			11.1n									
D6		11.4n										
D7				11.9n								
E1	11.2p	11.9p										
E4	10.7p											
E12					12.9n	12.1n	12.2n	12.2n	12.2n	12.1n	12.3n	12.5n

Note. p = positive, n = negative

Table 14

Standardized LD χ^2 Statistics (C2–E8)

Item	C2	C3	C4	C5	C6	C7	C8	C9	C10	D2	D4	E1	E2	E3	E4	E5	E6	E7	E8
C3	53.0p																		
C4	11.2p	120.5p																	
C5	12.5p	97.9p	148.5p																
C6		31.7p	25.8p	40.3p															
C7		24.6p	22.9p	37.2p	60.4p														
C8			12.1p	11.3p		22.8p													
C9		25.4p	15.5p	22.9p	32.1p	54.0p													
C10			14.4p			19.0p													
C11					16.7p	41.4p	13.3p	16.4p	10.9p										
D3										10.8p									
D5											10.8p								
E2												36.3p							
E3												30.3p	15.5p						
E4												23.3p	53.4p	11.9p					
E5												16.8p	36.7p	44.0p	87.1p				
E6													21.3p	12.1p	32.6p	39.7p			
E7												12.7p		25.0p		16.3p	18.0p		
E8															10.4p	15.7p		10.7p	
E9												12.6p	32.8p	13.1p	20.7p	36.1p	12.5p		13.5p

Note. p = positive, n = negative

VITA

Hongxuan (Nicole) Zhong

EDUCATION

Ph.D. in School Psychology, Expected 2019
Doctoral Minor in Educational Psychology
M.Ed. in School Psychology, 2015
The Pennsylvania State University, University Park, PA

Predoctoral Internship, 2017–2018
Center for Mental Health Policy and Services Research
Department of Psychiatry, Perelman School of Medicine
University of Pennsylvania

PUBLICATIONS

Oakland, T. D., **Zhong, N. H.**, & Kane, H. D. (2015). Gender differences in adaptive behavior among children and adolescents: Evidence from the USA. *Mankind Quarterly*, 56, 208–225.

International Test Commission. (2015). *ITC guidelines on test use*. (**N. H. Zhong** & J. H. Wang, Trans.). Beijing, China: The Chinese Psychological Society. (Original work published 2013)

International Test Commission. (2015). *ITC guidelines for translating and adapting tests*. (**N. H. Zhong** & J. H. Wang, Trans.). Beijing, China: The Chinese Psychological Society. (Original work published 2005)

PROFESSIONAL PRESENTATIONS

Zhong, N. H. (2017). *Development of an Abbreviated Version of the Raven's Matrices Based on Item Response Theory Using a Nationally Representative Sample in Gilgit, Pakistan.* Poster presented at the American Psychological Association (APA) 2017 Annual Convention, Washington, DC (August 3–6, 2017)

Oakland, T. D., **Zhong, N. H.,** & Kane, H. D. (2017). *Gender Differences in Adaptive Behavior among Children and Adolescents: Evidence from the USA*. Paper presented at the National Association of School Psychologists (NASP) 2017 Annual Convention, San Antonio, TX (February 21–24, 2017)

Zhong, N. H. (2016). *Structural validity of the Adjustment Scales for Children and Adolescents (ASCA).* Paper presented at the Harvard Student Research Conference, Cambridge, MA.