2021

ADAPTATION AND TRANSLATION OF AUTISM SPECTRUM DISORDER RATING SCALES FOR SCHOOL-AGE CHILDREN: EVALUATING ADHERENCE TO THE INTERNATIONAL TEST COMMISSION GUIDELINES

Tamanna Chhabra

ADAPTATION AND TRANSLATION OF AUTISM SPECTRUM DISORDER RATING SCALES FOR SCHOOL-AGE CHILDREN: EVALUATING ADHERENCE TO THE INTERNATIONAL TEST COMMISSION GUIDELINES

A dissertation submitted in partial

fulfillment of the requirements for the degree of

DOCTOR OF PSYCHOLOGY

to the faculty of the

DEPARTMENT OF PSYCHOLOGY

of

ST. JOHN'S COLLEGE OF LIBERAL ARTS AND SCIENCES

at

ST. JOHN'S UNIVERSITY

New York

by

TAMANNA CHHABRA

Date Submitted: _____ Date Approved: _____

_____ _____

Tamanna Chhabra Mark D. Terjesen, Ph.D.

ABSTRACT

ADAPTATION AND TRANSLATION OF AUTISM SPECTRUM DISORDER RATING SCALES FOR SCHOOL-AGE CHILDREN: EVALUATING ADHERENCE TO THE INTERNATIONAL TEST COMMISSION GUIDELINES

Tamanna Chhabra

The current review examines adaptations and translations of Autism Spectrum Disorder (ASD) rating scales for school-aged children around the world. ASD rating scales are a quick and efficient way to diagnosing ASD, and adaptations of ASD scales offer low-cost methods to make tools available around the world. The International Test Commission guidelines (ITC, 2017) provide a rigorous framework for researchers to adapt and translate psychological tests. This review used the X-CAGAM (Duke, 2019) a tool developed to operationalize the guidelines to examine how well the selected ASD measures adhere to the ITC guidelines. Fourteen adaptations and translations were included in this review and all scored as "poor quality" on the X-CAGAM. The measures showed strong psychometrics, however lacked a thorough analysis of the effects of culture and language on the items and the methods employed. This review presents and discusses the sample characteristics, psychometrics, and significant other findings related to the measures and their adherence to ITC guidelines. Finally, suggestions for future research and practice are provided.

ACKNOWLEDGEMENTS

My journey in psychology could make for a bildungsroman. I stepped into this field as a naïve 19-year-old and fourteen years later I am completing another chapter in this journey. I have learned that you can never stop learning. Each new placement, professor, colleague, or client teaches me something new and I am always humbled by how much there is still to know. I am happy to have chosen this path because I am assured to never have a dull moment.

As I complete my doctoral chapter, I have numerous people to thank. I want to start by thanking Dr. Mark D. Terjesen, my dissertation supervisor and mentor. He has guided me and encouraged me at every step of this program. I am grateful that he has always pushed me to do better and this dissertation is a result of his persistence and patience with me. I want to thank my committee members, Dr. Moskiwitz and Dr. Mouzakitis for their wisdom and support. I also want to thank the entire faculty and my colleagues in the school psychology department at SJU, especially Dr. Jenifer Mascolo. My program has been full of twists and turns and I would not be here without the support of my faculty and friends. I am grateful to my placement supervisors Dr. Miriam Varghese and Dr. Rebecca Neuhaus for their belief in me.

I am lucky to have the most supportive and loving family who have always encouraged me to follow my dreams. I am because of you. My mother Yamini is my best friend and biggest cheerleader. My brother Mohit and father Sherri have time and again come to my rescue. My New York family Bobby and Murphy, and my Atlanta family Rabi and Glenn, thank you for taking care of me.

My friends, old and new have always been there for me when I needed to talk, vent, or needed a pick me up. I thank my soul sisters, Abhimanyu, Trishna, and Ujjainee for being my pillars of support. My friends Saudamini, Archana Manvi, and

Garima thank you for the support and your faith in me. My friend and sister Chelsea, you have been a partner in crime. And my friends in New York, Vanessa, Persefoni, Harneet, Arjun, and Gulika. I am forever grateful to you.

Finally, I want to thank every colleague, child, individual, teacher, and family I have worked with. I grow with you. I am also filled with joy, gratitude, and hope for my next step and hope to better myself as a professional and a person every day.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Over the past decade, there has been a substantial increase in awareness about autism spectrum disorder (ASD) around the world (Elsabbagh & Hahler, 2015). Regrettably, this increase in awareness and research has been limited to certain sections of the world, resulting in limited public and policy level attention in lower-income and middle-income countries (Elsabbagh et al., 2012). One hypothesis that explains this disparity is that these countries have limited resources and face more urgent priorities related to survival that take precedence over mental health research (Scherzer et al., 2012). As a result, most tools developed to diagnose and help treat ASD, including screening questionnaires, rating scales, and observation schedules, are also limited to higher-income countries and specifically represent the cultural and linguistic characteristics of their populations (Elsabbagh et al., 2012) Furthermore, limited research and data related to ASD is available from lower-income countries (Hossain et al., 2017).

This paper aims to evaluate the existing adaptations and translations of ASD rating scales around the world. ASD rating scales are helpful and cost-effective tools that are used to diagnose ASD and further help with progress in monitoring and adapting intervention. Effective adaptations of ASD rating scales can prove to be efficient in aiding diagnosis and treatment for individuals with ASD from lower-income countries. The purpose of this review is to appraise the psychometric properties of the existing adaptations for their intended populations. The review will report gaps in the adaptation process, shed light on potential difficulties faced by researchers during adaptation of tests, and finally, inform future ASD research around the world about the utility of the available adapted tests. The following literature review provides a brief introduction to ASD, the impact of culture on the presentation

and diagnosis of ASD, the prevalence of ASD around the world, a need for early identification, information about ASD rating scales, and the recommended process for adapting and translating tests.

Autism Spectrum Disorders (ASD)

ASD refers to a group of neurodevelopmental conditions characterized by difficulties in social interaction, communication (including both verbal and non-verbal), and restricted and/or repetitive behaviors (APA, 2013). Typically, ASD distinctly presents itself by the age of three years and is considered a stable diagnosis that lasts the entire lifespan (Rice et al., 2012). Previously the Diagnostic and Statistical Manual (up until DSM-IV-TR) classified ASD as a group of disorders, including autistic disorder (AD), Asperger's Syndrome (AS), Childhood Disintegrative Disorder (CDD), and Pervasive Developmental Disorder not otherwise specified (PDD-NOS) (Matson & Neal, 2010; Tsai & Ghaziuddin, 2014). The changes introduced in the DSM-V (APA, 2013) collapsed these disorders into one overarching category: the diagnosis of Autism Spectrum Disorder.

Autism was first described by Kanner in 1943 based on his observation of 11 children with "a powerful desire for aloneness" and "an obsessive insistence on persistent sameness" (Kanner, 1943, p. 249). He named their condition "early infantile autism," but it was not until about four decades later, in the 1980s, that Autism was recognized as a neurodevelopmental disorder with biological underpinnings (Gillberg & Wahlstrom, 1985; Kim, et al., 2013; Wahlstrom, et al. 1986). As stated previously, an increase in ASD awareness around the world has been limited to approximately the last decade. To elucidate the concentration of research to higher-income countries, a recent research article created a bibliometric profile of scientific research on autism spectrum disorders from 2004-2015 (Sweileh et al.,

2016). The authors reported that from the 48,426 articles they cited, the largest number of articles were from the United States (46.48 %), followed by the United Kingdom (13.14 %), and Canada (5.8 %). No lower-income country appeared in the top 10 countries where the research was conducted. Hence, despite advances in science and research around ASD, awareness and knowledge are limited to certain parts of the world. To Highlight the importance of cultural understanding of ASD, the next section elaborates on what is known about culture and ASD.

Cultural Differences in ASD

The major diagnostic systems followed around the world, including the DSM V and the ICD 10 (World Health Organization, 1994), have largely overlapping presentations of the behavioral symptomology (Doernberg & Hollander, 2016) of ASD. Despite a seeming consensus in the medical domain, there is a recognized lack of literature concerning the cultural understanding of the presentation of ASD (Bernier et al., 2010). Additionally, there is little information on how culture affects the diagnosis of ASD and how parents make treatment decisions surrounding ASD (Mandell & Novak, 2005).

Bernier, Mao, and Yen (2010) present the micro and macro-level cultural influences on diagnosis and treatment choices for ASD. They refer to the variance in prevalence of reporting, as well as the age at diagnosis, to highlight the difference in diagnostic practices across cultures (Bernier, Mao & Yen, 2010). Further, knowledge of the differences in diagnostic and treatment services is essential in understanding ASD through a cultural lens (Bernier, Mao & Yen, 2010). For example, in Taiwan (Lung et al., 2009; Lung et al., 2008), researchers found that the process of diagnosis is much shorter in urban areas as compared to rural areas, and they argue that this finding is indicative of the differences in treatment services available.

There is also a recognition that the leading instruments used to diagnose ASD, for example, the ADOS-2 (Autism Diagnostic Observation Schedule; Lord et al., 2015) and the Autism Diagnostic Interview (Lord, Rutter, & Le Couteur, 1994), are not designed to consider cultural variables; thus, the potential cultural influences the validity and reliability of these measures have not been studied (Bernier, Mao & Yen, 2010). The ADOS-2 (Lord et al., 2015) is one of the few tests translated to and used in 12 different languages; sadly, there is limited research on instruments beyond the western world regarding culturally appropriate or adapted assessment instruments (Bernier, Mao & Yen, 2010). Most Asian languages do not have a term for ASD (Dobson et al., 2001), indicative of the gaps in recognition that may impact diagnosis and subsequent treatment of the disorder. Furthermore, though the ADOS is the gold standard for diagnosing ASD, it is an assessment that takes considerable time to administer and requires extensive and expensive training that may not be viable to adapt for use in lower-income countries.

In addition to variability in diagnostic procedures as a function of culture, cultural beliefs as they relate to the recognition of ASD have been studied. For example, in Asian cultures, avoiding eye contact is largely seen as a sign of respect and may hinder the recognition of early symptoms of ASD (Lian, 1996). Furthermore, it was reported in India that ASD could be recognized earlier due to the cultural emphasis on conforming to social norms; however, parents commonly seek help much later due to cultural beliefs surrounding mental illness, boys speaking later than girls, preference for seeking help from religious healers, and seeking medical intervention before behavioral intervention (Daley, 2004).

Furthermore, there appears to be a lack of training and understanding around ASD among medical professionals around the world, as reported in a study in the

United Kingdom (Shah, 2001). In India, significant variability was found within groups of psychiatrists, psychologists, and pediatricians about the characteristics most necessary for the diagnosis of ASD (Daley, 2004). Parents receiving conflicting information leads to inconsistencies in the understanding of the disorder, and subsequently, its course of treatment (Mandell & Novak, 2005). Adapted tools that are valid, reliable, and standardized through a culturally appropriate lens thus offer opportunities for a more relevant diagnosis in different parts of the world.

Prevalence of ASD Internationally

Research to estimate the global prevalence of ASD demonstrates considerable variability. Most recently, Baio et al. (2018) found that approximately one in 59 children (1.69 %) aged eight years are identified with ASD in the United States. On the global scale, Elsabbagh et al. (2012) reviewed epidemiological surveys and reported a global median ASD prevalence of 0.062%. Other reports about the global prevalence include a systematic review of ASD prevalence by Williams et al. (2006), estimating a prevalence of 0.002%, and more recently, Baxter et al. (2015) estimated a prevalence of 0.076 per 10,000 (0.00076 %) individuals worldwide. The stark variability in the prevalence reported worldwide could be attributed to inconsistency in knowledge, different research methodologies (e.g., inclusion criteria, age of participants), lack of culturally appropriate tools and scales used to screen for and diagnose ASD.

Given the insufficient prevalence data available in Asia, Sun and Allison (2010) conducted a systematic review of prevalence studies of autism from the year 1971 to 2008. They included 17 studies from six Asian countries (excluding South Asia) and reported a prevalence of ASD of 0.0148% since 1980. Further, they reported a higher prevalence rate of 0.0103% among children between 2–6 years old,

and a higher prevalence rate among boys. In addition, Sun and Allison indicated a higher prevalence in urban areas across countries, a higher prevalence in Japan as compared to China, and an increase in prevalence over time across countries (Sun & Allison, 2010). The authors did not include any articles from South Asia as these did not meet their inclusion criteria. Given this lack of reliable data, more research is needed in this area with well-validated measures that are appropriate to different cultural contexts.

More recently, Qiu et al. (2019) conducted a systematic review of the prevalence of ASD in Asia and reported ASD prevalence in East Asia (0.51%; China and Korea) was higher than that in West Asia (0.35%; Iran, Israel, and Lebanon) and South Asia (0.31%; India, Sri Lanka, and Bangladesh). Qiu et al. (2019) suggest the need for a more universal and standardized diagnostic process for ASD as early identification of ASD is linked to the quality of life for ASD children and their families and can inform future programs related to ASD in Asia.

The methodological and cultural differences in recognizing and collecting data may have impacted the difference in prevalence reported in Asia as compared to the prevalence of 1 in 54 (1.6% approx.) indicated in the US (Baio et al., 2018). The studies selected used different criteria for identification of autism, including the Autism Behavior Checklist (ABC; Volkmar et al., 1988), Childhood Autism Rating Scale (CARS, Schopler & Reichler, 1988), the DSM IV criteria (APA, 1994), and the ICD 10 criteria (WHO, 2010). Furthermore, it can be hypothesized that cultural beliefs, social norms, and the absence of an understanding of ASD may contribute to an underreporting of ASD by parents (Hossain et al., 2017), rendering a low prevalence rate. The variance in methodology calls for a need to employ a standardized approach; for example, using an adapted scale that is also normed to the

intended population to provide a clearer picture of the prevalence within and between countries.

Early Identification of ASD: A Need for Adapted Tools for School-Age Children

Recognition of the importance of early identification of symptoms as a means to achieving accurate diagnosis and educational and treatment planning has led to focused efforts to specify symptomatology and early trajectory of Autism at younger ages (Camarata, 2014; Lord & Jones, 2012). Earlier detection of neurodevelopmental disorders leads to earlier intervention and subsequently, better outcomes for individuals (Camarata, 2014). The plasticity of the brain in the first three years of life suggests that interventions at this time will be most effective to teach skills and behaviors (Caramata, 2014).

Ascribing to the view that ASD is pervasive (Koegel & Koegel, 2006), early intervention implies better outcomes relative to intervention delivered later in the developmental period (Camarata, 2014). Comprehensive interventions and behavioral interventions have been found to have better outcomes compared with other treatments or usual practice in terms of IQ and adaptive behaviors (Peters-Scheffer et al., 2011). Landa (2018) reviewed early intervention studies for children with ASD over the past 15 years and concluded that early intervention, especially when mediated by parental involvement in administering the intervention, is extremely beneficial. Bryson et al. (2003) concluded through their review that though early interventions are indicative of positive outcomes for children with ASD, the instruments available pose limitations, and children do not usually receive a diagnosis until the age of 4 years. Recent research in ASD has shown a stable diagnosis can be ascribed as early as 12 months of age (Pierce et al., 2019). These findings point to the wide gaps in the advancements of early diagnosis and early interventions for ASD.

On the other hand, administering psychological tests in early childhood poses a likelihood of false-positive results (Maddox et al., 2017). The early development period is characterized by rapid changes in development, causing difficulty in providing a diagnosis. There are also trans-diagnostic symptoms, common to different disorders, which can lead to a faulty or premature diagnosis (Camarata, 2014). This reinforces the importance of having psychometrically reliable and valid measures to reduce the likelihood of false-positive results. Such a measure will pave the way for early identification of ASD and allow for targeted interventions in the areas of concern presented.

Western, Educated, Industrial, Rich, and Democratic (WEIRD) Nations and Psychological Testing

Rad, Martingano, and Gingees (2018) state that psychology should aim to understand what is universally common in human behavior and how the variables of culture and context influence behavior. However, the science of psychology is primarily based on people who belong to Western, educated, industrialized, rich, and democratic (WEIRD) nations (Ginges, Martingano, & Rad, 2018; Henrich, Heine, & Norenzayan, 2010; Muthukrishna et al., 2020). To demonstrate further challenges with global generalizability of psychology research, Arnett (2016) reported that in the "Journal of Personality and Social Psychology", 67% of the American samples (and 80% of the samples from other countries) consisted primarily of under-graduates taking psychology courses. Hence the results from research studies created using a sample limited to the western world should not be generalized while studying non-WEIRD societies (Muthukrishna et al., 2020).

Similarly, psychological tests are also created within a specific culture and are loaded with cultural and linguistic factors from the WEIRD countries. Henrich and

colleagues (Henrich Heine, & Norenzayan, 2010) reviewed a range of domains often investigated within psychological research and compared WEIRD and non-WEIRD countries. More specifically, they examined the research that measured the constructs of visual perception, fairness, cooperation, spatial reasoning, categorization, and inferential induction, moral reasoning, reasoning styles, self-concepts and related motivations, and the heritability of IQ as reported in WEIRD and non-WEIRD countries. Overall the results from the non-WEIRD countries were statistically seen as outliers, implying that they do not fit into the "average" norms of WEIRD countries. This further suggests that factors outside of the construct being measured (for example, culture and language) have an impact on the scores. In the context of adaptation of ASD rating scales that the present study evaluating, the cultural presentation of ASD might lead to under or over-reporting in the adapted scales as a function for the context in which the research is conducted. While test adaptations pose as effective ways of measuring ASD in children in lower-income countries, the inability to meet guidelines regarding culture and language may undermine the validity of representation of the construct(s) being assessed.

ASD Rating Scales

The leading instruments presently used to diagnose ASD (ADOS-2 & ADI) require considerable time to administer and extensive and expensive training for the test administrator. Hence they do not pose to be viable options to adapt for use around the world, especially for lower-income countries. ASD rating scales, on the other hand, are narrowband rating scales designed to diagnose ASD and help plan treatment. Narrowband rating scales specifically measure a single area of concern, such as ASD or Attention Deficit Hyperactivity Disorder, etc. Rating scales are easy to administer, low cost, and time-efficient methods to help diagnose ASD in a wide

number of participants and hence can be particularly successful when adapted for use in lower-income countries (Lee and Stewart, 2017). Soto et al. (2014) conducted a review of the adaptation process of ASD screening tools around the world. Screening tools are primarily used to monitor for red flags and identify children at risk for ASD (Soto et al., 2014; Szatmari et al. 2003). Soto et al. (2014) reported that the articles that followed or stated a rigorous adaptation process also made more significant changes in the adapted tool to reflect the cultural and linguistic characteristics of the intended population. Hence robust adaptations likely follow a thorough adaptation process and result in nuanced modifications to the original tool.

The current review aims to examine the adaptations of ASD rating scales around the world. ASD rating scales pose many advantages as rating scales provide easily quantifiable results that are normed and standardized. These can assist further in progress monitoring of treatment, measuring treatment outcomes, and with ASD research (Ramsay, Reynolds, & Kamphaus, 2002). Rating scales also pose certain disadvantages. A reporter bias may affect the results, children may show a different presentation in different settings which may not be reflected by a single reporter, and narrowband rating scales do not account for differential diagnosis (Ramsay, Reynolds, & Kamphaus, 2002). Reported bias can also be affected by culture, as certain behaviors may not be reported as problematic for individuals in some cultures. Cultural adaptations can account for the differences in presentation and reporting styles of ASD symptoms. Furthermore, rating scales are effective ways to reach individuals that may require attention.

Adapting Psychological Tests

Developing a test simultaneously in different languages and countries with items that are relevant to the intended population is the least common but most

effective method to attain a test that is valid across cultures (Anderson, et al., 1996; Arafat et al., 2016). However, a majority of instruments are originally developed in English to be subsequently translated for use in other languages (De Beurs, et al., 2005). Adapting a test can be a particularly effective method for lower-income countries as it can significantly cut costs and speed up the process of establishing a psychometrically sound test (Ramsay et al., 2002). However, researchers adapting a test need to reach equivalence between the original and adapted versions, while providing for a culturally and linguistically valid test. Test adaptation includes the process of translation, adaptation, the assessment of reliability, and responsiveness (Duke, 2019). Adaptation moves beyond translation to address the cultural appropriateness of the instrument (Arafat et al., 2016). To address the need to improve testing practices globally, the International Test Commission (ITC) was set up in 1976 (Oakland et al., 2001; Oakland, 2004). In 1992, the ITC began developing a set of guidelines for translating and adapting psychological and educational tests. The International Test Commission guidelines for adapting and translating tests (first edition) were published in 2005 (ITC, 2005) and included 22 principles (ITC, 2017).

With the advancement in statistical methodologies, the ITC instituted a six-person committee to revise its guidelines in 2007. An updated and final version of the guidelines was published in 2017. These include 18 standards organized into six categories to facilitate their use: Pre-condition, test development, confirmation, administration, scoring and interpretation, and documentation (ITC, 2017; Muniz, 2013). The Pre-condition section includes considerations prior to the test adaptation process, including obtaining copyrights for the tool and determining if the tool is appropriate for adaptation (Duke, 2019; ITC, 2017; Muniz et al., 2013). The Test Development section provides information about the transition and adaptation

process, choosing a design, and piloting the adapted measure (Duke, 2019; ITC, 2017; Muniz et al., 2013). The Confirmation section addresses the adaptation's psychometric properties (Duke, 2019; ITC, 2017; Muniz et al., 2013). The Administration section highlights standardization of the testing practices, and the Score Scales and Interpretation section provides recommendations cautioning against misinterpretations of cross-cultural performance differences (Duke, 2019; Muniz et al., 2013; ITC, 2017). Finally, the Documentation section specifies the need for a detailed technical manual/information about the adaptation process (Duke, 2019; ITC, 2017; Muniz et al., 2013; Hambleton, 2005).

To operationalize the ITC guidelines (2017), Duke (2019) developed the Cross-Cultural Adaptation Guideline Adherence Measure (X-CAGAM; See Appendix A). The X-CAGAM is an 18 item rubric that operationalizes the ITC (2017) guidelines for practical use. Details about the tool are elaborated upon in the method section.

Method

Research Questions

The purpose of the present study is to evaluate how well test adaptations and translations of ASD rating scales for school-age children adhere to the International Test Commission (ITC, 2017) guidelines. Adherence to the ITC (2017) guidelines will be evaluated using the X-CAGAM (Cross-Cultural Adaptation Guideline Adherence Measure; Duke 2019), an objective tool designed to measure the same. The following research questions will be explored:

1. How well do test adaptations and translations of ASD rating scales adhere to the ITC (2017) guidelines of test adaptations and translations as reflected by the X-CAGAM?
2. What ITC (2017) guidelines, as reflected by the X-CAGAM, are predominantly followed by researchers while adapting and translating ASD rating scales?
3. What ITC (2017) guidelines, as reflected by the X-CAGAM, are predominantly not followed by researchers while adapting and translating ASD rating scales?

The following hypothesis will be investigated by the researcher

1. As low and middle-income countries do not have resources available to invest in mental health and subsequently test construction and adaptation, it is hypothesized that the test adaptations and translations of ASD rating scales will predominantly score on the X-CAGAM as “Poor Quality” (X-CAGAM Score 0-49 %) as a reflection of adherence to the ITC (2017) guidelines of test adaptations and translations.

2. Since there have been statistical advancements in research, and the ITC has updated its guidelines more recently (ITC, 2017), it is hypothesized that there will be a positive correlation between recent research as measured by year of publication and adherence to the guidelines on the X-CAGAM as indicated by the total X-CAGAM score.
3. The measures from WEIRD countries will score higher than non-WEIRD countries on the X-CAGAM.

Measure

The Cross-Cultural Adaptation Guideline Adherence Measure (X-CAGAM**;** Duke, 2019) is an 18 item rubric developed to be able to use the ITC (2017) guidelines as an objective measure in rating the degree to which translations and adaptations follow these guidelines. Each guideline is given the value of high (2), medium (1), and low (0) quality ratings based on the extent to which researchers adhere to the guideline. The X-CAGAM underwent three phases of development, the initial development, expert analysis, and finally, the expert pilot. The data from the expert pilot stage, as reported by Duke (2019), revealed that the X-CAGAM demonstrated strong construct validity and strong internal consistency; however, it displayed poor inter-rater reliability when used by the sample while scoring a fictitious adaptation study (Duke, 2019). However inter-rater reliability was good among experts using the same fictitious adaptation study. Since the X-CAGAM is the only available measure that operationalizes the ITC guidelines it will be used to evaluate how test adaptations and translations adhere to best practices. Duke (2019) also reported no significant relationships between cross-cultural experience factors and participant ratings, implying that practitioners can use the tool with differing experience and training.

Duke (2019) reported that the items that contain multiple criteria were more problematic. For example, the 10th item, (See Appendix A) addresses the statistical equivalence of the original and adapted measure and includes an appraisal of the construct, method, and item equivalence of the adapted measure. To attain the value of high on the X-CAGAM on this construct, all three equivalencies (construct, method, and item) must be addressed in test adaptation. Duke (2019) found discrepancies in expert scores specifically on such items on the X-CAGAM while scoring fictitious research papers as provided by Duke (2019). These findings will be considered while interpreting the results of this review by specifically looking at all three areas separately and then at the construct as a whole. Any observed difficulties will be discussed with the results. Despite some concerns in the measure, there exist no other measures that specifically evaluate the application of the ITC guidelines.

Eligibility Criteria

For this study, the following selection criteria of measures were followed. Published ASD rating scales that have been translated or adapted for use were selected. To select adaptions/translations that are statistically rigorous, articles published in peer-reviewed journals and student thesis and dissertations that have received an institutional review board (IRB) approval were selected. Finally, only translations or adaptations published after 1994 (after the publication of DSM IV) were selected as the concept of ASD evolved significantly with the publication of the DSM IV.

Selection criteria and Search strategy

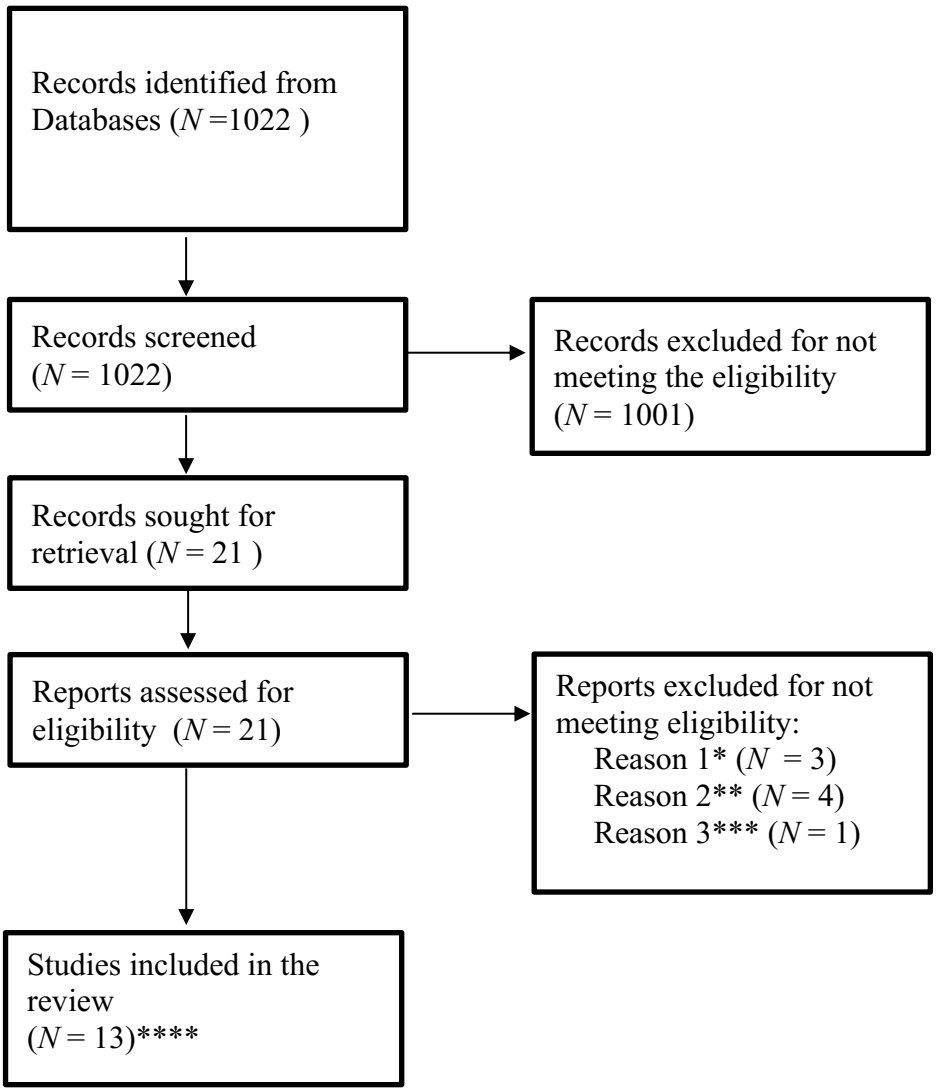
Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, Moher et al., 2009) guidelines were followed while selecting articles. PRISMA guidelines ensure a transparent approach to review and meta-analysis

searches by documenting specific guided steps followed during article searches and hence provide for a replicable review/meta-analysis. Articles and theses/dissertations were identified using electronic databases, including EBSCO, ProQuest, and APA. To search for articles, the following keywords were used: adaptation, translation, ASD rating scale, and autism rating scale. The abstracts of the articles and their method sections were screened to ascertain that they are empirical research studies that are adapting and/or translating ASD rating scales for school-age children and that they meet the eligibility criteria (Appendix B). The process is outlined in greater detail in Figure 1.

After retrieving the screened articles, some were excluded because they did not meet one of the three inclusion criteria. Three retrieved measures were adaptations and/or translations of screening instruments and not ASD rating scales (Reason 1: Guo et al., 2011; Mattila et al., 2012; Yousefi et al., 2015). Four retrieved measures were validation studies of ASD rating scales in different countries and they did not translate or adapt a measure (Reason 2: Nordin et al., 1998; Perry et al., 2005; Tachimori et al., 2003; Tafiadis et al., 2008). One article was deemed irrelevant as the authors used the Autism Rating Scale, which is a scale used to assess the schizophrenic phenotype, for patients with Schizophrenia. Since this review is concerned with ASD scales for ASD diagnosis and treatment planning, this article was excluded (Reason 3: Ballerini et al., 2015).

Figure 1

Screening process for adaptations and translations according to PRISMA Guidelines



** Adaptations and/or translations of screening instruments and not rating scales.*
*** Validation studies of ASD rating scales.*
**** Scale used to assess the schizophrenic phenotype.*
**** one article included two adaptations

Procedure

The researcher identified and screened the research methodology utilized in the translations and adaptations of ASD rating scales using the screening criteria (Appendix B). Articles and dissertations that did not meet the aforementioned eligibility criteria were not included in the analysis. The researcher then obtained permission from the author of the X-CAGAM to use the measure to evaluate the translations and adaptations of ASD rating scales. Since the X-CAGAM displayed poor inter-rater reliability during its development (Duke, 2019) the researcher followed the procedure that Duke used to achieve inter-rater reliability among experts while creating the X-CAGAM. Duke (2019) originally constructed three vignettes for the experts to score. Duke randomly assigned a vignette to the researcher of the present study to score and the researcher's scores for the vignette were compared to the scores intend by Duke and scores provided by the experts during the development of X-CAGAM. The present researcher's scores were in 89% agreement with the experts' scores and 95% agreement with the scores intended by Duke (2019). Hence the researcher proceeded with scoring the adaptations and translation on the X-CAGAM.

It was initially proposed that if there was a difference with the expert's scores or the scores assigned by Duke to the vignette, the researcher would discuss/receive additional training on the X-CAGAM scoring with the author of X-CAGAM, and subsequently score the second and third vignettes to establish reliability in scoring. As the scores were not significantly different, this process was not followed. All three vignettes were not attempted at first to give the further opportunity to establish reliability if it was not established with the first vignette.

Statistical Analysis

For the analysis, descriptive statistics regarding how the adapted and translated ASRS measures score on the overall X-CAGAM, distribution of scores across items on the X-CAGAM, and a correlation of the year of the publication with the score on the XCAGAM are reported in the results section. This will helps answer research questions 1, 2, and 3, and hypothesis 1. Since the ITC guidelines were first introduced in 2005 and subsequently updated in 2017, hypothesis 2 investigating whether more recent studies score better on the X-CAGAM was reported by examining correlations between the X-CAGAM scores the year of the article publication. For the third hypothesis, the researcher conducted T-tests to compare the X-CAGAM scores for WEIRD and non-WEIRD countries.

Results

The findings from the review of translations and adaptations of ASD rating scales are presented here. The first section documents the characteristics of the translations and adaptations reviewed. The second section describes the psychometrics of the adapted and translated measures included in this review. The third section examines the eighteen items on the X-CAGAM and their mean scores across the translations and adaptations. The fourth section describes the correlation between the year of publication and scores on the X-CAGAM. Lastly, the scores on the X-CAGAM of measures from WEIRD and non-WEIRD countries are compared.

Quantitative data from the review was analyzed using SPSS with values rounded to two significant figures. Significance testing was conducted at the $p < .05$ level.

Characteristics of the Translations and Adaptations of measures of ASD

Fourteen translations and adaptations of measures of ASD from thirteen articles and dissertations (one dissertation adapted two measures) met the selection criteria to be included in this review. The selected measures included different versions of five ASD rating scales. The characteristics and standardization data of these original measures are presented in Table 1.

Ten measures qualified as translations and four as translations and adaptations. An adaptation for the purpose of this research relied on the ITC guidelines and is defined as a measure modified in accordance to the culture and language of the country for which the measure is intended (ITC, 2017). A translation, on the other hand, follows either a forward and back translation design or a double translation reconciliation method for the use of the measure in a different language (ITC, 2017).

Table 2 depicts the characteristics of the adapted and/or translated measures including their intended country and language, age ranges, characteristics of the samples, and funding information, among other factors. The translations and adaptations included in this review are selected from ten countries and are in nine different languages. Nine researchers reported the gender breakdown of their samples of which led to approximately 84% male children with ASD in their clinical sample and approximately 66.7% male children in their total sample (including children with ASD, the non-clinical, and children with other diagnosis). Three researchers only report an ASD (clinical) sample, while others report clinical and non-clinical (n = 4) sample, or a clinical and an additional clinical group with a diagnosis other than ASD (n = 4), and the remaining report a clinical, an additional clinical group and a non-clinical sample (n = 3).

Table 1

Psychometrics of original measures reported in this review

Measure (Authors)	Ages (yrs.)	Sample			Reliability		Validity				
		ASD	Non-Clinical	Other	α	r	Discriminative	Convergent	Sensitivity	Specificity	Other
Constantino & Gruber, 2005 (Social Responsiveness Scale, SRS)	4-18	158	287	-	.93 - .97	.85	-	-	.78	.67	-
Gilliam, 2006 (Gilliam Autism Rating Scale-2, GARS-2)	3-22	1107	74	Other Dev. Disabilities 254	.84 - .94	.88	-	ABC .56-.78	1	.87	-
Gilliam, 1995 (Gilliam Autism Rating Scale, GARS)	3-22	1092	-	-	.90 - .93	.88	-	-	-	-	-
Goldstein & Naglieri, 2010 (Autism Spectrum rating Scale, ASRS 2-5 yrs.)	2-5	135	160 (M) 160 (F)	CD 35 DD 41	.97	.97	.90	.80	.89	.91	Kappa .80
Goldstein & Naglieri, 2010 (Autism Spectrum Rating Scale 6-18 yrs.)	6-18	214	480 (M) 480 (F)	ADHD 123 AD/MD 31 CD 39 Other 92	.97	.92	.91	.833	.90	.92	Kappa .80

Measure (Authors)	Ages (yrs.)	Sample			Reliability		Validity				
		ASD	Non-Clinical	Other	α	r	Discriminative	Convergent	Sensitivity	Specificity	Other
Schopler et al., 2010 (Childhood Autism Rating Scale-2, CARS-2 QPC)*	6-18	1034	-	-	-	-	-	-	-	-	-
Schopler et al., 2010 (Childhood Autism Rating Scale – 2 High Functioning, CARS-2-HF, 6-18 yrs.)	6-18	994	-	-	.96	994	.93	ADOS-.77	.88	.86	IRR .95
Schopler et al., 1988 Childhood Autism Rating Scale (CARS)	2+	537	-	-	.94	.90	-	-	-	-	IRR .71
Rutter et al., 2003 (Social Communication Questionnaire, SCQ)	4+	160	-	40 (diagnosis other than ASD)	.87		Non ASD .92 ID .70	-	.85	.75	Rutter et al., 2003 (Social Communication Questionnaire, SCQ)

Note: M, male; F, female subjects; ID, Intellectual Disability; ADHD, Attention Deficit Hyperactivity Disorder; AD, Anxiety Disorder; MD, Mood Disorder; IRR inter-rater reliability; ABC, Autism Behavior Checklist (Krug et al., 1980); ADOS,

** Originally a qualitative measure*

Table 2

Descriptions of the Translations and/or Adaptations and their Scores on the X-CAGAM

Author(s)	Measure	Country (Language)	Funding/ Grants Received	Adaptation &/or Translation	Sample				XCAGAM Score (%)
					Clinical	Non-Clinical	Other Diagnosis	Age Range	
Akoury-Dirani et al., 2013	CARS-2-HF (6-18 yrs.	Lebanon (Lebanese)	YES	Translation & Adaptation	20 (M) 4 (F)	-	ADHD 6(M)	6-18 yrs.	22.22
Al Jabery, 2008	GARS-2	Jordan (Arabic)	NO	Translation	38 (M) 12 (F)	-	ID 36 (M) 14 (F)	3-13 yrs.	41.67
Avcil et al. 2015	SCQ	Turkey (Turkish)	NO	Translation	50	-	ID 50	4-18 yrs.	16.67
Diken et al., 2012	GARS-2	Turkey (Turkish)	NO	Translation	331 (M) 105 (F)	49	ID 144; HI 44	3-21 yrs.	47.22
El Shourbagi, & Abd-El-Fattah, 2019	SCQ	Oman (Arabic)	NO	Translation	25 (M) 20 (F)	23 (M) 22 (F)	-	8-14 yrs.	11.11
Gau et al. 2013	SCQ	Taiwan (Chinese)	YES	Translation	590 (M) 146 (F)	221	-	2-18 yrs.	36.11
Jackson et al., 2013	GARS-2	USA (Spanish)	NO	Translation & Adaptation	77	9	ADHD 9; ID 2; Other 3	3-16 yrs.	27.78

Author(s)	Measure	Country (Language)	Funding/ Grants Received	Adaptation and /or Translation	Sample				XCAGAM Score (%)
					Clinical	Non-Clinical	Other Diagnosis	Age Range	
Li, 2012	CARS-2 QPC	USA (Chinese)	NO	Translation	20	-	-	2-17 yrs.	16.67
Li, 2012	GARS	USA (Chinese)	NO	Translation	20	-	-	2-17 yrs.	16.67
Pereira et al., 2008	CARS	Brazil (Portuguese)	NO	Translation	44 (M) 16 (F)	-	-	3-17 yrs.	19.44
Samadi & McConkey, 2014	GARS	Iran (Persian)	NO	Translation	358 (M) 84 (F)	67 (M) 35 (F)	ID 81 (M) 31 (F)	3-22 yrs.	30.56
Wang et al., 2012	SRS	Taiwan (Mandarin)	NO	Translation	307	32	ADHD 51; DD 51; Combined 33	4-6 yrs.	27.78
Zhou et al., 2019	ASRS (2-5 yrs.)	China (Chinese)	YES	Translation & Adaptation	146 (M) 21 (F)	908 (M) 841 (F)	-	2-5 yrs.	41.67
Zhou et al., 2017	ASRS (6-18 yrs.)	China (Chinese)	YES	Translation & Adaptation	161 (M) 26 (F)	752 (M) 713 (F)	-	6-18 yrs.	44.4

Note: M, male; F, female subjects; ID, Intellectual Disability; HI, Hearing Impairment; ADHD, Attention Deficit Hyperactivity Disorder; DD, Developmental Delay

Psychometrics of the Translations and Adaptations

Table 3 lists the psychometrics from the translations and adaptations. The central tendencies for the psychometrics across all adaptations are not reported as the researchers have adapted different scales and it was decided that a central tendency value will not be meaningful in discussing these measures. Furthermore, the samples and the psychometrics reported are not consistent across the measures. That is, as researchers adapted or translated different ASD rating scales, they cannot be directly compared as it relates to measures of central tendency and as a result, the ranges of values are reported.

Reliability

All the researchers reported reliability as indicated by Cronbach's alpha (α), which is a measure of the internal consistency of a measure (Bruin, 2006). The α values reported were between .73 and .96 indicating a high internal consistency. Eight translations and adaptations reported test-retest reliability by correlating the scores across the two administrations with administrations ranging between 2-4 weeks. The test-retest reliability was reported as high (scores between .77 and .98). Four measures reported intra-class correlation values (ICC) between .77 and .97, which indicate a high correlation between subscales and the total scale (Bruin, 2006). One adaptation also reported a high kappa value ($k = .90$; Pereira et al., 2008), which is a measure of test-retest reliability. In this study, fifty ($N = 60$) parents were re-administered the test after a maximum of four weeks and found a high test-retest reliability.

Sensitivity and Specificity

Seven researchers reported sensitivity and specificity values. Sensitivity is an indicator of the proportion of positives (children diagnosed with ASD) that are correctly identified by the measure and ranged between .66 to 1. Specificity is an

indicator of the proportion of negatives (children not diagnosed with ASD) that are correctly identified by the measure (Altman & Bland, 1994) and ranged between .82 and 1.

Fit Indices

Three measures reported the root mean square error of approximation (RMSEA), a statistic that accounts for latent model structures. A lower value indicates a better fit with the hypothesized model (Hooper et al., 2008). Zhou et al. (2019) studied the fit indices and found a two-factor structure (social communication and unusual behaviors) and Zhou et al. (2017) studied the fit indices for the three-factor structure (social communication, social responsiveness, and unusual behaviors) of the ASRS. Gau et al. (2013) studied the fit indices for the three-factor structure of the SCQ, and Diken et al. (2012) studied the fit indices for the three-factor structure (social interaction, repetitive behavior, and communication) for the ASRS. The RMSEA values ranged between .04 and .08 which indicated a good fit for the three-factor model. One study (Gau, et al. 2013) reported standardized root mean squared residual (*SMSR* = .06) which is a fit index like the RMSEA (Asparouhov & Muthén, 2018). Three translations and adaptations also reported the comparative fit index (CFI; Diken et al., 2012**;** Gau, et al. 2013; Wang et al., 2012). Two of the three (Diken et al., 2012**;** Gau, et al. 2013) had also reported the RMSEA. The CFI, similar to the RMSEA, is used to account for a small sample size by comparing the discrepancy of the obtained psychometrics of a measure to that of an optimal hypothetical model (Gatignon, 2010). A larger CFI value indicates a better fit with the hypothesized model. For this review, the values ranged between .89 and .98 indicating a good fit. Finally, one measure reported high goodness of fit index (*GFI* = .92; Gau, et al. 2013) which is a fit index like the CFI (Gatignon, 2010).

Factor Analysis

Zhou, et al. (2019; Chinese adaptation of the ASRS: 2-5 yrs.) conducted an exploratory factor analysis and found a two-factor structure. They excluded 8 items based on the factor analysis. Zhou et al. (2019) also reported the TLI (Tucker Lewis Index), a fit index similar to the CFI (*TLI*= .80). With an older age group, Zhou and colleagues (Zhou et al., 2017; Chinese adaptation of the ASRS 6-18 yrs.) conducted an exploratory factor analysis, which suggested a modification of 12 items and a subsequent confirmatory factor analysis revealed a stronger two-factor structure.

Validity

Information about the validity of the measures was reported in the form of convergent and discriminant validity. Convergent validity compares scores of a measure to another measure that assesses the same construct (Krabbe, 2016). Two sets of researchers reported convergent validity by correlating measures to the Autism behavior checklist(Al Jabery, 2008; El Shourbagi & Abd-El-Fattah, 2019)**.** The researcher reported a high and a moderate convergent validity respectively. Another researcher qualitatively reported good content and criterion validity. Discriminant validity compares scores of a measure to another measure that assesses a distinct construct (Krabbe, 2016). Seven measures reported good discriminant validity using Cohen's d or mean differences between groups.

Table 3

Psychometrics of the translations and Adaptations

		Reliability							Validity		
Author	Measure	α	*r*	ICC	Sens.	Spec.	RMSEA	CFI	Convergent	Discriminant	Other Statistics
Akoury-Dirani et al., 2013	CARS-2-HF	.92	.99	.97	1.00	1.00	-	-	-	-	-
Avcil et al. 2015	SCQ	.73-.83	-	.87-.96	.94	.84	-	-	-	-	-
Diken et al., 2012	GARS-2	.91	.94	-	-	-	.07	.89	-	Group Comparison	Diken et al., 2012
El Shourbagi & Abd-El-Fattah, 2019	GARS-2	.89	.93	-	-	-	-	-	*r*(ABC) = .52	Mean difference (between groups) = 1.1	-
Gau, et al. 2013	SCQ	≥.73	.77	.77-.78	-	-	.03	.98	-	-	Concurrent validity *r* = .65 SMSR = .06 GFI = .92
Jackson et al., 2013	GARS-2	.96	.98	-	-	-	-	-	-	*F* = 17.68	-

Author	Measure	Reliability			Sens.	Spec.	RMSEA	CFI	Validity		Other Statistics
		α	*r*	ICC					Concurrent	Discriminant	
Li, 2012	CARS-2 QPC	.91	-	-	-	-	-	-	-	-	-
Li, 2012	GARS	.89	-	-	-	-	-	-	-	-	
Pereira et al., 2008	CARS	.82	-	-	-	-	-	-	-	*r* (GAF) = -.75	-
Samadi & McConkey, 2014	GARS	.95	.99	-	.96	1	-	-	-	Percentage difference between groups	-
Wang et al., 2012	SRS	.73-.91	-	.77-.78	.66	.90	-	.98	-	Cohen's *d* >1.60	-
Zhou, et al., 2019	ASRS 2-5 yrs.	.91	-	-	.66	.85	.06	-	-	Cohen's *d* 2.27	TLI = .80
Zhou, et al., 2017	ASRS 6-18 yrs.	.93	-	-	.94	.82	.04 -.08	-	-	-	-

Note: N refers to number of subjects; ICC, intra-class correlation; CFI, comparative fit index; α, Chronbach's alpha; r, Pearson's correlation; RMSEA, root mean square error of approximation; GAF, Global Assessment of Functioning; ABC, Autism Behavior Checklist (Krug et al., 1980); GFI, goodness of fit Index; SMSR, Standardized Root Mean Squared Residual

Scores on the XCAGAM

Information about the selected articles and their scores on the Cross-Cultural Adaptation Guideline Adherence Measure (X-CAGAM**;** Duke, 2019) are presented in Table 3. Hypothesis 1 was supported with the results indicating that all fourteen translations and adaptations reviewed scored as "Poor Quality" on the X-CAGAM. Of note, six adaptations were from middle-income countries as defined by the World Bank (2020). Only three adaptations/translations were from a WEIRD country (USA) and were intended for use with culturally and linguistically diverse children living in the USA.

The present study examined whether ITC (2017) guidelines, as reflected by the X-CAGAM, are predominantly followed by researchers while adapting and translating ASD rating scales. Table 4 lists the guidelines and their corresponding minimum, maximum, mean scores, and standard deviations across the translations and adaptations.

To address the research questions regarding what guidelines are most and least followed, the guidelines with mean scores higher than 1 (moderate quality) or less than .5 (low quality) are reported here. Only two of the ITC guidelines as measured by the X-CAGAM obtained a mean score of more than 1 across the translations and adaptations. The first guideline is TD-1 (test development guideline 1; ITC, 2017), which addresses the choice of experts in the adaptation/translation process. Fifty percent of the measures scored as "good quality" and forty-three percent scored as "fair quality" for choosing experts well versed in the culture of the country adapting the test (M= 1.42*;* SD = .65). The second guideline, C-3 (confirmation guideline 3), addresses support for the norms, reliability, and validity of the adaptation/translation (ITC, 2017). Fifty percent of the measures scored as "fair quality" and thirty-six

percent scored as “good quality” as the researchers supported the measures with reliability and validity findings (*M*= 1.2*; SD* =.7).

Five guidelines obtained a mean score of less than .5 (poor quality) across the translations and adaptations. TD-4 (test developmental guideline 4) relates to the evidence that the item formats are suitable for the intended population. None of the translations and adaptations referred to the item/test format. TD-5 (test developmental guideline 5) refers to the use of a pilot test before the adaptation. Only three translations and adaptations attempted a pilot study (*M* = .29; *SD* =.29). C-4 (confirmation guideline 4) addresses the use of an equating design to link scores across different versions (using a bilingual group design, or a matched monolingual design, or a monolingual group design (Duke, 2019) of the test. Seventy percent of the translations and adaptations did not attempt the use of an equating design (*M* = .43; *SD* =.76). The A-2 (administration 2) guideline addresses whether standardizing testing instructions are provided by the adaptions and the eighty-six percent failed to indicate that here (*M* = .14; *SD* =.36).

Finally, for the two documentation guidelines (D-1 & D-2), all the translations and adaptations scored zero as there are no technical documents published for any of the translations and adaptations other than the articles presenting the findings.

Table 4

X-CAGAM items and Descriptive Statistics Scores from Translations and Adaptations

Item	Description	Minimum Score (*N*)	Maximum Score (*N*)	Mean (SD)
Precondition (PC)				
PC1	Obtain the necessary permission from the holder of the intellectual property rights relating to the test before carrying out any adaptation (ITC, 2017).	0 (5)	2 (8)	1.2 (.97)
PC2	Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the populations of interest is sufficient for the intended use (or uses) of the scores (ITC, 2017).	0 (6)	2 (2)	.71 (.73)
PC3	Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest (ITC, 2017).	0 (12)	1 (2)	.14 (.36)
Test Development Guidelines (TD)				
TD1	Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise (ITC, 2017).	0 (1)	2 (7)	1.42 (.65)
TD-2	Use appropriate translation designs and procedures to maximize the suitability of the test adaptation in the intended populations (ITC, 2017).	0 (4)	2 (5)	1 (.78)
TD-3	Provide evidence that the test instructions and item content have similar meaning for all intended populations (ITC, 2017)	0 (3)	2 (4)	.92 (.73)

Item	Description	Minimum Score	Maximum Score	Mean
TD-4	Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations (ITC, 2017).	0	0	0
TD-5	Collect pilot data on the adapted test to enable item analysis, reliability assessment and small-scale validity studies so that any necessary revisions to the adapted test can be made (ITC, 2017).	0 (11)	2 (1)	.29 (.6)
Confirmation Guidelines				
C-1	Select sample with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses (ITC, 2017).	0 (9)	2 (5)	.71 (.99)
C-2	Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations (ITC, 2017).	0 (8)	2 (1)	.5 (.65)
C-3	Provide evidence supporting the norms, reliability and validity of the adapted version of the test in the intended populations (ITC, 2017).	0 (2)	2 (5)	1.2 (.7)
C-4	Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test (ITC, 2017).	0 (10)	2 (2)	.43 (.76)

Item	Description	Minimum Score	Maximum Score	Mean
Administration guidelines				
A-1	Prepare administration materials and instructions to minimize any culture- and language-related problems that are caused by administration procedures and response modes that can affect the validity of the inferences drawn from the scores (ITC, 2017).	0 (8)	2 (1)	.5 (.65)
A-2	Specify testing conditions that should be followed closely in all populations of interest (ITC, 2017).	0 (10)	2 (2)	.14 (.36)
Score Scales and Interpretation Guidelines				
SSI	Interpret any group score differences with reference to all relevant available information (ITC, 2017).	0 (3)	2 (10)	.5 (.85)
SSI-2	Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported (ITC, 2017).	0 (8)	2 (2)	.57 (.75)
Documentation Guidelines				
Doc-1	Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population (ITC, 2017).	0	0	0
Doc-2	Provide documentation for test users that will support good practice in the use of an adapted test with people in the context of the new population (ITC, 2017).	0	0	0

X-CAGAM and WEIRD Countries

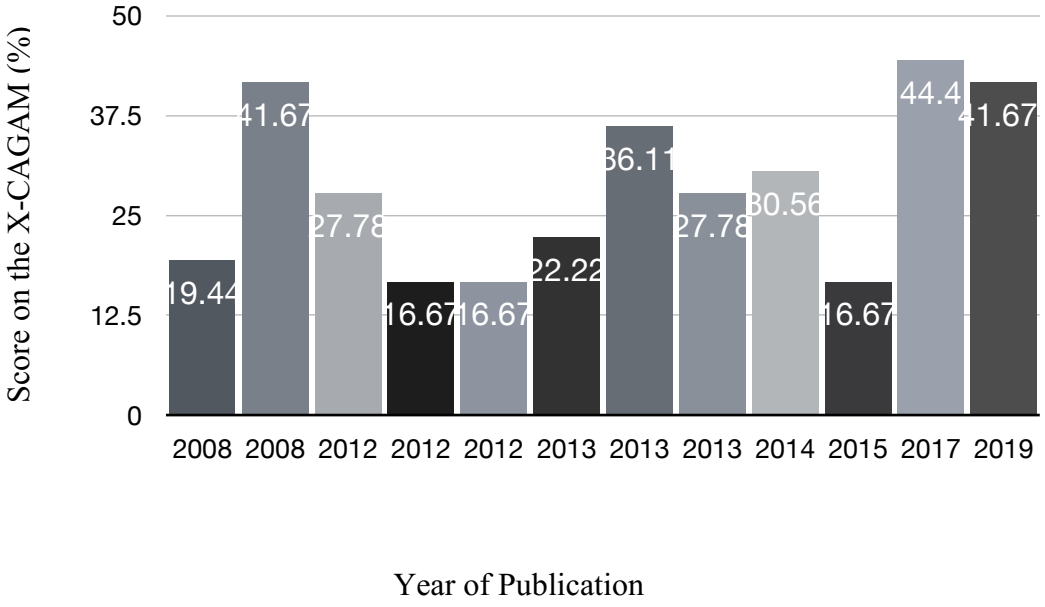
WEIRD countries were hypothesized to score better on the X-CAGAM than Non-WEIRD countries for the adapted tests. Three translations and adaptations included in this review were from WEIRD countries (USA), and eleven were from non-WEIRD countries (China, Oman, Turkey, Iran, Jordan, Lebanon, Taiwan, and Brazil). An independent sample T-test was conducted to compare scores on the X-CAGAM for WEIRD (n = 3; *mean*= 20.37), versus non-WEIRD (n = 11; *mean* = 20.45) countries. Results indicated no significant difference between the two $(T(14) = 2.43, p = .15)$, and as such, this hypothesis was not supported. It is possible that the correlation is not significant due to the small sample size (N = 14) (Hewitt et al., 2008). This is elaborated on further in the discussion.

X-CAGAM and Year of Publication

Hypothesis 3 predicted a positive correlation between recent research as measured by the year of publication and adherence to the guidelines on the X-CAGAM as indicated by the total X-CAGAM score. The selected adaptations and translations were conducted between 2008 and 2019 (*Median* = 2013). A Pearson correlation was conducted between the year of publication of the articles (adaptation/translation) and the total X-CAGAM score. The correlation was not significant (r (*14*) = .39, p = .17), and as such the hypothesis was not supported. Similar to the previous hypothesis, it is possible that the correlation is not significant due to the small sample size (N = 14) (Hewitt et al., 2008). Figure 2 depicts a visual of the year of publication and the score of X-CAGAM. Only one study indicated a score above 40% on the X-CAGAM prior to 2016, however both the studies after 2016 have scores higher than 40%. This could be indicative of a trend of better adaptations, however this can not be concluded due to a low number of studies.

Figure 2

A visual analysis of the year of publication and the corresponding score on the X-CAGAM



Discussion

Ideally, psychological measures would be developed in different countries independently to reflect culture and language variables (Sprangers, 1993; Van Widenfeld et al, 2005); however cross-cultural translations and adaptations of ASD assessments are a great way to apply existing measures to practice in lower and middle-income countries (Duke, 2019; Oakland, 2004). The current review evaluates the existing translations and adaptations of ASD rating scales around the world, intended for school-age children, using the X-CAGAM (Duke, 2019). In this chapter, the characteristics of the selected measures are discussed followed by significant findings of this review and consideration of these adaptations as they relate to the ITC guidelines (ITC, 2017). Following this chapter, the limitations of the study are presented, and finally, recommendations and future directions for translations and adaptations of ASD rating scales are outlined.

Characteristics of the Translations and Adaptations

Sample

Only thirteen studies were selected for this review. The low number of studies may be attributed to the limited information regarding ASD from low and middle income countries (Samms-Vaughan, 2014). Within the selected studies, ten measures included a non-ASD sample (non-clinical or sample of children with other developmental disabilities) and the age ranges replicated those of the original measures.

The sample sizes of selected measures ranged from 20 to 2233. Small sample sizes negatively affect the power and generalizability of the research findings (Tabachnick & Fidell, 2013). Furthermore, research in low or middle-income countries income countries is competitive, restrictive, and poorly funded (Acharya &

Pathak, 2019; Ciocca & Delgado, 2017). Large and diverse samples representing the population of the countries the measure is intended for would be ideal while adapting a test (ITC, 2017) however, researchers don't appear to have the resources to afford them.

Psychometrics

The psychometrics (reliability and validity) for the selected studies were strong across the translations and adaptations. However, only two studies conducted factor analysis and four studies reported fit characteristics (CFI, TLI, RMSEA, etc.). Two measures that reported a convergent validity compared scores on the adapted measures with the Autism Behavior Checklist or the DSM IV which represent the construct of ASD in WEIRD countries. Hence it can not be ascertained that the convergent validity results are valid.

Sass (2011) emphasized the importance of measurement invariance while comparing groups, including when researchers translate and adapt tests across cultures. Measures reliability and validity alone can not determine nonequivalence of a construct or items between diverse groups, and some variables such as ethnicity often go unaccounted for (Hancock et al., 2000; McDonald, et al., 2002; Sass, 2011). Testing for invariance can help create or modify instruments that account for group differences (Heggestaad et al., 2019; Sass, 2011). This is further illuminated by the following validation study conducted in the USA. McClain et al. (2020) examined the original ASRS (Goldstein & Naglieri, 2010) in a diverse (White, Black, and Latinx children) non-ASD sample (6-18 yrs.; $N = 405$). They reported a high internal consistency, test-retest reliability, and inter-rater reliability across their sample. They further conducted a confirmatory factor analysis (based on the three-factor structure of the original scale) and indicated a good fit for the White subgroup, a worse fit for

the Black, and Latinx subgroups (McClain et al., 2020). Furthermore, two items (one regarding showing emotion and another focusing on following directions), did not load significantly for the Black subgroup and loaded significantly but in the opposite direction for the Latinx subgroup (McClain et al., 2020).

These findings suggest that though the overall psychometrics of the scales were good, an in-depth inquiry into the factor structure revealed that the items did not function equally across diverse groups. Hence, the strong psychometrics (reliability, and validity, etc.) reported for the selected measures in this review should not be taken at face value in absence of further statistical investigation such as conducting exploratory and confirmatory factor analysis.

Gender

In the ASD literature, the prevalence of ASD is reported higher for biological males, with a male-to-female ratio of 4.5:1 (Christensen et al., 2018; Feri et al., 2018). However, there is growing awareness that ASD in biological females manifests differently, and hence females with ASD may be underreported or underrepresented (Mandy & Lai, 2017; Loomes et al., 2017). There is also evidence that repetitive and stereotyped behaviors are more prevalent in males (Ferri et al., 2018; May et al., 2014). The adaptations and translations appear to be following the WEIRD countries' understanding of ASD with a high prevalence in males based on the select samples as authors that reported gender characteristics included approximately eighty present males in their sample. This is further addressed in the future directions section.

Funding

Four selected measures received funding. Three of these four studies have the largest sample sizes among the selected studies. Adapting translating tests is an expensive process (Stansfield, 2013). All the measures selected in this review are

published by international publishing companies (Pearson and Western Psychological services) that charge per protocol used burdening researchers from low and middle income countries that have limited resources to start with. Furthermore the guidelines recommend for pilot tests, multiple experts for the process of translation and adaptation, added to the cost of time and space for conducting the assessments (ITC, 2017). Hence costs appear to be an important factor impeding adherence to the guidelines.

X-CAGAM: Significant Findings

Translations and adaptations ought to be developed and modified with statistical rigor, as a measure normed in one country may exclude cultural and linguistic factors of the intended population (Duke, 2019; Chia, 2012; Lynn & Vanhanen, 2012). The fourteen selected translations and adaptations overall scored in the "Poor Quality" range on the X-CAGAM which is consistent with what was hypothesized. Although adapting tests is a cost-effective way to make a test available (Lee & Stewart, 2017), low and middle-income countries that are likely to adapt existing tests might still have limited resources. This is evident in this review as only four selected measures reported receiving funding for their work and had among the larger sample sizes making the findings more generalizable in comparison to the measures that did not receive funding. While funding for the translations/adaptations did not affect the psychometrics of the measures, the mean score of the adaptations and translations receiving funding was higher on the X-CAGAM (*mean* = 36.1) as compared to those that did not receive funding (*mean* = 25.55) implying the availability of resources had an impact on the quality of the translations and adaptations.

In line with the ITC (2017) guidelines, Borsa et al. (2012) emphasize that an adapted test should be equivalent in terms of semantics, fit the intended culture and exhibit satisfactory psychometric properties. To address the "cultural fit" the measures also need to go beyond reporting reliability and validity measures to confirm equivalence (Borsa et al., 2012; Hambleton, 2005). The ITC guidelines recommend an extensive translation process followed by statistical procedures such as item analysis and equivalency testing in addition to measuring the psychometric properties of the adapted test including reliability (ITC, 2017).

Four selected measures reported fit indices and only two conducted an exploratory factor analysis. These analysis help determine that the same construct is measured across populations (Reise et al., 1993). However, none of the studies used an equating design to confirm the validity of the measure. Borsa and colleagues (Borsa et al., 2012) recommend analyses like the multi-group confirmatory factor analysis, differential item functioning, and multidimensional scaling to compare the measures factor structure between subgroups of the population and confirm whether the items measure the same construct within different groups. Such analysis can help determine the validity of a measure; or the need to modify a measure for use in different cultures (Milfont & Fischer, 2010; Sireci et al., 2005).

Another important consideration in test adaptation is the appropriateness of the methods used, which includes test procedure, instructions, test format, and example items for the intended culture (Hambleton, 2005). Rating scales and forced-choice formats are not as commonly used globally (Hambleton, 2005) as compared to the countries where these measures are originally developed. The selected measures described the methodology (TD-4), however did not present an accompanying technical or administration manual. Publishing manuals is an expensive process and

depends on risks and costs publishing companies are willing to take (Oakland, 2004). Furthermore, it is possible that researchers made technical manuals available locally in their countries however, did not report them in the articles due to word limits and restrictions of the journals where these adaptations and translations are published. Based on the ITC guidelines (ITC, 2017), the X-CAGAM recommends that researchers use surveys, cognitive interviewing, or try-out studies to assess familiarity with the format and instructions of the measure. This criteria was also not met by any researcher in this review (TD-5).

Considering the importance of culture and language on the assessment and diagnosis of ASD along with the evidence of how constructs, as measured in WEIRD countries, present differently in non-WEIRD countries (Henrich, et al., 2010; Muthukrishna et al., 2020), it is unfortunate that researchers have not given more importance to making the adapted measure more culturally appropriate. It appears the costs of the process and the application of statistically rigorous procedures are missing from the selected measures.

X-CAGAM and WEIRD Nations

The hypothesis comparing WEIRD and non-WEIRD countries on X-CAGAM for the adapted tests was not supported. The measures from WEIRD countries were adapted for use in the USA (Chinese versions of the GARS 2 and CARS QPC 2 (Li, 2002); and the Spanish version of the GARS 2 (Jackson et al. 2013). These three adaptations were conducted in marginalized communities and were small-scale studies conducted as a part of doctoral dissertations. All three scored in the “Poor” category on the X-CAGAM.

Fitzpatrick and Kind (2017) conducted a review in the USA to examine disparities in ASD health and health system quality and reported racial disparities in

access to services. Further evidence of racial disparities was reported by the centers for Disease Control Autism and Developmental Disabilities Monitoring (AADM, 2018); estimating the prevalence of ASD in White children in the US as 7% higher than Black children and 22% higher than ASD in Hispanic children. These findings indicate a lack of resources available to disadvantaged groups (Hill et al., 2015) in the WEIRD nations as well. It is noted that though the three studies were conducted in a WEIRD country, their samples did not represent a WEIRD population.

Six of the translations and adaptations were conducted in Middle Eastern countries. Mental health and developmental disabilities are neglected in the Gulf region (Osman & Afifi, 2010). Furthermore, the state of ASD research and public health initiatives in low and middle-income countries in the Middle East are negligible (Osman & Afifi, 2010). Given the state of ASD research and practice in Middle Eastern countries, it is encouraging that translations and adaptations from Middle Eastern countries comprised 43% of the measures included in this review. Further, given the lack of resources, the reviewed studies are a step forward in the direction of ASD diagnosis and research.

Four of the translation and adaptation studies were from East Asia (China and Taiwan). In China, there is less awareness about ASD among the general population but advancement in ASD is gradually increasing (Wang et al., 2012). Of note, for the translations and adaptations of the two ASRS versions in China, the authors modified the measures to better fit their intended population. Though the measures did not score high on the X-GAGAM, it appears that researchers are making advances in better ASD diagnosis and treatment.

ITC Guidelines

The hypothesis regarding X-CAGAM and the year of publication was not supported and the scores on the X-CAGAM show a wide gap between the guidelines and what is practiced. Rios and Sireci (2015) conducted a systematic review of published articles to study if the publication of the guidelines has improved the quality of adapted tests. They found that 93.4 % of their selected studies had been translated from English and the guidelines put forth by the ITC have not been practiced in the majority of the selected adaptations. This appears to be consistent with this review.

ASD assessment: Current practices

It is also important to consider tools and techniques currently being used to diagnose ASD in middle and lower-income countries. For example, in Arabic countries ASD is usually diagnosed by pediatricians using western criteria or by trained professionals using the CARS or the ADI (Hussein & Taha, 2013). The CARS and the ADI are both developed in WEIRD countries and do not provide valid results. In China, ASD assessment involves screenings, examination and an MRI (Wang et al., 2019). Recently, two Indian scales were developed to aid the diagnosis of Autism in India; namely the Indian Scale for Assessment of Autism (ISAA; Ministry of Social Justice and Empowerment, Government of India, 2008) and the INCLEN Diagnostic Tool for Autism Spectrum Disorder (INDT- ASD; Juneja et al., 2014). The ISAA has been criticized for including vague items and not providing subscale scores to target treatment (Dalwai et al., 2017) and the INCLEAN is based directly off the DSM-IV-TR which is based on the WEIRD understanding of ASD. The current global practices of diagnosing ASD do not appear robust and test adaptations as presented by this review are not valid for the

In conclusion, based on the results on the X-CAGAM and a review of the measures, it is recommended that adaptations and translations require further evaluations in terms of their appropriateness for their intended population. Since more robust measures may not be available in some countries, it is recommended that researchers and clinicians using these measures provide explanations and clarifications while reporting results during ASD testing or research for items that are not appropriate for the intended populations and are cautious while interpreting scores obtained on these measures.

Limitations

The current study evaluates translations and adaptations of ASD rating scales around the world using the X-CAGAM (Duke, 2019). Some limitations arise from the scope of this review and some from the X-CAGAM. The selected measures belonged to ten different countries and is a small estimate in terms of global cultures. Many studies may have not been selected during the search as they were not from peer-reviewed journals or may not be accessible in English, limiting the global scope of this review.

Since the world has seen an increase in research concerning ASD over the past decade (Elsabbagh & Hahler, 2015), it is surprising that the number of translations and adaptations is limited. This review has given us an insight into how ASD rating scales adhere to the ITC guidelines. Due to the limited number of translations and adaptations that met the selection criteria, it is difficult to make conclusions about the adaptation process in the field as a whole. Furthermore, there were only three translations and adaptations from WEIRD countries and these adaptations represented marginalized groups within the WEIRD country. Thus it is not feasible to make distinctions between WEIRD and non-WEIRD countries based on the findings of this

review. Though the measures belonged to diverse countries, the sample sizes were relatively small. These samples may not be representative of the diversity within the population of these countries and conclusions about the appropriateness of the measures can't be ascertained based on the publications reviewed.

Some limitations also arise pertaining to the use of the X-CAGAM for evaluating these adaptations/translations. The X-CAGAM is a recently developed tool that has not been used to review adaptations prior to this study. The X-CAGAM assigns equal weightage to scores given for each of the guidelines. This may not give us a true picture of the quality of the adaptation. For example, the guidelines regarding obtaining copyright and regarding technical documents are scored between 0 and 2 and the guideline regarding meeting three types of equivalence (construct, method and item) is also scored between 0 and 2. While both are important, it is argued that the guideline regarding equivalence could be broken up into three or have more score points assigned since it measures the quality of the adaptation unlike the guideline regarding copyright and regarding technical documents.

Finally, certain guidelines for example PC-1 (obtaining copyright of the measure) may not have been met by many articles in our review due to the lack of reporting by the authors. That is, they may have done it but just not reported it for space limitations. Furthermore, we argue that the guidelines regarding cultural and linguistic factors in the adaptation, administration, and interpretation should be given more weight in the X-CAGAM scoring to assess the quality of adaptations. Despite the criticism, the X-CAGAM was the only tool available at the time of conducting this review that operationalizes the ITC (2017) guidelines and as such was used despite its limitations.

Future Directions

The accurate understanding and diagnosis of ASD in school-age children can assist with early interventions and more positive outcomes for children with ASD (Camarata, 2014; Lord & Jones, 2012). This review gives us an insight into the quality of translations and adaptations of ASD rating scales. It is important to note that though the translations and adaptations have their weaknesses, they do attempt to make early identification and diagnoses of ASD more feasible in their communities and are an important first step towards creating awareness of ASD around the world.

Given that ITC guidelines are not followed as intended, the ITC committee could focus on how to aid the implementation of these guidelines for researchers around the world. The X-CAGAM has taken the step to operationalize the guidelines and the next step would be for the international testing community to provide academic support and find ways of providing financial support to researchers to follow these guidelines, especially in lower and middle-income countries.

As a next step, it is recommended that researchers conduct exploratory and confirmatory factor analyses, have bigger and more representative sample sizes, and use equating designs to review their cultural and linguistic appropriateness. We recommended that researchers study how ASD presents differently within their countries and across diverse groups within their countries. Some qualitative and focused group studies may help inform such research. Furthermore, researchers need to be cognizant about the health disparities that exist for marginalized groups and include such groups while planning larger-scale studies.

Finally, though understanding the gender differences in ASD is a relatively recent area of research, how gender differences of roles and expectations affect the identification of ASD across cultures, especially societies like the Middle East that

view gender roles and expectations differently (Robins & Thomas, 2018) are important next steps.

Implications for the Practice of School Psychology

For the practice of school psychology, this review addresses and emphasizes the importance of cultural and linguistic differences in the presentation and assessment of ASD. The adapted tools that are available for use require more investigation into validity across culture and language, which in turn has many implications for the practice of school psychology. The present tools are not sensitive to diversity around the world and sadly, neither within WEIRD countries. Hence, the current rating scales may not provide an accurate diagnosis of ASD in school-age children from culturally and linguistically diverse and marginalized backgrounds. Furthermore, there is also a lack of scientific investigation into how ASD presents across cultures and countries (Mandell & Novak, 2005; Tincani et al., 2009).

It would be important for practitioners to use caution against under or over diagnosing ASD in children belonging to diverse backgrounds. Since culturally and linguistically sensitive rating scales are not available, practitioners should enquire into such differences during assessing for ASD and provide diagnosis and recommendations keeping these differences in mind. Furthermore, practitioners within WEIRD Countries should also consider that a tool created or adapted for use in their countries may not represent the diversity in the population.

Practitioners from non-WEIRD countries should use the adapted tools as they might be the only tools available to them, however, it is recommended that practitioners take contextual factors of the child into consideration while using these tools. Psychologists engaged in practice can also contribute to research using practice-based studies to contribute to the knowledge and testing of ASD. ASD rating scales are a great way to diagnose and document progress (Payakachat et al., 2012),

however, ecological factors need to be addressed while providing an ASD diagnosis and treatment recommendations (Bernier et al., 2010; Hock & Ahmedani, 2012).

APPENDICES

APPENDIX A

X-CAGAM (Duke, 2019)

The following rubric is designed to allow the user to assess a test developer's adherence to each of the ITC (2017) Guidelines for Translating and Adapting Tests. Please read each guideline and its operational definition carefully. Next, locate the portion(s) of the text describing each guideline within the vignette. Select the rating that you believe best captures the quality of the evidence describing adherence to each guideline, with 2 representing good quality, 1 representing fair quality, and 0 representing poor quality.

Part One: Pre-Condition Guidelines

PC-1 (1) Obtain the necessary permission from the holder of the intellectual property rights relating to the test before carrying out any adaptation (ITC, 2017).		
Test developers or researchers must respect the copyright laws of the original measure. Intellectual property rights protect educational and psychological tests. Test developers or researchers conducting adaptations must retain the original characteristics of the test (structure, material, format, scoring, etc.) to avoid violating this copyright UNLESS they have a specific agreement with the test publisher/author/copyright holder to make modifications of these characteristics.		
2	1	0
The test developers describe the establishment of a signed agreement with the test author/publisher before beginning the adaptation, which specifies acceptable modifications and describes who owns the rights to the adapted test.	The test developers mention the establishment of a license or agreement with the original publisher, but do not provide more detail about this agreement.	The test developers do not describe the establishment of a copyright license or agreement with the original publisher **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.
PC-2 (2) Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the populations of interest is sufficient for the intended use (or uses) of the scores (ITC, 2017).		
Test developers and researchers need to ensure that the construct being assessed is understood in the same way across language and cultural groups.		

2	1	0
The test developers recruit experts in the target construct (e.g., intelligence) and individuals who are familiar with the target cultural group to determine if the construct is legitimate in that context. Methods used to achieve this goal could include surveys, focus groups, and interviews.	The test developers or researchers compile and cite previous research evidence regarding the target construct in the culture of interest (e.g., discussion of other, similar tests) to support the measure's suitability for the cultural/linguistic group. However, an independent investigation of construct overlap is not conducted.	The test developers do not provide information about the overlap between the construct, as measured by test and the definition of the construct in the target culture **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.

PC-3 (3) Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest (ITC, 2017).		
Test developers need to identify and attempt to control irrelevant variables related to culture and language that may impact the participants' performance on the adapted measure. Specifically, qualitative information about the "cultural distance" between the source and target language groups should be collected.		
2	1	0
The test developers conduct observations, interviews, focus groups, or surveys with potential participants in the target culture to determine potential irrelevant variables such as: motivational levels, understanding of the test instructions, experience and familiarity with psychological tests/rating scales, the speediness of test administration, and cultural differences (e.g.,	The test developers or researchers seek feedback (e.g., through interview, survey, or think-aloud session) from their translators, who are native to the target language or culture, to determine potential irrelevant variables in the participant group such as: motivational levels, familiarity and experience with psychological tests,	The test developers or researchers do not describe efforts to identify and/or control irrelevant variables, **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.

family structure, religion, lifestyle, values). If any factors are identified as problematic, they are controlled for in later empirical analyses.	speediness of test-taking, and family structure, religion, values, and lifestyle. If any factors are identified as problematic, they are controlled for in later empirical analyses.	

Part Two: Test Development Guidelines

TD-1 (4) Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise (ITC, 2017).		
Test developers must identify translators with expertise in the languages involved, the cultures involved, the content of the test, and general principles of psychological testing.		
2	1	0
The test developers identify a team of **two or more** translators, at least one of whom is native in the target language and lives in the target country. Additionally, the team of translators includes at least one individual who has an in-depth knowledge of the target locale's culture and one individual who is an expert in the test content and assessment principles. Training in test development (e.g., item writing) is provided and documented for all translators who lack background knowledge in this area.	The test developers identify a team of at least two translators, but none currently live in the target locale. The team of translators includes at least one individual with expert knowledge of the target locale's culture and one individual with some knowledge of the test content and assessment principles. However, training in test development (e.g., item writing) is not provided to translators who lack this background experience.	The test developers identify only one translator **OR** employ a translator that does not have native knowledge of the language and culture of the target country **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.
TD-2 (5) Use appropriate translation designs and procedures to maximize the suitability of the test adaptation in the intended populations (ITC, 2017).		

The translators or translating team must translate the test or measure of interest such that the language is natural, acceptable, and functional. An exact literal translation of the measure is unlikely to be suitable.		
2	1	0
The test translators combine multiple translation designs. For example, first a double-translation and reconciliation procedure is used. In this approach, two obtain independent forward translations of the measure are reviewed by a third translator or expert panel to resolve discrepancies and produce a single adapted version of the test. Next, a backward translation is used to check the accuracy of the first version to produce a final version. Additionally, a checklist or rating scale is used to evaluate any adapted items.	The test translators use a double- translation and reconciliation procedure.	A forward or back translation alone is used to translate and adapt the test **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.

TD-3 (6) Provide evidence that the test instructions and item content have similar meaning for all intended populations (ITC, 2017).		
During the adaptation process, test developers must collect data about whether the items and instructions make sense to the target population.		
2	1	0
The test developers provide two sources of **converging** evidence supporting the meaning of the test in the target	The test developers provide one source of evidence supporting the meaning of the test in the target population. This	The test developers do not describe a systematic approach for evaluating the meaning of the adapted item content and

population. This could include data from: 1) previous studies and test adaptations 2) a small try-out study of the test, 3) a cognitive interviewing procedure, 4) feedback from reviewers native to the target country 5) a pilot test with bilingual respondents or 6) survey administration	could include data from: 1) previous studies and test adaptations 2) a small try-out study of the test, 3) a cognitive interviewing procedure, 4) feedback from reviewers native to the target country 5) a pilot test with bilingual respondents or 6) survey administration	instructions to the target population **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.
TD-4 (7) Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations (ITC, 2017).		
The test developers must ensure that the item format of the adapted test or mode of administration (e.g., computerized) is familiar to target respondents to avoid bias in measuring the construct of interest.		
2	1	0
The test developers use a systematic method (e.g., surveys, cognitive interviewing, try-out study) to assess **ALL of** the following: 1) are respondents familiar with the format of the test, 2) are conventions of the test familiar to the respondents (e.g., marking responses), and 3) do the practice items or exercises represent enough training to prepare the respondents to respond appropriately to the actual test items.	The test developers use a systematic method (e.g., surveys, cognitive interviewing, try-out study) to assess **at least one** of the following: 1) are respondents familiar with the format of the test, 2) are the conventions of the test familiar to the respondents (e.g., marking responses), and 3) are the practice items or exercises enough training to prepare the respondents to respond appropriately to the actual test items.	The test developers do not evaluate whether these aspects of the test are suitable for the intended population **OR** the evidence is lacking such that it does not fulfill the necessary criteria to merit a rating of 2 or 1.

TD-5 (8) Collect pilot data on the adapted test to enable item analysis, reliability assessment and small-scale validity studies so that any necessary revisions to the adapted test can be made (ITC, 2017).		
Before administering the adapted test to a large number of individuals for norming and psychometric investigations, the test developers must provide initial evidence of the adapted test's psychometric quality. The purpose of this analysis is to determine whether it is appropriate to move forward with the test adaptation.		
2	1	0
The test developers conduct a pilot study with a moderately sized sample (e.g., 100 participants) and analyze **at least 2/3** of the following: 1) item analysis to obtain information about discrimination and item level means 2) basic reliability analysis (e.g., coefficient alpha) 3) at least one validity study to evaluate factors such as different modes of administration, instruction phrasings, or test length (ITC, 2017).	The test developers conduct a pilot study with a moderately sized sample (e.g., 100 participants) and analyze **ONE of** the following1) item analysis to obtain information about discrimination and item level means 2) basic reliability analysis (e.g., coefficient alpha) 3) at least one validity study to evaluate factors such as different modes of administration, instruction phrasings, or test length (ITC, 2017).	The test developers do not collect pilot data **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.

Part Three: Confirmation Guidelines

C-1 (9) Select sample with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses (ITC, 2017).		
Test developers must ensure that they collect a large enough sample to conduct statistical analyses that will allow them to establish norms, make judgments about reliability and validity, and analyze differential item functioning.		
2	1	0
The test developers collect a sample of at least 300 participants that is representative of the target population for which the	The test developers collect a sample of above 200 but fewer than 300 participants.	The test developers collect a sample of fewer than 200 respondents **OR** the evidence is lacking such

test will be used.		that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.
C-2 (10) Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations (ITC, 2017).		
Test developers must investigate and establish the construct, method, and item equivalence of the adapted measure, especially if their goal is to compare test-taker performance across two language versions of the test.		
2	1	0
Test developers adequately investigate all three forms of equivalence as follows: 1) **Construct Equivalence**: test developers conduct a confirmatory factor analysis (CFA) or Weighted Multi-Dimensional Scaling (WMDS) procedure AND examine the convergent and discriminant validity of the adapted measure 2) **Method Equivalence:** test developers assess and describe possible sources of method bias (e.g., differential motivation, experience with testing, speediness). 3) **Item equivalence:** Test developers complete a Differential Item Functioning analysis using a standard procedure that best matches their data structure and sample size (e.g., IRT for large	Test developers adequately investigate **2 out of 3** of the following: 1) **Construct Equivalence**: test developers conduct a confirmatory factor analysis (CFA) or Weighted Multi-Dimensional Scaling (WMDS) procedure AND examine the convergent and discriminant validity of the adapted measure 2) **Method Equivalence:** test developers assess and describe possible sources of method bias (e.g., differential motivation, experience with testing, speediness). 3) **Item equivalence:** Test developers complete a Differential Item Functioning analysis using a standard procedure that best matches their data structure and sample size (e.g., IRT for large samples, Mantel-Haenszel	Test developers adequately investigate **1 or fewer** of the following: 1) **Construct Equivalence**: test developers conduct a confirmatory factor analysis (CFA) or Weighted Multi-Dimensional Scaling (WMDS) procedure AND examine the convergent and discriminant validity of the adapted measure 2) **Method Equivalence:** test developers assess and describe possible sources of method bias (e.g., differential motivation, experience with testing, speediness). 3) **Item equivalence:** Test developers complete a Differential Item Functioning analysis using a standard

samples, Mantel-Haenszel for smaller samples) AND provide hypothetical reasons for any observed DIF	for smaller samples) AND provide hypothetical reasons for any observed DIF	procedure that best matches their data structure and sample size (e.g., IRT for large samples, Mantel-Haenszel for smaller samples) AND provide hypothetical reasons for any observed DIF. **OR** there is insufficient information provided to merit a rating of 2 or 1.

C-3 (11) Provide evidence supporting the norms, reliability and validity of the adapted version of the test in the intended populations (ITC, 2017).		
Test developers must present empirical information about the validity and reliability of the adapted measure, with special consideration of five sources of validity evidence (e.g., test content, response processes, internal structure, relations to other variables, and consequences of testing).		
2	1	0
The test developers fulfill the following **three** criteria: 1) Evidence is provided that the existing test norms can be used appropriately and fairly for the adapted measure or new norms are developed 2) evidence is provided that the new test is reliable for its intended use (e.g., internal consistency estimate of .80 or higher for clinical use) 3) at least 3 sources of validity evidence are described that are related to the intended use of the test scores (e.g., factor analysis,	The test developers fulfill **two out of three** of the following criteria: 1) Evidence is provided that the existing test norms can be used appropriately and fairly for the adapted measure or new norms are developed 2) evidence is provided that the new test is reliable for its intended use (e.g., internal consistency estimate of .80 or higher for clinical use) 3) at least 3 sources of validity evidence are described that are related to the intended use of the test scores (e.g., factor analysis,	The test developers fulfill **one or fewer** of the following criteria: 1) Evidence is provided that the existing test norms can be used appropriately and fairly for the adapted measure or new norms are developed 2) evidence is provided that the new test is reliable for clinical use, as indicated by internal consistency estimates of .80 or higher 3) at least 3 sources of validity evidence are described that are related to the

convergent validity)	convergent validity) **OR** there is insufficient information about any of the three criteria to merit a rating of 2.	intended use of the test scores (e.g., factor analysis, convergent validity) **OR** there is insufficient information about any of the three criteria to merit a rating of 1.

C-4 (12) Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test (ITC, 2017).		
Test developers should place the original and adapted versions of the test on a single reporting scale before making cross-cultural comparisons. This should only be done if there is strong evidence of equivalence.		
2	1	0
If there is strong evidence of equivalence and cross-cultural comparison is desired, the test developers perform ONE of the following to link the two test versions: 1) a bilingual group design, 2) a matched monolingual design, or 3) a monolingual group design **OR** Linking is not attempted due to1) shortcomings in addressing equivalence or 2) no need or intent to compare performance across cultural groups	There is strong evidence of equivalence and linking is used, but the test developers do not provide details about the methods used to achieve a common reporting scale **OR** a linking design is not used, but no explanation for this choice is given.	A linking design is attempted without sufficient evidence of equivalence **OR** there is insufficient information to merit a rating of 1 or 2.

Part Four: Administration Guidelines

A-1 (13) Prepare administration materials and instructions to minimize any culture- and language-related problems that are caused by administration

procedures and response modes that can affect the validity of the inferences drawn from the scores (ITC, 2017).		
Test developers must carefully anticipate problems that might arise during test administration with the target cultural group due to cultural and linguistic variables.		
2	1	0
Test developers systematically assess and address **at least four** of the following factors: 1) the clarity of test instructions, 2) the appropriateness of the test response method, 3) the timing of the test, 4) motivation to participate in testing, 5) knowledge of the test's purpose, and 6) appropriateness of scoring methods for the target cultural group. Necessary changes in these factors from the original version are detailed and justified.	Test developers systematically assess and address **at least two** of the following factors: 1) the clarity of test instructions, 2) the appropriateness of the test response method, 3) the timing of the test, 4) motivation to participate in testing, 5) knowledge of the test's purpose, and 6) appropriateness of scoring methods for the target cultural group. Necessary changes in the four described factors from the original version are detailed and justified.	Test developers fail to address test administration factors during the adaptation process **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.
A-2 (14) Specify testing conditions that should be followed closely in all populations of interest (ITC, 2017).		
Test developers must develop standardized testing procedures and conditions (e.g., time limits, location) that will be followed in both the source and target culture to reduce the impact of these factors on performance, which can limit interpretations of an individual's scores.		
2	1	0
Test developers provide detailed, specific information about the standardized procedures used. They also describe the procedures used to training test administrators in these procedures.	Test developers state that they developed standardized procedures for the adapted test, but do not provide enough information about the specific procedures used to merit a rating of 2.	No information is provided regarding adaptation of test instructions or standardized procedures **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.

Part Five: Score Scales and Interpretation Guidelines

SSI-1 (15) Interpret any group score differences with reference to all relevant available information (ITC, 2017).		
Test developers must make sure that they present appropriate interpretations of intergroup differences in scores on the test of interest.		
2	1	0
Test developers consider multiple possible score interpretations (e.g., motivation, education) and systematically investigate reasons for intergroup score differences (e.g., administer both versions to bilingual respondents)	Test developers consider multiple possible score interpretations, but do not systematically investigate the reasons for intergroup differences.	Test developers do not present multiple possible interpretations for intergroup score differences **OR** there is insufficient evidence to warrant a rating of 1 or 2.
SSI-2 (16) Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported (ITC, 2017).		
Test developers must avoid over-interpreting score differences by making score comparisons between cultural groups in the absence of strong validity evidence.		
2	1	0
Test developers accurately interpret their results based only on the validity evidence that they have collected, such that they only make intergroup comparison statements when measurement invariance is established (see C2) **AND** test developers also caution users about accurate interpretation explicitly in text and discourage misinterpretation/misuse.	Test developers either explicitly caution users about accurate interpretation OR provide guidance for accurately interpreting respondent group differences.	Test developers inaccurately interpret their scores by making comparisons across large groups in the absence of established measurement invariance (see C2) **OR** the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2.

Part Six: Documentation Guidelines

Doc-1 (17) Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population (ITC, 2017).		
Test developers must document the adaptation process to promote future replication and to support the measure's use in the new population.		
2	1	0
Test developers provide a technical manual with both qualitative and quantitative evidence describing the following: 1) Utility of the construct and test in the new population 2) sample characteristics and item data 3) evidence of content, criterion- related, and construct validity 4) a description of data analyses and results	Test developers provide a technical manual detailing **three of the following** pieces of evidence: 1) Utility of the construct and test in the new population 2) sample characteristics and item data 3) evidence of content, criterion-related, and construct validity 4) a description of data analyses and results	The manual does not provide sufficient detail for at least three pieces of evidence OR the evidence is lacking such that it does not fulfill the necessary criteria to obtain a rating of 1 or 2

Doc-2 (18) Provide documentation for test users that will support good practice in the use of an adapted test with people in the context of the new population (ITC, 2017).		
The technical manual should be accessible to users in practical settings.		
2	1	0
The test developers include **ALL** of the following information about the adapted test in a format that is accessible to users (e.g., manual): 1) describe the construct 2) describe the adaptation process 3) summarize evidence	The test developers include **at least 6** of the following details about the adapted test in a format that is accessible to users (e.g., manual): 1) describe the construct 2) describe the adaptation process 3) summarize	The test developers fail to include at least 6 pieces of information in a format that is accessible to users (e.g., manual) **OR** the evidence is lacking such that it does not fulfill the necessary

supporting the need for the adaptation 4) summarize evidence for the cultural suitability of the item content, test instructions, and response format 5) describe the suitability of using the test with various subgroups and any restrictions to use 6) explain issues related to good practice in test administration 7) explain how intergroup comparisons can be made if appropriate 8) provide the information needed for scoring and norming or describe computer-based scoring 9) provide guidelines for the interpretation of results (e.g., reliability, validity, etc).	evidence supporting the need for the adaptation 4) summarize evidence for the cultural suitability of the item content, test instructions, and response format 5) describe the suitability of using the test with various subgroups and any restrictions to use 6) explain issues related to good practice in test administration 7) explain how intergroup comparisons can be made if appropriate 8) provide the information needed for scoring and norming or describe computer-based scoring 9) provide guidelines for the interpretation of results (e.g., reliability, validity, etc).	criteria to obtain a rating of 1 or 2.

X-CAGAM Overall Score Calculation:

[Sum of All Item Scores] X 100/36

1. Sum the scores for each individual item

2. Divide this sum by 36

3. Multiply the result by 100 to obtain a percentage

X-CAGAM overall Score= _____ / 36 X 100 = _____ %

Score	Category	Description
0 - 49%	**Poor Quality**	The measure received a score of zero on more than half of the guideline items (raw score of 17 or lower)
50 - 75%	**Fair Quality**	The measure received a score of two on at least half of guideline items, a score of one on all items, or a combination of scores higher than zero on the majority of items, such that the raw score is between 18 and 27
76 - 100%	**Good Quality**	The measure received a score of two on the majority of guideline items (raw score of 28 or higher)

APPENDIX B

Selection criteria

INCLUSION CRITERIA
Adaptations of ASD rating scales.
Translations of ASD rating scales.
Articles published in peer reviewed journals.
Published/unpublished students thesis/dissertations.
Research conducted after DSM-IV was published (1994)

References

Acharya, K. P., & Pathak, S. (2019). Applied research in low-income countries: why and how?. *Frontiers in Research Metrics and Analytics*, *4*, 3.

Akoury-Dirani, L., Alameddine, M., & Salamoun, M. (2013). Validation of the Lebanese Childhood Autism Rating Scale–Second Edition–High Functioning Version. *Research in Autism Spectrum Disorders*, *7*(11), 1332-1338.

Al Jabery, M. A. (2008). *The examination of validity and reliability indicators of the Jordanian translated Arabic version of the Gilliam Autism Rating Scale (GARS-2)*. Wayne State University.

Altman, D. G. & Bland, J. M. (1994). "Diagnostic tests. 1: Sensitivity and specificity". *BMJ*. **308**(6943): 1552. doi:10.1136/bmj.308.6943.1552

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., Text Revision).* Washington, DC: American Psychiatric Association.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders (4th ed.).* Washington, DC: Author.

Anderson, R. T., Aaronson, N. K., Bullinger, M., & McBee, W. L. (1996). A review of the progress towards developing health-related quality-of-life instruments for international clinical studies and outcomes research. *Pharmacoeconomics*, *10*(4), 336-355.

Asparouhov, T., & Muthén, B. (2018). SRMR in Mplus. *Retrieved from Mplus Web Notes website: http://www. statmodel. com/download/SRMR2. pdf*.

Arafat, S. Y., Chowdhury, H. R., Qusar, M. S., & Hafez, M. A. (2016). Cross cultural adaptation & psychometric validation of research instruments: A methodological review. *Journal of Behavioral Health*, *5*(3), 129-136.

Avcil, S., Baykara, B., Baydur, H., Münir, K. M., & İnal Emiroğlu, N. (2015). 4-18 Yaş Aralığındaki Otistik Bireylerde Sosyal İletişim Ölçeği Türkçe Formunun Geçerlik ve Güvenilirlik Çalışması [The validity and reliability of the Social Communication Questionnaire- Turkish form in autistics aged 4-18 years]. *Turk psikiyatri dergisi = Turkish journal of psychiatry*, *26*(1), 56–64.

Ballerini, M., Stanghellini, G., Chieffi, M., Bucci, P., Punzo, P., Ferrante, G., ... & Galderisi, S. (2015). Autism rating scale (ARS)-Italian version. *JOURNAL OF PSYCHOPATHOLOGY-GIORNALE DI PSICOPATOLOGIA*, *21*(3), 297-308.

Baxter, A. J., Brugha, T. S., Erskine, H. E., Scheurer, R. W., Vos, T., & Scott, J. G. (2015). The epidemiology and global burden of autism spectrum disorders. *Psychological medicine*, *45*(3), 601-613.

Bernier, R., Mao, A., & Yen, J. (2010). Psychopathology, families, and culture: autism. *Child and Adolescent Psychiatric Clinics*, *19*(4), 855-867.

Borsa, Juliane Callegaro, Damásio, Bruno Figueiredo, & Bandeira, Denise Ruschel. (2012). Cross-cultural adaptation and validation of psychological instruments: some considerations. *Paidéia (Ribeirão Preto)*, *22*(53), 423-432.

Bruin, J. (2006). Newtest: command to compute new test. UCLA: Statistical Consulting Group. *Retrieved*, *4*, 2021.

Bryson, S. E., Rogers, S. J., & Fombonne, E. (2003). Autism spectrum disorders: early detection, intervention, education, and psychopharmacological management. *The Canadian Journal of Psychiatry*, *48*(8), 506-516.

Camarata, S. (2014). Early identification and early intervention in autism spectrum disorders: Accurate and effective?. *International Journal of Speech-Language Pathology*, *16*(1), 1-10.

Chia, M. Y. (2012). *Fidelity of implementing an assessment translation and adaptation framework in a study of an emerging international assessment* (Order No. 3549171). Available from ProQuest Dissertations & Theses Global. (1283387484). Retrieved from http://jerome.stjohns.edu:81/login?url=https://search-proquest-com.jerome.stjohns.edu/docview/1283387484?accountid=14068

Christensen, D. L., Braun, K. V. N., Baio, J., Bilder, D., Charles, J., Constantino, J. N., ... & Yeargin-Allsopp, M. (2018). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2012. *MMWR Surveillance Summaries*, *65*(13), 1.

Ciocca, D. R., and Delgado, G. (2017). The reality of scientific research in Latin America; an insider's perspective. *Cell Stress Chaperones* 22, 847–852. doi: 10.1007/s12192-017-0815-8

Constantino, J. N., & Gruber, C. P. (2012). *Social responsiveness scale: SRS-2*. Torrance, CA: Western Psychological Services.

Dalwai, S., Ahmed, S., Udani, V., Mundkur, N., Kamath, S. S., & Nair, M. K. C. (2017). Consensus statement of the Indian Academy of Pediatrics on

evaluation and management of autism spectrum disorder. *Indian pediatrics*, *54*(5), 385-393.

Daley, T. C. (2004). From symptom recognition to diagnosis: children with autism in urban India. *Social science & medicine*, *58*(7), 1323-1335.

Diken, I. H., Diken, O., Gilliam, J. E., Ardic, A., & Sweeney, D. (2012). Validity and Reliability of Turkish Version of Gilliam Autism Rating Scale-2: Results of Preliminary Study. *International Journal of Special Education*, *27*(2), 207-215.

Dobson, S., Upadhyaya, S., McNeil, J., Venkateswaran, S., & Gilderdale, D. (2001). Developing an information pack for the Asian carers of people with autism spectrum disorders. *International journal of language & communication disorders*, *36*(S1), 216-221.

Doernberg, E., & Hollander, E. (2016). Neurodevelopmental Disorders (ASD and ADHD): DSM-5, ICD-10, and ICD-11. *CNS spectrums*, *21*(4), 295-299.

Duke, D. (2019). Developing a rubric to measure adherence to the international test commission guidelines for translating and adapting tests (Unpublished doctoral dissertation). St. John's University, New York.

El Shourbagi, S., & Abd-El-Fattah, S. M. (2019). The Adaptation of Gilliam Autism Rating Scale-2 within an Omani Context: Some Initial Findings. *AsianJournal of Education and Social Studies*, 1-10.

Elsabbagh, M., Divan, G., Koh, Y. J., Kim, Y. S., Kauchali, S., Marcín, C., ... & Yasamy, M. T. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism research*, *5*(3), 160-179.

Ferri, S. L., Abel, T., & Brodkin, E. S. (2018). Sex differences in autism spectrum disorder: a review. *Current psychiatry reports*, *20*(2), 1-17.

Bishop-Fitzpatrick, L., & Kind, A. J. (2017). A scoping review of health disparities in autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *47*(11), 3380-3391.

Gatignon, H. (2010). *Confirmatory Factor Analysis in Statistical analysis of management data*. DOI: 10.1007/978-1-4419-1270-1_4

Gau, S. S. F., Lee, C. M., Lai, M. C., Chiu, Y. N., Huang, Y. F., Kao, J. D., & Wu, Y. Y. (2011). Psychometric properties of the Chinese version of the Social Communication

Questionnaire. *Research in Autism Spectrum Disorders*, *5*(2), 809-818.

Gillberg, C. & Wahlstrom, J. (1985). Chromosome abnormalities in infantile autism and other childhood psychoses: a population study of 66 cases. *Developmental Medicine & Child Neurology*. 27, 293–304. doi: 10.1111/j.1469-8749.1985.tb04539.x

Gilliam, J. E. (2006). *Gilliam autism rating scale-2*. Austin, TX: PRO-ED.

Gilliam J. E., (1995). Gilliam autism rating scale. Austin: PRO-ED.

Goldstein, S., & Naglieri, J. A. (2010). *Autism Spectrum Rating Scales: ASRS*. North Tonawanda, NY: Multi-Health Systems.

Guo, Y. Q., Tang, Y., Rice, C., Lee, L. C., Wang, Y. F., & Cubells, J. F. (2011). Validation of the autism spectrum screening questionnaire, Mandarin Chinese version (CH-ASSQ) in Beijing, China. *Autism*, *15*(6), 713-727.

Hahler, E. M., & Elsabbagh, M. (2015). Autism: A global perspective. *Current Developmental Disorders Reports*, *2* (1), 58-64.

Hambleton, R.K. (2005). Issues, Designs, and Technical Guidelines for Adapting Tests Into Multiple Languages and Cultures. In Hambleton, R.K., Merenda,

P.F. & Spielberger, C. D. (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* New York, NY: Psychology Press

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and manova in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling, 7*, 534-556.

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management*, *45*(6), 2596-2627.

Henrich, J., Heine, S. J., & Norenzayan A. The weirdest people in the world? Behav Brain Sci. 2010 Jun;33(2-3):61-83; discussion 83-135. doi: 10.1017/S0140525X0999152X. Epub 2010 Jun 15. PMID: 20550733.

Hewitt, C. E., Mitchell, N., & Torgerson, D. J. (2008). Listen to the data when results are not significant. *BMJ (Clinical research ed.)*, *336*(7634), 23–25. https://doi.org/10.1136/bmj.39379.359560.AD

Hill, C. V., Pérez-Stable, E. J., Anderson, N. A., & Bernard, M. A. (2015). The National Institute on Aging health disparities research framework. *Ethnicity & disease*, *25*(3), 245.

Hock, R., & Ahmedani, B. K. (2012). Parent perceptions of autism severity: Exploring the social ecological context. *Disability and Health Journal*, *5*(4), 298-304.

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *Journal of Business Research Methods*, *6*, 53–60.

Hossain, M. D., Ahmed, H. U., Uddin, M. J., Chowdhury, W. A., Iqbal, M. S., Kabir, R. I., ... & Hossain, S. W. (2017). Autism Spectrum disorders (ASD) in South Asia: a systematic review. *BMC psychiatry*, *17*(1), 281

Hussein, H., & Taha, G. R. (2013). Autism spectrum disorders: a review of the literature from Arab countries. *Middle East Current Psychiatry*, *20*(3), 106-116.

International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). [www.InTestCom.org]

International Test Commission (2005). International Guidelines on Test Adaptation. [www.intestcom.org]

Jackson, L. S., Little, S. G., & Akin-Little, A. (2013). The Spanish adaptation of the Gilliam Autism Rating Scale-2: Translation and psychometric analysis. *Research in Autism Spectrum Disorders*, *7*(9), 1160–1167.

Juneja, M., Mishra, D., Russell, P. S., Gulati, S., Deshmukh, V., Tudu, P., ... & Inclen Study Group. (2014). INCLEN diagnostic tool for autism spectrum disorder (INDT-ASD): Development and validation. *Indian pediatrics*, *51*(5), 359-365.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous child*, *2*(3), 217-250.

Koegel, R. L., & Koegel, L. K. (2006). *Pivotal response treatments for autism: Communication, social, & academic development*. Paul H Brookes Publishing.

Krabbe, P. (2016). *The measurement of health and health status: concepts, methods and applications from a multidisciplinary perspective*. Academic Press.

Krug, D. A., Arick, J., & Almond, P. (1980). Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. *Journal of Child Psychology and Psychiatry*, *21*(3), 221-229.

Landa, R. J. (2018). Efficacy of early interventions for infants and young children with, and at risk for, autism spectrum disorders. *International Review of Psychiatry*, *30*(1), 25-39.

Lord, C., & Jones, R. M. (2012). Annual Research Review: Re‐thinking the classification of autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *53*(5), 490-509.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., Bishop, S. L., & Schedule, A.A.D.O. (2015). ADOS-2. *Manual (Part I): Modules*, 1-4.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, *24*(5), 659-685.

Li, N. (2012). Preliminary validation of the Childhood Autism Rating Scale-questionnaire for parents or caregivers (CARS2-QPC) and the Gilliam Autism Rating Scale (GARS-2) with a Chinese-speaking Population.

Lian, M. (1996). *Teaching Asian American children. England. In: Duran E, editor. Teaching students with moderate/severe disabilities, including autism: strategies for second language learners in inclusive settings-2nd edition.* Springfield (IL): Charles C Thomas, Publisher; p. 239.

Lung, F. W., Shu, B. C., Chiang, T. L., & Lin, S. J. (2009). Parental mental health, education, age at childbirth and child development from six to 18 months. *Acta Paediatrica*, *98*(5), 834-841.

Lung, F. W., Shu, B. C., Chiang, T. L., & Lin, S. J. (2008). Efficient developmental screening instrument for 6‐and 18‐month‐old children in the Taiwan Birth Cohort Pilot Study. *Acta Paediatrica*, *97*(8), 1093-1098.

Lynn, R., & Vanhanen, T. (2012). National IQs: A review of their educational, cognitive, economic, political, demographic, sociological, epidemiological, geographic and climatic correlates. Intelligence, 40(2), 226-234. doi:10.1016/j.intell.2011.11.004

Maddox, B. B., Brodkin, E. S., Calkins, M. E., Shea, K., Mullan, K., Hostager, J., … Miller, J. S. (2017). The Accuracy of the ADOS-2 in Identifying Autism among Adults with Complex Psychiatric Conditions. *Journal of autism and developmental disorders*, *47*(9), 2703–2709. doi:10.1007/s10803-017-3188-z

Mandell, D. S., & Novak, M. (2005). The role of culture in families' treatment decisions for children with autism spectrum disorders. *Mental retardation and developmental disabilities research reviews*, *11*(2), 110-115.

Matson, J. L., & Neal, D. (2010). Differentiating communication disorders and autism in children. *Research in Autism Spectrum Disorders*, *4*(4), 626-632.

Mattila, M. L., Jussila, K., Linna, S. L., Kielinen, M., Bloigu, R., Kuusikko-Gauffin, S., ... & Moilanen, I. (2012). Validation of the Finnish Autism Spectrum Screening Questionnaire (ASSQ) for clinical settings and total population screening. *Journal of autism and developmental disorders*, *42*(10), 2162-2180.

May, T., Cornish, K., & Rinehart, N. (2014). Does gender matter? A one year follow-up of autistic, attention and anxiety symptoms in high-functioning children with autism spectrum disorder. *Journal of autism and developmental disorders*, *44*(5), 1077-1086.

McClain, M. B., Harris, B., Schwartz, S. E., & Golson, M. E. (2020). Evaluation of the autism spectrum rating scales in a diverse, nonclinical sample. *Journal of Psychoeducational Assessment*, *38*(6), 740-752.

McDonald, R. P., Seifert, C. F., Lorenzet, S. J., Givens, S., & Jaccard, J. (2002). The effectiveness of methods for analyzing multivariate factorial data. *Organizational Research Methods, 5*, 255-274.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of psychological research*, *3*(1), 111-130.

Ministry of Social Justice and Empowerment. Scientific Report on Research Project for Development of Indian Scale for Assessment of Autism. New Delhi: Government of India; 2009.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med*, *6*(7), e1000097.

Muniz, J., Elosua, P., Hambleton, R. K., & International Test Commission. (2013). International Test Commission Guidelines for test translation and adaptation. *Psicothema*, *25*(2), 151.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological science*, *31*(6), 678-701.

Norris, M., & Lecavalier, L. (2010). Screening accuracy of level 2 autism spectrum disorder rating scales: A review of selected instruments. *Autism*, *14*(4), 263-284.

Oakland, T., Poortinga, Y. H., Schlegel, J., & Hambleton, R. K. (2001). International Test Commission: Its history, current status, and future directions. *International Journal of Testing*, *1*(1), 3-32.

Oakland, T., & Lane, H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing*, *4*(3), 239-252.

Oakland, T. (2004). *Use of Educational and Psychological Tests Internationally. Applied Psychology, 53(2), 157–172.* doi:10.1111/j.1464-0597.2004.00166.x

Payakachat, N., Tilford, J. M., Kovacs, E., & Kuhlthau, K. (2012). Autism spectrum disorders: a review of measures for clinical, health services and cost-effectiveness applications. *Expert review of pharmacoeconomics & outcomes research*, *12*(4), 485–503. https://doi.org/10.1586/erp.12.29

Pereira, A., Riesgo, R. S., & Wagner, M. B. (2008). Childhood autism: translation and validation of the Childhood Autism Rating Scale for use in Brazil. *Jornal de Pediatria*, *84*(6), 487-494.

Perry, A., Condillac, R. A., Freeman, N. L., Dunn-Geier, J., & Belair, J. (2005). Multi-site study of the Childhood Autism Rating Scale (CARS) in five clinical groups of young children. *Journal of autism and developmental disorders*, *35*(5), 625-634.

Peters-Scheffer, N., Didden, R., Korzilius, H., & Sturmey, P. (2011). A meta-analytic study on the effectiveness of comprehensive ABA-based early intervention programs for children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *5*(1), 60-69.

Pierce, K., Gazestani, V. H., Bacon, E., Barnes, C. C., Cha, D., Nalabolu, S., ... & Courchesne, E. (2019). Evaluation of the diagnostic stability of the early

autism spectrum disorder phenotype in the general population starting at 12 months. *JAMA pediatrics*, *173*(6), 578-587.

Qiu, S., Lu, Y., Li, Y., Shi, J., Cui, H., Gu, Y., ... & Cheng, Y. (2019). Prevalence of autism spectrum disorder in Asia: A systematic review and meta-analysis. *Psychiatry research*, 112679.

Osman, O. T., & Afifi, M. (2010). Troubled Minds in The Gulf: Mental Health Research in the United Arab Emirates (1989-2008). *Asia Pacific Journal of Public Health*, *22*(3_suppl), 48S-53S.

Qoronfleh, M. W., Essa, M. M., Alharahsheh, S. T., Al-Farsi, Y. M., & Al-Adawi, S. (2019). Autism in the Gulf states: A regional overview. *Frontiers in Bioscience, Landmark*, *24*, 324-336.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401-11405.

Ramsay, M. C., Reynolds, C. R., & Kamphaus, R. W. (2002). *Essentials of behavioral assessment* (Vol. 37). John Wiley & Sons.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin*, *114*(3), 552.

Rice, C. E., Rosanoff, M., Dawson, G., Durkin, M. S., Croen, L. A., Singer, A., & Yeargin-Allsopp, M. (2012). Evaluating changes in the prevalence of the autism spectrum disorders (ASDs). *Public Health Reviews*, *34* (2), 17.

Rios, J. A., & Sireci, S. G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing*, *14*(4), 289-312.

Robbins, M., & Thomas, K. (2018). Women in the Middle East and North Africa: A divide between rights and roles. *Retrieved January*, *23*, 2020.

Rutter, M., Bailey, A., & Lord, C. (2003). *The social communication questionnaire: Manual*. Western Psychological Services.

Samadi, S. A., & McConkey, R. (2014). The utility of the Gilliam autism rating scale for identifying Iranian children with autism. *Disability and rehabilitation*, *36*(6), 452-456.

Samms-Vaughan, M. E. (2014). The status of early identification and early intervention in autism spectrum disorders in lower-and middle-income countries. *International Journal of Speech-Language Pathology*, *16*(1), 30-35.

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347-363.

Schopler, E., Van Bourgondien, M., Wellman, G., & Love, S. (2010). *Childhood autism rating scale-second edition (CARS2): Manual*. Los Angeles: Western Psychological Services.

Schopler, E., Reichler, R. J., & Renner, B. R. (1988). *CARS: The childhood autism rating scale*. Western Psychological Services.

Scherzer, A. L., Chhagan, M., Kauchali, S., & Susser, E. (2012). Global perspective on early diagnosis and intervention for children with developmental delays and disabilities. *Developmental Medicine & Child Neurology*, *54*(12), 1079-1084.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. *Adapting educational and psychological tests for cross-cultural assessment*, 93-115.

Shah, K. (2001). What do medical students know about autism?. *Autism*, *5*(2), 127-133.

Sprangers, M. A. G., Cull, A., Bjordal, K., Groenvold, M., & Aaronson, N. K. (1993). The European Organization for Research and Treatment of Cancer approach to quality of life assessment: guidelines for developing questionnaire modules. *Quality of Life Research*, *2*(4), 287-295.

Soto, S., Linas, K., Jacobstein, D., Biel, M., Migdal, T., & Anthony, B. J. (2015). A review of cultural translations and adaptations of screening tools for autism spectrum disorders. *Autism*, *19*(6), 646-661.

Sweileh, W. M., Al-Jabi, S. W., Sawalha, A. F., & Sa'ed, H. Z. (2016). Bibliometric profile of the global scientific research on autism spectrum disorders. *Springerplus*, *5*(1), 1480.

Sun, X., & Allison, C. (2010). A review of the prevalence of autism spectrum disorder in Asia. *Research in Autism Spectrum Disorders*, *4*(2), 156-167.

Stansfield, C. W. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, *20*(2), 189-207.

Stewart, L. A., & Lee, L. C. (2017). Screening for autism spectrum disorder in low- and middle-income countries: A systematic review. *Autism*, *21*(5), 527-539.

Szatmari, P., Bryson, S. E., Boyle, M. H., Streiner, D. L., & Duku, E. (2003). Predictors of outcome among high functioning children with autism and Asperger syndrome. *Journal of Child Psychology and Psychiatry*, *44*(4), 520-528.

Tabachnick, B., & Fidell, L. (2013). Using Multivariate Statistics. Boston, MA: Pearson Education Inc. [Paper reference 11]

Tachimori, H., Osada, H., & Kurita, H. (2003). Childhood Autism Rating Scale−Tokyo Version for screening pervasive developmental disorders. *Psychiatry and clinical neurosciences*, *57*(1), 113-118.

Tafiadis, D., Loli, G., Tsanousa, E., & Tafiadi, M. (2008). The Gilliam Autism Rating Scale (GARS-2), a pilot study for the Greek autistic population. *Annals of General Psychiatry*, *7*(1), 1-1.

The World Bank Group (2020). https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Tincani, M., Travers, J., & Boutot, A. (2009). Race, Culture, and Autism Spectrum Disorder: Understanding the Role of Diversity in Successful Educational Interventions. *Research and Practice for Persons with Severe Disabilities*, *34*(3–4), 81–90.

Uwaezuoke, Samuel. (2015). Autism Spectrum Disorder in Children: The Disparities between the Developed and Developing Countries. *Autism-Open Access. 05*. 10.4172/2165-7890.1000152.

Van Widenfelt, B. M., Treffers, P. D., De Beurs, E., Siebelink, B. M., & Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical child and family psychology review*, *8*(2), 135-147.

Volkmar, F. R., Cicchetti, D. V., Dykens, E., Sparrow, S. S., Leckman, J. F., & Cohen, D. J. (1988). An evaluation of the autism behavior checklist. *Journal of Autism and Developmental Disorders*, *18*(1), 81-97.

Wahlstrom, J., Gillberg, C., Gustavson, K. H., and Holmgren, G. (1986). Infantile autism and the fragile X. A Swedish multicenter study. *American Journal of Medicine Genet*ics 23, 403–408. doi: 10.1002/ajmg.1320230132

Wang, J., Zhou, X., Xia, W., Sun, C., Wu, L., & Wang, J. (2012). Autism awareness and attitudes towards treatment in caregivers of children aged 3–6 years in Harbin, China. *Social psychiatry and psychiatric epidemiology*, *47*(8), 1301-1308.

Wang, J., Lee, L., Chen, Y., & Hsu, J. (2012). Assessing autistic traits in a taiwan preschool population: Cross-cultural validation of the social responsiveness scale (SRS). Journal of Autism and Developmental Disorders, 42(11), 2450-9.

Williams, J. G., Higgins, J. P., & Brayne, C. E. (2006). Systematic review of prevalence studies of autism spectrum disorders. *Archives of disease in childhood*, *91*(1), 8-15.

Won, H., Mah, W., & Kim, E. (2013). Autism spectrum disorder causes, mechanisms, and treatments: focus on neuronal synapses. *Frontiers in molecular neuroscience*, *6*, 19.

World Health Organization (1992). *The ICD- I0 Classification of Mental and Behavioral Disorders- Clinical descriptions and Diagnostic Guidelines Descriptions.* Geneva: WHO.

Yousefi, N., Dadgar, H., Mohammadi, M. R., Jalilevand, N., Keyhani, M. R., & Mehri, A. (2015). The validity and reliability of Autism Behavior Checklist in Iran. *Iranian journal of Psychiatry*, *10*(3), 144.

Zhou, H., Li, C. P., Huang, Y., Zou, X. B., Luo, X. R., Wu, L. J., ... & Wang, Y. (2019). Reliability and validity of the translated Chinese version of Autism Spectrum Rating Scale (2–5 years). *World Journal of Pediatrics*, *15*(1), 49-56.

Zhou, H., Zhang, L., Luo, X., Wu, L., Zou, X., Xia, K., ... & Wang, Y. (2017). Modifying the Autism Spectrum Rating Scale (6–18 years) to a Chinese context: an exploratory factor analysis. *Neuroscience bulletin*, *33*(2), 175-182.

Vita

Name	*Tamanna Chhabra*
Baccalaureate Degree	*Bachelor of Arts, University Of Delhi, India. Major: Psychology Honors.*
Date Graduated	*March, 2009*
Other Degrees	*Master of Arts, Tata Institute of Social Sciences, Mumbai, India: Major: Counseling*
Date Graduated	*March, 20111*
Other Degrees	*Master of Philosophy, Amity University Uttar Pradesh India, Major: Clinical Psychology*
Date Graduated	*July, 2015*
Other Degrees	*Master of Science, St. John's University, Queens, NY, Major: School Psychology*
Date Graduated	*May, 2019*