ABSTRACT

Michaela DuBay: Translation and Cultural Adaptation of Autism Screening Tools
(Under the direction of Linda R. Watson)

Parent-report screening tools for autism spectrum disorder (ASD) are widely used as a way to promote early identification of children with or at risk for ASD. Results from these questionnaires provide information about the presence of early symptoms of ASD and can aid in decisions about referrals for evaluation and treatment. Most of these screening tools were developed in English in the US or UK. In order to meet the needs of culturally and linguistically diverse populations, nationally and internationally, these tools require translation. Traditional translation methods (“forward-back approach”) include a forward translation, back translation, and review. The first paper in this dissertation is a psychometric analysis of the forward-back translation methodology. We compared parent report data from the original English Modified Checklist for Autism in Toddlers – Revised (Robins et al., 2014) to its forward-back Spanish Western-Hemisphere translation to examine the psychometric properties of a tool translated using the traditional methodology. Results indicated significant differences between the two versions, including a higher overall risk score and screen positive rate among the Spanish records, as well as more items left blank among Spanish-speaking respondents. Traditional translation methods did not appear sufficient to maintain similar psychometric properties between the language versions. The second paper is a review of literature proposing and supporting a more rigorous cultural adaptation approach to the translation process. This paper
presents a specific set of guidelines from the literature along with rationales for each step in the process. A quality appraisal tool is presented to serve as a systematic measure of translations or to guide future translation and adaptation teams. The final clinical paper is an implementation of this rigorous cultural adaptation approach, using the *First Years Inventory v 3.1* (Baranek et al., 2013) as a test case. This parent-report screening tool was translated and culturally adapted for a target population of US-based Spanish-speaking caregivers. The paper describes the translation process, highlights specific areas of the instrument that presented the most challenges and required the most adaptations, and examines the value of the individual steps in the translation process to inform future translation and adaptation teams.
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Linda Watson, who saw the seed of my interest in this topic and expertly cultivated it into a series of distinct research questions. She encouraged me to attempt difficult but rewarding and meaningful work, answered endless questions, revised and refocused, and generally expressed full and complete confidence in my abilities, the impact of which will surely last my whole career.

I owe much gratitude to the rest of my dissertation committee, Grace Baranek, Lucía Mendez, Betsy Crais, and Sharon Williams. They gave me thoughtful guidance and knowledgeable support throughout this process, and I thank them for granting me their expertise, their perspectives, and their work for this dissertation.

These studies would not have come to fruition without the cooperation and collaboration of many medical practices, community organizations, and families. I am indebted to the nurses and staff at medical practices who meticulously saved M-CHAT forms from their patients and scrubbed names and birthdates. I am eternally grateful to the local community organizations who trusted me, allowed me to work with their families, and helped me find participants for these studies. And I would like to express endless gratitude to the mothers, fathers, grandmothers, and aunts who welcomed us into their homes and helped us understand their perspectives. Y quiero expresar un agradecimiento infinito a las madres, los padres, las abuelas, y las tíos que nos recibieron en sus hogares y nos ayudaron a comprender sus perspectivas.
I had the great pleasure of working with many research assistants, translators, and collaborators. Thank you to Danna Smith, Will Brinson, Carolina Rojevic, Nuria Vicens, Mary Jones, Yulieth Peña, Helen Lee, Valentina Vera Carrasquero, Elizabeth Choi, and Allison Phillips. I could not have done this work without your sincere dedication.

Many thanks also to my family. To Kateri for professional moral support and accountability. To Lizabeth for all-encompassing moral support and encouragement. To my dear parents for their unending confidence and support, and for their enthusiasm for editing and grammar. And to Josh for most everything. Specifically, all the logistical and moral support, the confident assurance and reassurance, and above all, the infinite patience.
# TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION ........................................................................................................1

References ..................................................................................................................................4

CHAPTER 2: PSYCHOMETRIC COMPARISON OF ENGLISH AND SPANISH VERSIONS OF THE M-CHAT-R ..........................................................................................6

Introduction ................................................................................................................................6

Methods ....................................................................................................................................9

Results ....................................................................................................................................13

Discussion .................................................................................................................................17

References .................................................................................................................................28

Supplementary Materials .........................................................................................................32

Supplementary Materials References .......................................................................................40

CHAPTER 3: TRANSLATION AND CULTURAL ADAPTATION OF PARENT-REPORT DEVELOPMENTAL ASSESSMENTS: IMPROVING RIGOR IN METHODOLOGY ..............................................................................42

Introduction ..............................................................................................................................42

Considerations in Equivalence .................................................................................................47

Translation and Cultural Adaptation Guidelines ......................................................................50

A Quality Appraisal Tool ..........................................................................................................67

Final Considerations ...............................................................................................................69

References .................................................................................................................................70
CHAPTER 4: RIGOROUS TRANSLATION AND CULTURAL ADAPTATION OF AN AUTISM SCREENING TOOL: FIRST YEARS INVENTORY AS A CASE STUDY

Introduction.................................................................................................................75
Methods..........................................................................................................................80
Results..............................................................................................................................86
Discussion.........................................................................................................................100
References.........................................................................................................................110
LIST OF TABLES AND FIGURES

Table 2.1. Respondent Demographics ............................................................. 13
Table 2.2. Risk Score and Missingness Comparisons........................................ 14
Table 2.3. Itemized Risk Score and Missingness Comparisons............................... 16
Figure 3.1 Representation of the translation and cultural adaptation process........... 52
Table 3.1 Quality Appraisal Tool for Translated and Adapted Caregiver-Report
Developmental Instruments.............................................................................. 68
Table 4.1 Patterns of challenges encountered during the translation and
-cultural adaptation process............................................................................ 88
Table 4.2 Frequency of responses for each response option selected by
participants who were presented with either 3 or 5 labeled response options............. 92
LIST OF ABBREVIATIONS

ASD  Autism Spectrum Disorder
ES   English Speaking
FYI  First Years Inventory
M-CHAT Modified Checklist for Autism in Toddlers
M-CHAT-R Modified Checklist for Autism in Toddlers – Revised (parent-report only)
M-CHAT-R/F Modified Checklist for Autism in Toddlers – Revised with Follow Up
SS   Spanish Speaking
SWH  Spanish Western-Hemisphere
CHAPTER 1: INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder impacting social communicative development and restrictive and repetitive behaviors. ASD affects around 1% of the global population (Elsabbagh et al., 2012). Many parent-report ASD screening tools have been developed with the goal of identifying children who are showing early behavioral symptoms of ASD so that they can receive appropriate referrals for evaluation and treatment (Aldridge et al., 2012; Hampton & Strand, 2015; Reznick et al., 2007; Robins et al., 2014; Wiggins et al., 2007). While most of these tools were developed in English, they are frequently translated into other languages to meet the needs of culturally and linguistically diverse global populations (Al Maskari et al., 2018). The traditional methodological approach to these translations is the “forward-back approach.” In the forward-back approach, one person translates the instrument into the new target language, then a second person translates the new version back into the original language. Finally, this back-translated version is compared to the same-language original version and revisions are made to achieve adequate equivalence between versions according to the authors (Soto et al., 2015). There are no universally accepted guidelines for the translators’ background or knowledge, nor are there specific standards for how to compare the two language versions. The translated screening tool does not typically undergo psychometric testing prior to clinical use (Al Maskari et al., 2018; El-Behadli et al., 2015; Soto et al., 2015). Psychometric testing includes examination of an instrument’s statistical properties, such as its reliability, validity (including concurrent, diagnostic, or predictive validity), factor structure,
cutoff points, and norms. Knowing these properties allows clinicians to make appropriate referral decisions and provide families with useful information.

**Cultural Influence**

Culture impacts the way people understand and answer questions (Beaton et al., 2000). The way parents conceptualize their own child’s behaviors, including expectations for timing of developmental milestones and for appropriate child behaviors, is also influenced by culture (van Kleeck, 1994). Additionally, culture often dictates norms around social communication and social interaction, which are core deficit areas in ASD (van Kleeck, 1994). Thus, cultural perspectives around child development and child rearing are likely to influence parents' understanding of questions and response patterns within the context of ASD screening. Yet, the forward-back approach to translation does not commonly account for cultural differences between the two language populations (Al Maskari et al., 2018; El-Behadli et al., 2015; Soto et al., 2015). Further, psychometric properties may differ across cultures even on culturally adapted measures deemed equivalent on all content. Therefore, re-assessment of local norms, cutoffs, reliability, and validity is required after tools are adapted (Beaton et al., 2000). Professionals in the fields of ASD, early child development, and developmental disorders are beginning to recognize this concern. A small number of research teams have begun to incorporate cultural adaptation methods that are used more widely in other fields (Cuesta-Gomez et al., 2016; Grinker et al., 2015; Kayihan et al., 2015; Magán-Maganto et al., 2018). This alternative approach incorporates rounds of quality checking for initial drafts of the translation as well as pre-testing of the instrument with target population members (DuBay & Watson, 2019). However, the vast majority of ASD screening tool translations continue to be implemented using a traditional forward-back approach (Al Maskari et al., 2018; Soto et al., 2015).
Aims

The overall aim of this dissertation is to examine the process of translating and culturally adapting parent-report ASD screening tools so that these tools can be used effectively in a linguistically and culturally diverse global society. First, an empirical paper psychometrically compares parent-report data from the Spanish and English versions of the Modified Checklist for Autism in Toddlers – Revised (Robins et al., 2014) to assess whether psychometric equivalence is maintained following the forward-back translation of an ASD screening tool. A methodology paper then reviews literature from a broad range of fields related to translation and cultural adaptation methods and presents a comprehensive methodological framework for translating parent-report screening tools. Finally, a clinical paper implements this rigorous translation and cultural adaptation methodology using the First Years Inventory v 3.1 (Baranek et al., 2013) as a test case to illustrate the translation and adaptation process. This process paper also highlights specific areas of the instrument that presented the most challenges and required the most adaptation during the process to inform future projects.
REFERENCES


CHAPTER 2: PSYCHOMETRIC COMPARISON OF ENGLISH AND SPANISH VERSIONS OF THE M-CHAT-R

Early identification of autism spectrum disorder (ASD) is of global concern, due to its estimated global prevalence of 1% and associated impairments in social communication and restrictive/repetitive behaviors (Elsabbagh et al., 2012). Consequently, widespread screening efforts have been made to identify children showing early signs of ASD. If identified at a young age, these children may be able to receive early intervention and potentially reduce the cascading impact of neurodevelopmental processes (Dawson, 2008). The American Academy of Pediatrics recommends that all children are screened twice for ASD between 18 and 24 months of age (Johnson & Myers, 2007). A large number of ASD screening tools have been developed to heed these screening recommendations, the vast majority of which utilize a parent-report questionnaire format (Thabtah & Peebles, 2019). While most of these tools were developed in English-speaking countries, many are being explored internationally as well. At a minimum, these tools require translation in order to be used in linguistically diverse societies, both internationally and within majority English-speaking countries. Traditional translation methods, known as the forward-back approach, include a forward translation made by one translator, a back translation by a separate translator, and comparison of the two same language versions (Peña, 2007). In practice, this method does not typically include a separate examination of psychometric properties before clinical or research use (El-Behadli et al., 2015; Soto et al., 2015; Wild et al., 2005). Rather, the psychometric properties, such as reliability, diagnostic or
predictive validity, factor structure, and norms, are assumed to be equivalent to the original instrument. In the fields of early child development and developmental disabilities, a small body of recent literature recommends an alternative cultural adaptation approach to these translations, including quality checking rounds and pre-testing with target population members, although traditional methods continue to be the most commonly used approach (Al Maskari et al., 2018; DuBay & Watson, 2019; El-Behadli et al., 2015; Soto et al., 2015).

**M-CHAT-R/F**

The *Modified Checklist for Autism in Toddlers – Revised with Follow-Up* (M-CHAT-R/F; Robins et al., 2014) is an ASD screening tool used widely in the US and internationally. It was originally developed from the nine-item *Checklist for Autism in Toddlers* (CHAT), which derived items from behaviors present in 18-month old toddlers who later developed ASD (Baron-Cohen et al., 1991). In its first version, the *Modified Checklist for Autism in Toddlers* (M-CHAT; Robins et al., 2001), added 14 items to the CHAT and expanded the age range to 18 to 30 months. Later, the M-CHAT-R/F (Robins et al., 2014) removed 3 items with poor diagnostic predictability, simplified language, added examples within items, and included a structured follow-up interview to minimize false positives. The multi-site validation study of the M-CHAT-R/F showed a sensitivity of .911, a specificity of .955, and a positive predictive value of .138 when using the parent-report questionnaire only among 15,612 toddlers attending well-checkup visits. Adding the follow-up interview increased specificity to .993 and positive predictive value to .475 (Robins et al., 2014).

The M-CHAT-R/F website (www.mchatscreen.com/mchat-rf/translations; retrieved February 7, 2020) has at least 65 translations available for download, 7 of which are in Spanish. The “Spanish – Western Hemisphere” translation is the version intended for use in the US. This
version was translated specifically for Latin American Spanish-speaking populations. The traditional forward-back approach was used, with one forward translator, one back translator, and additional translators for same-language comparison and revisions. Translators were recruited from a range of Latin American countries in order to represent multiple Spanish dialects (D. Robins, personal communication, Oct. 4, 2019). To date, there have been no published direct assessments of this Spanish translation’s validity or diagnostic predictability, despite its common clinical use in the US. Western-Hemisphere Spanish versions of the M-CHAT and M-CHAT-R/F are also commonly used in research, including studies that pool data from both Spanish and English versions (Campbell et al., 2017; Chlebowski et al., 2013). However, there is no empirical evidence to examine whether or not the English and Spanish versions should be treated as equivalent measures. Two studies have psychometrically analyzed data from the Western-Hemisphere Spanish translation of the M-CHAT (Robins et al. 2001), both revealing relatively poor psychometric properties, including lower internal consistency (Scarpa et al., 2013), lower specificity and positive-predictive values, and higher screen positive rates (Guthrie et al., 2019) in comparison to data from the original English version of the M-CHAT.

**Study Aim**

This study sought to compare the psychometric properties of the original English M-CHAT-R/F to the “Spanish – Western Hemisphere” (SWH) version. This study evaluated only the initial 20 parent-report items, without data from the follow-up interview; thus, this report will use the acronym M-CHAT-R to reflect the use of parent-report data only. We compared risk scores, variability, initial screen positive rates, and missingness across language groups using data collected from a community sample of caregivers who either took the M-CHAT-R in
Spanish or in English. Our aim was to examine the psychometric properties of a parent-report developmental screening tool that was translated using traditional methods. These analyses will also provide empirical evidence of the clinical performance of the SWH M-CHAT-R within a US-based Spanish-speaking population.

**Methods**

A retrospective medical chart review design was used to collect M-CHAT-R records. Records were collected through a university healthcare system’s data warehouse of electronic medical records as well as directly from nine pediatric practices and public health clinics. A small number of M-CHAT-Rs were collected via in-person recruitment at local area festivals (n = 17). A total of 3,724 records were received and reviewed. Records were excluded if they were from patients outside of the 16- to 30-month age range or had missing age data (n = 270), were not in English or Spanish (n = 53), or used the older M-CHAT (n = 46). Current guidelines recommend repeat administrations as a child develops (Johnson & Myers, 2007). Consequently, we identified multiple M-CHAT-R records for 381 patients in the data warehouse. In these cases, one record for each patient was randomly selected to remain in the dataset, while others were excluded. As opposed to the data warehouse, data collected directly from medical practices or at festivals did not contain identifying information; therefore, we were not able to link individual patient data across multiple administrations in this subset. Data collected directly from medical practices was collected over a shorter period than records in the data warehouse. Therefore, likelihood of repeat M-CHAT-R administrations was lower in the subset of directly collected records, but was still possible in some cases. Nonetheless, all records collected directly from medical practices that otherwise met inclusion criteria were retained for analyses.
Records were grouped by language of M-CHAT-R. English-speaking (ES) participants are those who completed the original M-CHAT-R (in English) and Spanish-speaking (SS) participants are those who completed the SWH version of the M-CHAT-R (in Spanish). SS participants had children who were statistically significantly older than ES participants by two weeks on average ($p < .05$). Child gender was approximately evenly split in both language groups. Respondents self-reported their relation to the child in 76% of the records. Of these, 83% reported “mother,” 14% reported “father,” and 2% reported “grandmother.” The remaining caregivers reported that they were foster parents, step-parents, legal guardians, an aunt, a great grandmother, and a social worker. See Table 2.1 for additional demographic information.

Only data from the respondent’s independent completion of the 20 parent-report items of the M-CHAT-R were collected for this study, as information from follow-up interviews was not available. Therefore, initial screen positive rates, defined as total scores > 2 on the parent-report form, were used in analyses instead of actual screen positive rates. Initial screen positive rates for the entire sample of children was 6.0%, including 3.7% of girls and 8.0% of boys.

**Measures**

*The Modified Checklist for Autism in Toddlers – Revised with Follow-Up* (M-CHAT-R/F; (Robins et al., 2014) is a screening tool designed to identify children 16-30 months of age who show symptoms of autism. The screening instrument includes 20 items with dichotomous response options (“yes” or “no”). Many items provide descriptions or examples to aid in understanding. For example, the first item reads, “If you point at something across the room, does your child look at it? (FOR EXAMPLE, if you point at a toy or an animal, does your child look at the toy or animal?)” All items for the original M-CHAT-R/F can be accessed here: https://www.cpqcc.org/sites/default/files/M-CHAT-R_F_1.pdf. The SWH version can be
The M-CHAT-R/F uses a summated scoring algorithm, where the number of item responses reflecting risk ("no" for all items except 2, 5, and 12, for which the risk response is "yes") are added together for the total score. Children are considered "low-risk" if their score is 0-2, "medium-risk" if their score is 3-7, and "high-risk" if their score is 8-20. For scores in the "medium-risk" range, there is a set of follow-up questions to be administered in an interview format to clarify the informant’s at-risk responses. For scores in the “high-risk” range, or for scores that remain 2 or higher after the follow-up interview, an evaluation and referral for intervention are recommended. In the absence of follow-up interview data, initial screen positive rates of > 2 were retained for analysis in this study.

**Data Analysis**

All statistical analyses were completed in R (R Core Team, 2013). Two main approaches were employed to determine whether caregivers responded to the items on the M-CHAT-R in the same way across languages. The first approach was to compare risk-score means, initial screen positive rates, and variability between groups. Prevalence of autism, and symptoms associated with autism, are theoretically assumed to be equivalent across populations, although no data exist to prove this hypothesis (Bernier et al., 2010; Zaroff & Uhm, 2012). If this assumption is accurate, then total risk-score means, initial screen positive rates, individual item fail rates, and floor-rates (i.e. proportion of participants indicating no ASD risk on any item) should be similar across groups. A significant difference in these psychometric properties may suggest possible differences in the way caregivers are interpreting and responding to items in the scale. This first analytical approach will be referred to as the “risk-score analyses.”
The second approach was an analysis of data quality, specifically the frequency of items respondents left blank (i.e.; missingness). Significantly more missingness for an individual item in one language group would suggest that that language group may have had difficulty in understanding or responding to that item. Significantly more missingness overall in one language group would suggest more difficulty understanding or responding to the scale as a whole. To answer these questions, we compared the total number of items left blank per participant between language groups. To examine individual item differences, we compared the proportion of missingness per item between language groups.

A combination of paper and electronic records was reviewed. Electronic records contained no missingness due to forced-choice settings on electronic records. Paper copies of records, however, allowed for items left blank to be retained for our review. Therefore, “missingness analyses” were performed with the subset of the sample containing only paper copies of records. This subset contained 745 English records and 279 Spanish records.

Additionally, two of the medical practices where data were being collected directly also contributed their records to the university data warehouse. That is, we likely received paper copies and electronic copies of the same records from these two practices. Only electronic records from these two practices were retained for risk-score analyses. Only paper copies from these two practices were retained for missingness analyses. The final subset for the risk-score analyses contained 1,964 English records and 627 Spanish records.
Table 2.1. Respondent Demographics

<table>
<thead>
<tr>
<th></th>
<th>For Risk Score Analyses</th>
<th>For Missingness Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Spanish</td>
</tr>
<tr>
<td>n</td>
<td>1,964 (75.8%)</td>
<td>627 (24.2%)</td>
</tr>
<tr>
<td>Child age</td>
<td>21.7 months (sd = 3.4)</td>
<td>22.1 months (sd = 3.5)</td>
</tr>
<tr>
<td></td>
<td>Range 16-30</td>
<td>Range 16-30</td>
</tr>
<tr>
<td>Child sex</td>
<td>Boys / Girls</td>
<td>50.4% / 49.6%</td>
</tr>
</tbody>
</table>

Results

Risk-Score Analyses

Each participant was assigned a total risk score by summing the number of at-risk responses across all items. The lowest possible risk score is 0, indicating low likelihood of ASD, and the highest possible score is 20, indicating high likelihood of ASD. Because data came from community-based screening, most children scored at low risk, resulting in a significant negative skew of the risk score data. To address this skew, we employed resampling methods using a bootstrapped t-test and a Welch modification to correct for unequal variances between groups (see Table 2.2). Results indicated that SS respondents had significantly higher total risk scores than ES respondents ($t = 4.01, p < .001$). We also completed a post hoc negative binomial regression to determine if any demographic factors influenced this difference. SS respondents were 1.4 times more likely to have a higher risk score than ES respondents when controlling for the respondent’s relation to the child, the respondent’s gender, the child’s sex, and the child’s age ($p < .001$).
Table 2.2. Risk Score and Missingness Comparisons

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD)</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Risk Score</td>
<td>English</td>
<td>0.62 (1.3)</td>
<td>4.01</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>0.85 (1.5)</td>
<td></td>
</tr>
<tr>
<td>Total Missingness</td>
<td>English</td>
<td>0.16 (0.5)</td>
<td>5.05</td>
</tr>
<tr>
<td>Score</td>
<td>Spanish</td>
<td>0.57 (1.4)</td>
<td></td>
</tr>
</tbody>
</table>

Floor rates are the proportion of records with a risk score of 0, where the respondent does not endorse any at-risk responses. The floor rate was 66.1% among ES respondents and 56.0% among SS respondents. A $\chi^2$ test revealed that ES respondents had a significantly higher floor rate than SS respondents ($\chi^2 = 20.8, p < .001$).

Screen positive rates are the proportion of records with risk scores above the cut-off point for risk level determinations. Because data from follow-up interviews was not available, we used $> 2$ as the determination for this initial screen positive cut-off. The proportion of records with total risk scores $> 2$ was 7.9% for SS respondents and 5.5% for ES respondents. A $\chi^2$ test revealed that ES respondents had a significantly lower initial screen positive rate than SS respondents ($\chi^2 = 4.4, p < .05$).

**Itemized Risk-Score Analyses**

Itemized analyses were conducted to determine which individual items were driving differences in risk scores between language groups. Fisher’s Exact Tests were performed on 2 x 2 contingency tables containing frequencies of at-risk versus not-at-risk responses for each language group. Table 3 contains these itemized results. False Discovery Rates were used to correct for multiple analyses. Six items remained with significant differences in proportions of
at-risk responses, specifically items 2, 3, 4, 5, 17, and 18 (see Table 2.3). For each of these items except item 17, SS respondents were more likely to endorse the at-risk response. For item 17, SS respondents were less likely to endorse the at-risk response.
Table 2.3. Itemized Risk Score and Missingness Comparisons

<table>
<thead>
<tr>
<th>Item</th>
<th>At-Risk Response Comparisons</th>
<th>Missing Response Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% at-risk response (English/Spanish)</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>1. Point following</td>
<td>1.0 / 2.1</td>
<td>2.07</td>
</tr>
<tr>
<td>2. Wondering hearing</td>
<td>1.5 / 4.0</td>
<td>2.65</td>
</tr>
<tr>
<td>3. Pretend play</td>
<td>4.4 / 13.0</td>
<td>3.30</td>
</tr>
<tr>
<td>4. Climbing</td>
<td>0.4 / 1.4</td>
<td>4.08</td>
</tr>
<tr>
<td>5. Finger movements</td>
<td>10.6 / 17.7</td>
<td>1.75</td>
</tr>
<tr>
<td>6. Imperative pointing</td>
<td>3.2 / 4.2</td>
<td>1.34</td>
</tr>
<tr>
<td>7. Declarative pointing</td>
<td>3.2 / 4.2</td>
<td>1.33</td>
</tr>
<tr>
<td>8. Interest in other children</td>
<td>1.9 / 2.7</td>
<td>1.45</td>
</tr>
<tr>
<td>9. Show object</td>
<td>1.4 / 1.6</td>
<td>1.13</td>
</tr>
<tr>
<td>10. Response to name</td>
<td>0.6 / 1.4</td>
<td>2.59</td>
</tr>
<tr>
<td>11. Social smile</td>
<td>0.5 / 0.6</td>
<td>1.26</td>
</tr>
<tr>
<td>12. Noise sensitivity</td>
<td>13.0 / 13.7</td>
<td>1.05</td>
</tr>
<tr>
<td>13. Walking</td>
<td>0.7 / 1.0</td>
<td>1.46</td>
</tr>
<tr>
<td>14. Eye contact</td>
<td>0.9 / 1.3</td>
<td>1.50</td>
</tr>
<tr>
<td>15. Imitation</td>
<td>0.7 / 1.0</td>
<td>1.38</td>
</tr>
<tr>
<td>16. Gaze-following</td>
<td>4.2 / 3.7</td>
<td>0.88</td>
</tr>
<tr>
<td>17. Gaining parents’ attention</td>
<td>7.8 / 3.7</td>
<td>0.46</td>
</tr>
<tr>
<td>18. Understands what is said</td>
<td>2.8 / 5.0</td>
<td>1.82</td>
</tr>
<tr>
<td>19. Social reference</td>
<td>3.6 / 4.5</td>
<td>1.26</td>
</tr>
<tr>
<td>20. Being swung</td>
<td>0.3 / 0.6</td>
<td>2.52</td>
</tr>
</tbody>
</table>

*Remained significant after false discovery rate reduction for multiple analyses. See https://www.cpqcc.org/sites/default/files/M-CHAT-R_F_1.pdf (English), and https://mchatscreen.com/wp-content/uploads/2015/05/M-CHAT-R_F_Western_Spanish.pdf (Spanish) for full text of each item.
**Missingness Analyses**

We created a new variable to represent total missingness by summing the number of items that the respondent selected either no response option or both response options. Similar to total risk scores, total missingness data were also negatively skewed with unequal variances between groups. Therefore, a bootstrapped t-test with a Welch modification was also used to compare total missingness scores between language groups (see Table 2.2). Results indicated that SS respondents left more items blank than ES respondents ($t = 5.05, p < .001$). In order to determine if any demographic factors influenced this result, we completed an additional post hoc negative binomial regression. SS respondents were 3.6 times more likely to leave more items blank than ES respondents while controlling for the respondents’ relation to the child, the informant’s gender, the child’s sex, and the child’s age ($p < .001$).

**Item-by-Item Missingness Analysis**

Similar to itemized analyses of risk scores, Fisher’s Exact Tests were performed on 2 x 2 contingency tables containing frequencies of missing versus not missing responses per item for each language group. False Discovery Rates were also used to correct for multiple comparisons. SS were more likely to contain missing data for items 2, 3, 5, 12, 13, 15, and 17 than ES respondents (see Table 2.3).

**Discussion**

This psychometric comparison between the original English M-CHAT-R and the SWH translation revealed that respondents who took SWH version endorsed a greater number of ASD symptoms, were less likely to report no ASD symptoms, and were more likely to leave items blank. Additionally, children whose caregivers took the SWH M-CHAT-R were more likely to screen positive initially than children whose caregivers took the M-CHAT-R in English. If we
assume that the overall actual prevalence of ASD symptoms is the same within the two language populations tested here, these results suggest that the M-CHAT-R SWH translation does not perform the same way as the original English M-CHAT-R. These observed differences possibly reflect the influence of factors that were not addressed adequately in the traditional translation procedures.

**Differences in Risk Scores**

Other studies of forward-back translations of the M-CHAT family (M-CHAT or M-CHAT-R) have found an increased likelihood of higher risk scores and initial screen positive rates among respondents using translated versions. For forward-back Spanish translations of the M-CHAT, two studies have reported higher overall risk scores as well as higher at-risk endorsement rates on specific items (Kimple et al., 2014; Windham et al., 2014). Unexpectedly high at-risk endorsement rates (>40%) among low-risk samples were also documented on two items of the M-CHAT-R for a forward-back Chinese translation (Guo et al., 2019) and a forward-back Albanian translation (Brennan et al., 2016). In a forward-back Arabic translation for an Egyptian context in a low-risk sample, Mohamed et al. (2016) observed extremely high endorsement rates (>20%) for over half of items.

Similar to our findings, studies that have used a traditional forward-back translation approach for the M-CHAT family of screeners in low-risk samples also frequently have reported unusually high initial screen positive rates (scoring > 2 prior to the follow-up interview), as compared to original M-CHAT validation studies. For the M-CHAT, initial screen positive rates were reported to be 26.4% in South Korea with a Korean version (Seung et al., 2015), 23.8% in Egypt with an Arabic version (Mohamed et al., 2016), 49.2% in Turkey with a Turkish version (Kara et al., 2014), and 22% in Chile with a Spanish version (Eugenin et al., 2015). Similarly,
multiple studies have reported high initial screen positive rates for the SWH translation of the M-CHAT, including 15.0% (Guthrie et al., 2019), and 30.3% (Windham et al., 2014). For the M-CHAT-R, initial screen positive rates were reported to be 14.4% in China with a Mandarin translation (Guo et al., 2019), 9.75% in Albania with an Albanian translation (Brennan et al., 2016), and 25% in Indonesia with an Indonesian translation (Windiani et al., 2016). Comparatively, original English M-CHAT and M-CHAT-R validation studies reported initial screen positive rates of 9.1% (Chlebowski et al., 2013) and 7.2% (Robins et al., 2014), respectively, in low-risk community-based samples.

Alternatively, Magán-Maganto et al. (2018) incorporated quality checking and cultural adaptation procedures when translating the M-CHAT-R for a Spanish context rather than using a strictly traditional approach. Their adaptation resulted in what appeared to be typical endorsement rates of items (respondents selected at-risk responses less than 4% of the time on 18 items), but a lower initial screen positive rate (2.3%). Cuesta-Gomez et al. (2016) also used a cultural adaptation approach for the Spanish M-CHAT in Chile and reported a 4.3% initial screen positive rate. In India, TS et al. (2018) used a rigorous cultural adaptation approach for the M-CHAT-R and achieved a reasonable initial screen positive rate (5.5%), although they did observe an unexpectedly high endorsement rate for one item (35.3%). Improved psychometric characteristics from these studies may suggest that a more rigorous translation and cultural adaptation approach could yield more linguistically and culturally appropriate measures.

**Differences in Missingness**

In the current study, SS respondents were also more likely to leave an item blank as compared to ES respondents. Large portions of such missing data may be an indication that respondents do not understand a particular item, do not understand the instructions, or are not
able to choose between the answer choices provided (Ware & Gandek, 1998; Ware et al., 1996, 1995). Missing data may also be an indication that the linguistic complexity of the tool does not align with the literacy levels of the population sampled (Ware et al., 1995). Because this missingness pattern was more frequent among the SS respondents, these respondents may have experienced difficulty understanding the translation of the instrument. Amount of missing data is not typically reported in M-CHAT translation studies. However, one research team in India reported some missing responses on three items but otherwise good psychometric properties after translating the M-CHAT-R into Malayalam using a rigorous cultural adaptation approach (TS et al., 2018).

**Item-Specific Differences**

Items 2, 3, 5, and 17 showed significantly different response patterns between groups, where SS respondents were more likely to leave items blank and were more likely to endorse the at-risk response, with the exception of 17, which showed the opposite trend in endorsement rates. There were two additional items (4 and 18) where SS caregivers were more likely to report at-risk responses. There were three additional items (12, 13, and 15) that SS respondents were more likely to leave blank. Based on itemized results from other translations in the M-CHAT family, items 2, 4, 5, and 12 frequently present problems in endorsement rates and diagnostic discrimination (Brennan et al., 2016; Carakovac et al., 2016; Guo et al., 2019; Mohamed et al., 2016; TS et al., 2018; Tsai et al., 2019; Wong et al., 2018), including Spanish translations (Canal-Bedia et al., 2011; Eugenin et al., 2015; Kimple et al., 2014; Magán-Maganto et al., 2018). Some difficulty has been reported for items 3, 15, 17, and 18 (Canal-Bedia et al., 2011; Eldin et al., 2008; Kara et al., 2014; Kimple et al., 2014; TS et al., 2018), but no translation studies have reported difficulties with item 13.
We have identified some patterns across the items observed to show differential responses between groups in this study. While we describe these patterns below, a complete linguistic and cultural analysis of each of these items, along with possible rationales for observed differences, can be found in the Supplementary Materials for this article.

Several linguistic differences between the English and SWH versions were identified within these problematic items. Some of these differences consisted of grammatical or semantic variances that may have differentially influenced respondents’ interpretation of items. For example, item 5 uses the term “wiggle” in English, while the SWH version reads “mueve/move.” This verb lacks the unique characteristic of the movement that is present in the English version. Additionally, the SWH version does not contain the underlined font for the word “unusual,” as it does in English. These differences may have made it more likely for those reading the SWH version to endorse “yes,” the at-risk response. Similarly, item 15 contains the phrase “copy what you do,” which is translated to “imitar sus movimientos/copy your movements.” “Movements” is more limited in meaning as compared to “what you do,” and fails to encompass the non-movement example provided for this item (i.e. sounds). These changes, and other similar changes within items that showed differential responses between groups, may have been chosen to simplify the grammar or complexity of the SWH version; however, they appear to have influenced interpretation differentially across language groups.

Differences between the English and SWH versions were also noted in the area of specific connotations of words. For example, item 3 uses the terms “fantasía” and “imaginación” to describe the concepts “play pretend” and “make-believe.” These Spanish terms are not frequently used in reference to children’s play, and in addition to their English cognates “fantasy” and “imagination,” they can contain the connotations of “romantic intimacy” and
“mental-health related delusion,” respectively. Item 12 translates the adjective “everyday” as “cotidiano,” which can contain a connotation more closely aligned with the English terms “customary” or “traditional” as opposed to the intended English connotation “ordinary.” These unintended connotations, along with others present in other items that showed differential responses between groups, may have influenced Spanish speakers’ interpretations differently than English speakers’, leading to differences in endorsement and missingness rates. McClure et al. (2018) noted an unintended connotation in their Nepali translation of the M-CHAT-R which resulted in inaccurate responses and required modification.

There were several instances where the grammar of the Spanish language may have complicated respondents’ interpretation. Verb tense and possessive terms for the third person and formal second person are equivalent in Spanish. In some items, the pronoun is not provided, requiring the reader to interpret whether a phrase refers to the child, the adult, or an object. For example, item 15 could be interpreted as “Does your child try to imitate his/her movements?” instead of “your movements.” These ambiguities in intent may have confused respondents, leading to their increased likelihood to leave these items blank. In a Spanish translation study of a similar ASD screening tool, DuBay, Watson, Lee, et al. (2019) also identified the need to repeat nouns or pronouns in this context to avoid subject misinterpretations.

Each of the reverse-scored items (2, 5, 12) showed differential responses between groups. DuBay, Watson, Lee, et al. (2019) found that respondents often exhibited a positive response bias in which Spanish-speakers assumed all items described typical behaviors. A similar pattern may have caused more SS respondents in this study to choose “yes,” which was the at-risk response for these items. Additionally, some SS respondents may have recognized that the behaviors in these items sounded atypical, but hesitated to answer “no” because of this positive
response bias, increasing their likelihood of leaving the item blank. In fact, these three items typically present problems in most M-CHAT family translation studies, with unexpectedly high endorsement rates for each item (Brennan et al., 2016; Guo et al., 2019; Wong et al., 2018), including Spanish translations (Eugenin et al., 2015; Kimple et al., 2014; Magán-Maganto et al., 2018; Windham et al., 2014). TS et al. (2018) also noted high missingness on each of these three items.

Finally, there may have been some additional cultural factors that differentially influenced response to items. For example, item 4 asks whether the child likes “climbing on things” and includes furniture as an example. Literature suggests that Latin American families hold high behavioral compliance expectations of children (Calzada et al., 2010; Halgunseth et al., 2006). This example may have been interpreted as a non-compliant behavior. SS respondents may have viewed “no” as the more socially appropriate response as compared to ES respondents, increasing their likelihood of choosing this response. Additionally, item 2 asks whether the adult has wondered if their child is deaf. The SWH translation reads “si su hijo es sordo/if your child is deaf.” There is a common expression in Spanish using the term “sordo” to refer to listening compliance. Consequently, SS respondents may have interpreted the item as asking “Have you ever wondered if your child is selectively listening?” These factors, and other similar cultural factors, may significantly change the conceptual meaning of items in this new cultural context, possibly leading to differential responses across groups.

Similar to previous studies of Spanish translations of the M-CHAT family, no apparent problems existed for items 1, 6, 7, 8, 9, 10, 11 (Canal-Bedia et al., 2011; Kimple et al., 2014; Magán-Maganto et al., 2018; Windham et al., 2014). Some of these studies did note high
endorsement rates, poor diagnostic discrimination, or initial need to adapt items 16, 19, and 20, although results here did not indicate problems with these items.

**Clinical Implications**

Differences in the initial screen positive rate between the two language groups, while significant, were lower than previous studies of the SWH version of the M-CHAT. In the current study, 55 of 1,000 children screened positive if their caregiver took the M-CHAT-R in English, where 79 of 1,000 children screened positive if their caregiver took the M-CHAT-R in Spanish. Clinically, our results suggest a potential increased false positive rate in the SS population, a hypothesis that appears to be supported by one recent study. Guthrie et al. (2019) found that specificity and PPV, both of which are lowered by higher rates of false positives, were significantly lower in a SS group as compared to an ES group in a community-based study of the M-CHAT. If the false positive rate is indeed higher in this population, this would provide further support for using the follow-up interview, already deemed necessary for all children who score between 3 and 7. Notably, there is some research to suggest that physicians may not frequently perform follow-up interview procedures as recommended (Campbell et al., 2017). The follow-up interview has been shown to identify false positives from the initial screen so that these are changed to accurate screen negatives (Robins et al., 2014). Theoretically, the follow-up interview could correct an inflated false positive trend in this population by identifying low-risk children among these false positive screens as professionals provide examples of concepts and clarify the meaning of items with caregivers. However, it should be noted that the follow-up interview for the M-CHAT-R/F was translated using the same forward-back approach as the initial parent-report items. Therefore, the follow-up interview could also contain similar linguistic or cultural influences. Nonetheless, the use of multiple examples, and the presence of a professional who
may be able to clarify item meaning may help to identify false positive cases and avoid unnecessary referral.

**Limitations**

We were unable to access data from follow-up interviews for this study. It is possible that the follow-up interviews may have corrected the potentially greater number of false positive initial screens. We were also unable to access diagnostic outcomes for those who went on to receive ASD evaluations. Therefore, we do not know the actual number of true and false positives and negatives to examine differences in diagnostic validity between the language versions. We were also unable to identify repeat administrations of the M-CHAT-R within data collected directly from medical practices or festivals; thus, we may have included a few repeated administrations. Additionally, we were not able to examine missingness among the data collected electronically.

Due to the lack of data on the demographic characteristics of the sample, we were unable to examine how dialect or country of origin, education or income levels, acculturation levels, or other demographic variables may have influenced response patterns for endorsement rates or missingness. Nonetheless, the sample size and data collection procedures are likely to have resulted in a sample that is representative general population of Spanish speakers in this region of the US.

The conclusion that these results indicate a difference in the way these populations are responding to the M-CHAT-R translations is based on the assumption that the actual prevalence of ASD and ASD symptoms is similar across the two populations. This assumption is generally accepted in the literature, though there is no empirical evidence to support it. Observed prevalence rates of ASD are often lower in SS and Latino/Hispanic populations (Baio et al.,
This observed trend is typically assumed to be explained by healthcare access differences rather than actual prevalence differences (Mandell et al., 2009). If actual prevalence rates reflect current observed prevalence rates, then the results found here would be in the opposite direction of that trend. Specifically, observed prevalence rates would suggest that fewer ASD symptoms should be observed in the SS sample, when in fact SS caregivers reported more ASD symptoms. Therefore, although a difference in actual ASD prevalence cannot be ruled out as an explanation for these results, this is unlikely due to the opposite direction of the trend observed here.

Conclusions/Future Directions

Due to this study and others’ observed differences in initial screen positive rates, floor rates, endorsement rates, and item missingness rates between the original M-CHAT-R (or M-CHAT) and forward-back translations of other languages, it appears that the traditional forward-back translation methodology may not be sufficient to retain similar psychometric properties between language versions of parent-report ASD screening instruments. Linguistic differences between versions and cultural factors may influence caregivers’ understanding of the concepts contained within the items of the instrument. This could lead caregivers with differing cultural backgrounds to systematically respond to items differently when compared to US-based English-speaking caregivers, for whom the original tool was developed. Consequently, these findings support the need to include more rigor in the translation process, to ensure intelligibility in the new population, and to address cultural factors that may influence response to items. At a minimum, we recommend that the M-CHAT-R undergo re-translation of problem items for the SWH version using a rigorous cultural adaptation approach. Future research should also examine
the SWH version of the follow-up interview to determine if interviews are able to reduce false positives.

We wish to emphasize that although the findings presented here do raise concerns regarding the validity of the SWH M-CHAT-R/F in a US-based population, they do not justify discontinuing the use of this tool. A more adequate alternative screening measure has neither been identified nor developed for this population, which is already at risk of delayed or missed diagnosis (Baio et al., 2018; Mandell et al., 2009; Ratto et al., 2016). A new SWH version (“Spanish – Western Hemisphere version 2”) has been added to the M-CHAT-R/F translations website, which contains modifications to some of the problematic items identified here. It is possible that this version has improved psychometric properties; however, similar to the SWH version, there are no known publications to describe its translation process or to support its clinical use. To avoid exacerbating existing health disparities in this area, we should continue to use available Spanish-language screening tools with the SS population in the US, while simultaneously pursuing the development of tools with strong psychometric properties.
REFERENCES


SUPPLEMENTARY MATERIALS

Linguistic and Cultural Analysis of Items Showing Endorsement or Missingness Differences Between Groups

Item 2

Item 2 asks whether the parent has ever wondered if their child might be deaf. Spanish-speaking (SS) respondents were more likely to respond “yes” and to leave the item blank. This is the second item that respondents encounter, and the first that is reverse-scored (i.e. a “yes” response to this item increases the risk score). In a qualitative study of SS respondents’ understanding of similar concepts, caregivers frequently had difficulty understanding such reverse-score items (DuBay, Watson, Lee et al., 2019). They assumed all items described typical behaviors and tended to respond positively to all items. Similarly, previous studies of translations of either version of the M-CHAT have encountered high endorsement rates or other poor performance of reverse-score items (Brennan et al., 2016; Kara et al., 2014; Kimple et al., 2014; Stenberg et al., 2014; TS et al., 2018; Wong et al., 2018). Two translations of the M-CHAT-R/F have documented unusually high endorsement rates for this item, including the Spain Spanish translation (Brennan et al., 2016; Magán-Maganto et al., 2018). One study noted significantly more missing data from this item for the Malayalam translation (TS et al., 2018).

Two slight linguistic differences are noted between the English and Spanish – Western Hemisphere (SWH) versions: 1) To convey the meaning “ever” as in “have you ever wondered,” the SWH version uses the phrase “alguna vez/any time.” This wording utilizes simple Spanish grammar, but the resulting phrase introduces a slightly different meaning, and is not as common as the phrase “have you ever wondered” is in English. Secondly, the English M-CHAT-R version uses “might be deaf” where the SWH version uses “es sordo/is deaf.” There is also a common phrase in the Spanish language that uses the term “sordo” to mean “selective listening,” or
choosing to ignore someone rather than experiencing hearing difficulty. Respondents may have interpreted the item as asking, “Have you ever wondered if your child is selectively listening?” The phrase “es sordo/is deaf” aligns with the syntax used in the common expression, where the inclusion of a qualifier like “might” (tal vez, puede ser, etc.) may have reduced this pattern.

**Item 3**

Item 3 asks about pretend play. SS respondents were more likely to respond “no” and to leave the item blank. Two previous studies have reported less frequency of pretend play behaviors in Latin American toddlers as compared to English Speaking (ES) toddlers from the US (Farver & Howes, 1993; Göncü et al., 2000). It may be that toddlers of SS caregivers in fact exhibit lower trait levels for this item, causing more respondents to answer “no.” If pretend play is less frequent in this population as a whole, the item is likely to be less predictive of autism diagnoses.

Participants in DuBay et al. (2019) experienced difficulty understanding the concept of pretend play. There was no identified direct translation of the term “pretend play,” and many caregivers reported that they understood the concept as imitation of daily activities. The terms “fantasía/fantasy” and “imaginación/imagination,” are also not frequently used in reference to children’s play. When used with adults, they can contain connotations related to “romantic intimacy” and “mental-health related delusion,” respectively. In DuBay et al. (2019), caregivers sometimes interpreted these terms as referring to more advanced dramatic play, such as play including differing roles, when used in a play context. Both of these potential interpretations would conflict with the original intent of the item.

Additionally, “fingir” is used in the SWH translation for the verb “to pretend” in each of the examples in the item. *Fingir* can be more directly translated as “to feign,” which may have
introduced an unintended connotation to the item, causing confusion or misunderstanding on the part of SS respondents.

Interestingly, Canal-Bedia et al. (2011) reported that parents had difficulty understanding their initial translation of the M-CHAT version of this item. Their resulting modification included “juegos imaginativos/imaginative games” for “play pretend,” and “hacer como si/act as if” for “pretend” in the examples, as opposed to the term “fingir.” The Spain Spanish M-CHAT-R/F translation includes the same initial description as the SWH M-CHAT-R/F version, but uses “hacer como que/act as if” for the examples (Magán-Maganto et al., 2018). No other translations of the M-CHAT or M-CHAT-R/F have reported psychometric challenges with this item.

**Item 4**

Item 4 asks whether the child likes to climb on things. SS respondents were more likely to respond “no.” Because “furniture” is the first example provided, SS respondents may initially consider this item to be related to compliant behaviors. There is some literature to suggest that Latin caregivers place high value on behavioral compliance (Calzada et al., 2010; Farver & Howes, 1993). This could have possibly influenced SS respondents’ interpretation of this item such that they would perceive the appropriate social response to this item to be “no.”

Some other translation studies have reported poor psychometrics of this item, both for the M-CHAT-R/F and for the M-CHAT (Guo et al., 2019; Wong et al., 2018), including in Spanish (Albores-Gallo et al., 2012; Magán-Maganto et al., 2018). Fortunately, this item was meant to be a foil item and was not intended predict ASD, which may explain some of its lower psychometric properties. However, the item still contributes to the risk score in a clinical context. If SS respondents are more likely to fail this item, that would unnecessarily contribute to a higher risk score.
**Item 5**

Item 5 asks whether the child makes unusual finger movements near their eyes. SS respondents were more likely to respond “yes” and to leave the item blank. This is the second reverse-scored item that respondents encounter. Similar to findings in item 2, more frequent “yes” responses may be a result of a general positive response bias and lack of reversing that bias for reverse-scored items.

One difference between the two versions is the term “wiggle,” which is translated to “mueve/move,” removing the unique characteristic of the action. Another minor difference between the SWH translation of the M-CHAT-R and the original English version is the lack of underlined font on the word “unusual.” In English, this font likely focuses respondents’ attention to the negative nature of the item, and may cue respondents to reverse their positive response bias. In its absence, SS respondents may perceive this cue less frequently. However, the phrase “de manera inusual/in an unusual way” is added to the SWH version at the end of the example. Theoretically, this change would cue respondents to reverse their positive response bias, though the results here indicate it the addition may not have been sufficient. According to qualitative data of the Chilean Spanish translation, which uses slightly different phrasing, caregivers specifically reported difficulty understanding the movement described in the item (Eugenin et al., 2015). Caregivers presented with a Nepali translation of this item also reported difficulty understanding its intent (McClure et al., 2018).

This item is frequently found to result in poor psychometric properties in translations of the M-CHAT and M-CHAT-R/F (Brennan et al., 2016; Carakovac et al., 2016; Guo et al., 2019; Mohamed et al., 2016; TS et al., 2018; Tsai et al., 2019; Wong et al., 2018), including Spanish
translations (Canal-Bedia et al., 2011; Eugenin et al., 2015; Kimple et al., 2014; Magán-Maganto et al., 2018).

**Item 12**

Item 12 asks about noise sensitivity. SS respondents were more likely to leave this item blank. This is the third and final reverse-scored item, possibly causing confusion among the respondents. Additionally, the adjective “everyday” is translated as “cotidiano,” which directly translates to “quotidian,” but has connotations more closely related to “customary” or “traditional.” SS respondents may have had difficulty interpreting the concept of “ruidos cotidianos” if they understood the term as “customary noises” or “traditional noises.”

This item does consistently prove problematic in translation studies, both for the M-CHAT and M-CHAT-R/F versions. Specifically, it is frequently endorsed by a large proportion of respondents and does not differentiate well between children with and without ASD (Brennan et al., 2016; Cuesta-Gomez et al., 2016; Eugenin et al., 2015; Guo et al., 2019; Kara et al., 2014; Kimple et al., 2014; Magán-Maganto et al., 2018; Mohamed et al., 2016; Stenberg et al., 2014; TS et al., 2018; Tsai et al., 2019; Wong et al., 2018).

**Item 13**

Item 13 asks whether the child is walking. SS respondents were more likely to leave this item blank. This item was translated directly, reflecting the present tense syntax used in the English version. However, the present tense is often used for ongoing activity in Spanish. SS respondents may not have known whether to interpret this wording as “Is your child walking now” or “Does your child walk typically,” which may have resulted in them leaving the item blank more frequently.
Alternatively, this item is only three words long and takes up one line of the questionnaire, whereas most other items contain many more words and take up at least two lines of text. SS respondents may have missed this item due to formatting, although the formatting is similar in the English version. Fortunately, item 13 is another foil item. Because items left blank typically do not increase the risk score in a clinical setting, this difference may not result in major clinical implications.

**Item 15**

Item 15 asks about imitation of adult actions. SS respondents were more likely to leave this item blank. There are two differences between the SWH and English versions. First, “what you do” was changed to “sus movimientos/your movements.” This change maintains general conceptual equivalence and is not likely to contribute to significant misunderstanding; however, the final example in the item is imitation of “sounds,” which are not typically considered to fall in the category of “movements.” Second, the examples in the item each include a verb in the English version, while the SWH includes verbs in the first two examples but no verb in the final example. Literally translated back into English, the examples would read “say goodbye with their hand, clap, or some silly sound that you make.” The lack of parallel structure for these examples may have contributed to confusion among SS respondents. Additionally, the pronoun “sus/your” could refer to the adult’s movements or to the child’s own movements, potentially complicating interpretation further and leading to an increased likelihood of being left blank.

The Spain Spanish version uses the exact same language, but no difficulty was reported for this item in Magán-Maganto et al. (2018). Two forward-back translations of the M-CHAT did report poor diagnostic predictability with this item (Eldin et al., 2008; Kara et al., 2014), though no other problems have been reported for this item in other publications.
Item 17

Item 17 asks whether the child attracts the respondent’s attention for social purposes. SS respondents were more likely than ES respondents both to respond “yes” and to leave the item blank. This was the only item endorsement difference where SS respondents were less likely to endorse the high risk response than ES respondents.

The verb “halagar” is used to describe the concept of getting “praise.” Halagar is a fairly uncommon word in Spanish and may have not been understood by respondents with low literacy levels. Additionally, one connotation of this term is “complement,” which introduces a slightly different conceptual meaning than the English version.

TS et al. (2018) reported difficulty initially translating this item to Malayalam and they found that respondents endorsed the at risk response (“no”) more frequently than expected. Caregivers presented with a Nepali translation also reported difficulty with this item due to its length and complexity (McClure et al., 2018).

Item 18

Item 18 asks about receptive language and uses several simple commands as examples. SS respondents were more likely to respond “no.” The same underlying values of behavioral compliance may have influenced this item as well. SS respondents may have had higher expectations for children complying with parents’ directives than ES respondents. Therefore, a “yes” response may have required a higher threshold level of the latent trait among SS respondents. This item did have higher at-risk endorsement rates in Spanish translations when compared to the English version of the M-CHAT (Kimple et al., 2014).

Additionally, the term “hacer señas/make signals” is the translation used for “point.” The concept of pointing has multiple Spanish translations across dialects. “Hacer señas” can mean to
point, however it can also mean to make gestures of any kind. SS respondents may have interpreted this concept as gesturing in any way to support the child’s understanding, while ES respondents considered gestures other than pointing to be allowed in this context. This pattern of interpretation could make the item less difficult in English, possibly explaining the more frequent “no” response among Spanish records.
REFERENCES


CHAPTER 3: TRANSLATION AND CULTURAL ADAPTATION OF PARENT-REPORT DEVELOPMENTAL ASSESSMENTS: IMPROVING RIGOR IN METHODOLOGY

Parent-report screening and assessment tools are designed to identify children who are showing symptoms associated with various developmental conditions, including autism spectrum disorder (ASD). Diagnosis of these developmental conditions requires knowledge of the way the children function in their everyday life. Because observations in a clinical setting do not always provide a valid representation of a child’s daily functioning, parent report is one of the most common data collection methods for screening and assessment of developmental conditions. Most available tools were created and validated with non-minority English-speaking individuals in western countries such as the United States and England. However, the need to identify children who show signs of developmental difficulties is not unique to these populations. In fact, a resolution of the World Health Organization (WHO, 2014) urged its member states to launch coordinated and comprehensive efforts to build a global capacity to meet the needs of individuals with ASD and other developmental disabilities. Building such capacity requires assessment and screening tools that are valid for use with culturally and linguistically diverse populations.

Because of the resources required to develop new tools, modifying existing tools for use with new populations is often preferred as a more feasible and cost-effective strategy (Guillemin, 1993; Ware, Gandek, & Keller, 1996). Therefore, many tools originally developed in English are

---

being translated into new languages. Traditionally, these self-report tools are translated using a basic forward-back methodology that was originally taken from a wider set of recommendations provided by Brislone (1970). However, simple translation methods are not sufficient to begin using an instrument with a new population (Acquadro et al., 2008; Gjersing, Caplehorn, & Clausen, 2010; Guillemin, 1993). Factors such as translation errors and, more importantly, cultural differences complicate the task of translating an instrument. Traditional methods do not sufficiently address these problems (Beaton et al., 2000; Gjersing et al., 2010; Nelson McDermott & Palchanes, 1994). Culture acts as a lens through which we view the world, including parenting and child development. Cultures often have distinct perspectives around when children learn certain skills, how children should interact with others, and what is valued as appropriate behavior from children (van Kleeck, 1994). Further, culture influences the way people understand and answer questions (Beaton et al., 2000). Cultural perspectives around child rearing will in turn influence parents' understanding of questions as well as their response patterns. Thus, cultural adaptation should be included in the translation process of parent-report tools. Without appropriate translation and adaptation methods, the translated version of an instrument may differ in over- or under-identification or diagnosis rates as compared to the original instrument (Soto et al., 2015), possibly resulting in a poor quality screener or assessment tool.

Self-report instruments are designed such that a layperson independently responds to a series of questions or statements about themselves or a family member. The Modified Checklist for Autism in Toddlers – Revised (M-CHAT_R/F; Robins et al., 2014), the Infant Toddler Checklist (Wetherby & Prizant, 2002), and the Social Communication Questionnaire (Rutter, Bailey & Lord, 2003) are examples of commonly used self-report developmental screening tools
designed to be completed independently by a caregiver. Self-report instruments present a specific challenge in measurement. The data collected is completely dependent on the wording of the items themselves (along with instructions and answer choices). No professional with training, expertise, or experience in the constructs of interests, is available to clarify the meaning of items as parents are answering the questions. Thus, the reliability and validity of these instruments is completely dependent upon wording and administration method. A quality translation and cultural adaptation process is required in order to appropriately modify a self-report instrument for use with a new population (Acquadro et al., 2008; Beaton et al., 2000). Brislon (1970) first introduced a translation methodology involving back-translation, among other procedures, in the 1970s. Back-translation was then frequently adopted as the sole step necessary in the translation process, although Brislon himself specifically cautioned against this practice (1970). In this traditional methodology, one person translates the instrument into the new target language, one person takes this target translation and back-translates it into the source language, then compares the two source-language versions and resolves discrepancies. Research and clinical teams typically begin using the translation once the two source-language versions appear “equivalent,” though most research teams fail to provide a full description in their publications of the way they conceptualize equivalence.

The fields of ASD and developmental disabilities generally continue to use simple translation methods as standard practice. Specifically, minimal quality checking methods are utilized, pre-testing and re-testing of psychometric properties are frequently omitted, and little information about the process is included in peer-reviewed publications. Three recent reviews in the fields of ASD and developmental disabilities illustrate this trend. Soto et al. (2015) and Al Maskari et al. (2018) both reviewed translations and cultural adaptations of ASD screening tools.
published in peer-reviewed journals. They found limited descriptions of the translation process, frequent omissions of important steps in the process, and overall low rigor in translation and cultural adaptation methodology (Soto et al., 2015; Al Maskari, Melville, & Willis, 2018). El-Behadli and colleagues (2015) published an evidence map analyzing the quality of 63 translations of the 9 developmental screening tools recommended by the American Academy of Pediatrics. In this review, there were no supporting peer-reviewed publications for 30% of the translations found. Information on translation methodology for these translations was often collected through personal communications or from instrument manuals. Further, researchers found no information regarding the methods of translation for almost half of the translations. A majority of translations omitted quality checking or pre-testing methods, and 11% specifically stated that culture was not considered during the translation process (El-Behadli et al., 2015). The M-CHAT-R (Robins et al., 2014), is used widely internationally and has at least 60 official translations. The website (www.mchatscreen.com) provides a recommended methodology for translation similar to traditional back-translation methods described earlier.

These findings indicate that a rigorous translation and cultural adaptation process for self-report instruments has not been widely adopted in the fields of ASD and developmental disabilities, although some recent reports suggest that the limitations of simpler, more traditional translation methods are becoming more widely recognized in these fields (Kondolot et al., 2016; Windham et al., 2014). The risk of using an instrument that is poorly adapted to a population’s culture is measurement bias. Specifically, level of risk or impairment in various areas of development, including social communication and repetitive and restrictive behaviors, may be overestimated or underestimated when using a maladapted instrument (Harrison et al., 2017).
This paper seeks to present a set of guidelines written from the perspective of early child development and developmental disabilities, including ASD. While there are numerous guidelines in the literature that describe rigorous translation and cultural adaptation methodologies (see Beaton, 2000; Gjersing et al., 2010; Nelson McDermott & Palchanes, 1994; Sousa & Rojjanasrirat, 2011; Wild, 2005), none highlight the unique experience of parents and other caregivers responding to questions that relate to child behaviors and abilities, a concept that is ripe with cultural meaning. Our field is also in need of specific criteria to assess the quality of the translation and adaptation process for developmental screening and assessment tools. The three reviews described earlier used criteria that were either based only on early works (Soto et al., 2015, El-Behaldi et al., 2015), or designed for intervention adaptation rather than assessment adaptation (Al Maskari et al., 2018). We propose that a quality appraisal tool, designed for developmental screening and assessment tools and based on the most up to date literature, will support research and clinical efforts in ASD and developmental disabilities by providing a means to identify the most appropriate tools for use with different culturally and linguistically diverse populations.

There are three main goals of this paper. First, we describe the various dimensions of equivalence that should be considered when adapting a parent-report developmental instrument for a new population. Second, we present clear and concise guidelines in translation and cultural adaptation methodology, synthesized from the array of recommendations that can be found across related fields such as public health and social sciences. Guidelines are organized by the key stages found in the literature (instrument reproduction, pre-testing, and psychometric testing), with descriptions and rationales for the various ways to complete each stage. Figure 2.1 presents a visual summary of this process. Third, we propose a basic set of criteria, formulated as
a quality appraisal tool (Table 2.1), to examine the quality of translated and culturally adapted parent-report developmental screening tools and assessments.

**Considerations in Equivalence**

Language plays a central role in the vast majority of self-report instruments, where a respondent answers a series of questions (written or oral). Language itself is a representation of a group’s worldview, where word meanings are dependent on the speaker’s understanding of each concept (Castro et al., 2010; McKenna & Doward, 2005). Caregiver views of child development are heavily driven by culture, and the language used to probe caregiver observations of children must be interpretable within the context of their specific cultural lens and experiences.

Translation must take into account this cultural lens rather than translating each word into the target language. A literal translation of words may fail to produce an instrument that has meaning to speakers of the target language (Behr, 2018; Chapman & Carter, 1979; Colina et al. 2017; Nelson McDermott & Palchanes, 1994). The goal in reproducing an instrument in the target language is to maintain equivalence to the foundational concepts within the source instrument, while changing only the linguistic code, such that the same concepts are fully represented in the new cultural context (Acquadro, Jambon, Ellis, & Marquis, 1996; Gjersing et al., 2010). Thus, the term *cultural adaptation* as opposed or in addition to *translation* is used to represent the process of maintaining complete equivalence between versions (McKenna & Doward, 2005; Ware et al., 1996). Equivalence should be maintained on a variety of key dimensions to sufficiently adapt the new version to the target culture. Most descriptions of these dimensions come from either Flaherty et al. (1988) or Guillemim (1993), though the specific descriptions and terminology used by these authors differed. A synthesis of their equivalence dimensions, with the addition of descriptions from later authors, is presented here.
Linguistic equivalence consists of two levels: semantic and idiomatic. Semantic equivalence requires that the sentence structure and words of two instruments have the same meaning (Guillemin, 1993), including the conceptual definitions of the terms used (Beaton et al., 2000). Translators must consider multiple meanings of words and language differences where grammatical structures have no equivalent (Beaton et al., 2000; Guillemin, 1993). In traditional forward-back translation methodology, semantic equivalence is the primary, and sometimes only, dimension of equivalence that is considered. Idiomatic equivalence refers to the translations of colloquialisms, idioms, and expressions (Beaton et al., 2000; Flaherty et al., 1988; Sousa & Rojjanasrirat, 2011). For example, if translated literally, expressions such as “in his own world,” or “reading between the lines” may become meaningless in another language. These expressions need to be adapted to maintain the underlying meaning within the target language and culture (not responding to external stimuli or interpreting unstated information, respectively). Considerations of idiomatic equivalence may be more common in items describing social or emotional states (Guillemin, 1993; Ware, Gandek, & Keller, 1996), which are likely to be included in measures of social and communication development.

There are three closely related levels of construct equivalence, including conceptual, experiential, and content equivalence. For conceptual equivalence, the concept that the instrument measures must exist within the target culture (Gjersing et al., 2010; McKenna & Doward, 2005; Stewart & Napoles-Springer, 2000). To maintain experiential equivalence, Guillemin (1993) states that items describing experiences, situations, or contexts should fit the target cultural context. One example is a concept included in many screening tools for ASD: “peek-a-boo,” a common infant game in western, English-speaking cultures. An equivalent experience should be substituted (a common infant social communication game, ideally with
turns) rather than attempting to translate the term “peek-a-boo” to the target language. Maintaining experiential equivalence may require significant changes in items describing activities or experiences (Guillemin, 1993; Ware et al., 1996). Finally, content equivalence signifies that the concept itself has the same relevance and acceptance within the target culture and that items within the instrument appropriately represent the concept without significant omissions (Gjersing et al., 2010; Squires et al., 2013; Stewart & Napoles-Springer, 2000).

For technical, or operational, equivalence, the methods of data collection must yield the same data in the target population as with the source instrument (Squires et al., 2013). This often refers to the format of administration, such as pencil/paper, telephone, or face-to-face interview (Gjersing et al., 2010; Sartorius & Kuyken, 1994). Cultural norms dictate what type of information is shared with whom, where, and how (Carter et al., 2005; Flaherty et al., 1988). The mode of administration used in the source instrument may be inadequate to collect the intended data in the target culture and may need to be adapted (Guillemin, 1993; Flaherty et al., 1988). For example, if one cultural group deems it inappropriate for strangers to ask questions about a child’s development, administration strategies may use an interview format, incorporating time to familiarize the caregiver with the interviewer. Technical equivalence also refers to the most appropriate format (e.g., grid vs list), and level of language (i.e., formality, complexity) for the target population (Stewart & Napoles-Springer, 2000; Ware et al., 1996). For parent report tools that assist in screening or diagnosing children with ASD or other developmental disabilities, it is imperative for the tool to have wide applicability across the population. Consequently, the variability in caregiver education and literacy within the general population will influence the types of adaptations required to achieve technical equivalence.
Translation and Cultural Adaptation Guidelines

In the growing literature on translation procedures designed to yield valid tools for another language and culture, most authors describe three basic steps in the adaptation process: 1) reproduction of the source instrument into the new target language, 2) pre-testing with members of the target population, and 3) psychometric analysis of the new version. A graphic representation of this process appears in Figure 3.1. A long and resource-heavy reproduction and pre-testing process is typically recommended, involving many quality assurance measures (Beaton et al., 2000; Sousa & Rojjanasrirat, 2011; Ware et al., 1996); however, this body of literature is largely based on theory rather than empirical data. One older empirical study does suggest that methods using a single forward translation with no further quality checking or pre-testing procedures is not sufficient (Berkanovic, 1980). However, few studies have compared translation and adaptation methods. From comparison studies that do exist, da Mota Falcao, Ciconelli, and Ferraz (2003) and Perneger, Leplège, and Etter (1999) found few statistical differences between two translated versions of instruments using more or less rigorous methods, and concluded that translations of self-report instruments may be valid without undergoing as many quality assurance measures as once hypothesized. More recently, Acquadro et al. (2008) conducted a literature review of translation and adaptation studies and found some evidence to suggest that more rigorous approaches utilizing multi-step procedures result in higher quality translations. However, they found no empirical evidence to suggest that any specific method is ideal. While these studies provide some insight into the effectiveness of some strategies, more such research is needed to identify the most optimal reproduction and pre-testing methods (Lenderking, 2005). Regardless of how rigorous or complex the reproduction and pre-testing methods are, psychometric properties of an instrument cannot be assumed to be maintained.
following translation and adaptation (Beaton et al., 2000; McKenna & Doward, 2005). Multiple studies have documented differences in psychometric properties between different language versions of the same instrument, both in traditional and more rigorous translations (Fourie & Feinauer, 2005; Granas, Norgaard, & Sporrong, 2014; Kondolot, Ozmert, & Oztop et al., 2016). Consequently, psychometric properties should be reassessed using the new version with the new target population to ensure that the tool is reliable and valid. Our proposed quality appraisal tool presents a minimum set of criteria thought to produce a developmental screening or assessment tool that is sufficiently equivalent to the source instrument for both clinical and research purposes.
Figure 3.1. Representation of the translation and cultural adaptation process
Step 1: Reproduction of the Instrument

Process of Reproduction

The first phase of this process involves reproducing the instrument in the target language. Pre-planning is commonly recommended as a quality assurance method that occurs prior to initiating forward translation. Pre-planning involves examining the source version to determine its translatability in the new cultural context in an effort to identify potential challenges to maintaining construct equivalence, often referred to as a translatability assessment (Acquadro et al., 2018). Prior to initiating translation, teams may annotate the source instrument by offering wording alternatives, clarifying item intent, defining key terms, or replacing culture-specific terminology or concepts with universally understood concepts (Acquadro et al., 1996; Acquadro et al., 2018; Behr & Scholz, 2011; Chapman & Carter, 1979; Wild et al., 2005). Teams may also perform literature reviews, discuss concepts with experts, or conduct focus groups with parents or other caregivers in the target population (Flaherty et al., 1988; Gjersing et al., 2010; Sartorius & Kuyken, 1994; Stewart & Napoles-Springer, 2000). Such efforts allow researchers to determine how the concepts are defined and understood in the target population and may prevent threats to all dimensions of equivalence.

When the instrument is ready for reproduction, almost all authors recommend using multiple translators trained in dimensions of equivalence. Using a single translator increases the likelihood of including idiosyncrasies or individual language style in the translation (Guillemin, 1993; Wild et al., 2005). Alternatively, creating and comparing multiple translations by multiple translators helps to identify discrepancies in translator interpretation, especially when there are wording ambiguities within the source instrument (Beaton et al., 2000; Sousa & Rojjanasrirat, 2011). Thus, many authors recommend multiple forward translations to be made either
independently (Beaton et al., 2000; Sousa & Rojjanasrirat, 2011; Wild et al., 2005) or with a team of translators (Sartorius & Kuyken, 1994; Squires et al., 2013). It should be noted that in the context of the team approach, shared misconceptions may be exacerbated and social norms may make members reluctant to contradict each other (Maneesriwongul & Dixon, 2004). Translators should be native to the target language and culture (Beaton et al., 2000; Guillemin, 1993; Sousa & Rojjanasrirat, 2011), and ideally reside in the target country to ensure the use of local wording (Acquadro et al., 1996; Wild et al., 2005). Translators should have adequate translational competence, including sufficient knowledge in both languages and cultures and training in dimensions of equivalence (Behr, 2018). Multidisciplinary teams can include both individuals who are knowledgeable in the terminology and constructs of the instrument and laypersons who are blind to the concepts of interest, yet knowledgeable in colloquial terminology, such as child care providers or others who regularly interact with parents in the target culture (Acquadro et al., 1996; Beaton et al., 2000; Guillemin, 1993; Sousa & Rojjanasrirat, 2011). This ensures that concepts are appropriately represented in the translated instrument, but that the target population can understand items. Translators should translate items, instructions, and response options. If translations are made independently, forward translations are then collaboratively synthesized into a single version, resolving any discrepancies (Beaton et al., 2000; Sousa & Rojjanasrirat, 2011).

**Checking the Quality of the Reproduction**

Once an instrument has been reproduced in the new language, the quality of the translation should be inspected using quality assurance methods. The purpose of quality assurance methods is to identify and resolve discrepancies in linguistic, construct, or technical equivalence. A variety of methods are discussed in the literature, including but not limited to
back-translation, preliminary pre-testing, bilingual equivalence assessment, and expert review. Pre-planning, discussed above, is also considered a quality assurance method. A final consensus meeting takes place where the translation team reviews data collected in these quality-checking phases and makes revisions to improve equivalence across versions.

The vast majority of guidelines within the literature call for back-translation methods to test the quality of the reproduction. Back-translation consists of one or multiple translators, blind to the original instrument, who translate the forward reproduction back into the source language, resulting in one or more back-translations (Beaton et al., 2000; Sousa & Rojjanasrirat, 2011). Some recommend using only informed translators with knowledge of the concepts of interest in order to create a translation that would closely reflect the source instrument (Bracken & Barona, 1991; Sartorius & Kuyken, 1994). Others recommend using only individuals who are blind to the concept of interest in order to better elicit unexpected meanings and highlight imperfections (Beaton et al., 2000; Guillemin, 1993). This latter suggestion may be especially useful for developmental screening tools and assessments, as caregivers may not have any knowledge of the concept of interest when they answer the questions. The source and back-translated versions are then compared for the various dimensions of equivalence by the entire translation and adaptation team. This methodology, called the forward-back (FB) translation method, is meant to highlight translation errors that may have occurred in the forward translation that would impact the validity of the measure (Chapman & Carter, 1979; Guillemin, 1993). Thus, any errors can be identified and resolved during a final consensus step.

Although the FB methodology is recommended and used by most translation teams across fields, several authors point out that it has no scientific basis (Hagell et al., 2010; McKenna & Doward, 2005; Swaine-Verdier et al., 2004). Unfortunately, back-translation can
provide misleading information about the forward translation (Acquadro et al., 1996; McKenna & Doward, 2005; Swaine-Verdier et al., 2004). Often times, a high-quality forward translation may significantly alter the wording and structure such that it appears dissimilar to the source version when back-translated. Similarly, a high quality forward translation with a poor quality back-translation would give a false negative impression (Acquadro et al., 1996; McKenna & Doward, 2005; Swaine-Verdier et al., 2004). More damaging is the impression of a false positive. When the forward translation inappropriately retains the source language structure and wording, a back-translator may correct these errors, leaving the impression of a high quality forward translation (Acquadro et al., 1996; Maneesriwongul & Dixon, 2004; Nelson McDermott & Palchanes, 1994; Swaine-Verdier et al., 2004; Wang et al., 2006). In addition, back translation does not easily identify complex or formal language in a forward translation. In this case, an accurate back translation would leave the impression of linguistic equivalence, but the reproduction would lack intelligibility to monolingual caregivers in the target population (Chapman & Carter, 1979; Hagell et al., 2010; Nelson McDermott & Palchanes, 1994; Wang et al., 2006). One study comparing translation methodologies determined that back-translation sometimes identified translation errors that did not actually exist and missed problems that did exist (Behr, 2017). Some authors have recently called for the abandonment of back-translation methods altogether (Behr, 2017; Colina et al. 2017). Instrument developers should use caution if they choose back-translation and may consider adding additional quality checking measures.

Pre-testing is a method used in the second phase (described below), to be completed after reproduction and quality checking. However, some authors recommend using preliminary pre-testing methods as a quality check during the reproduction process (Nelson McDermott & Palchanes, 1994; Ware et al., 1995). Preliminary pre-testing can be combined with other quality
checking measures, and is sometimes used prior to back-translation so that intelligibility in the target population is checked before back-translation methods check for equivalence (Nelson McDermott & Palchanes, 1994). One comprehensive method, developed as an alternative to the FB process, utilizes preliminary pre-testing in a panel format. This strategy, the Dual Panel approach, has been shown to produce translations that are preferable to monolingual lay audiences (Hagell et al., 2010). In the Dual Panel approach, a panel of five to seven translators makes a consensus translation. Translators are native to the target language, have knowledge of the instrument’s underlying model, and have frequent contact with the target population (Hagell et al., 2010; Swaine-Verdier et al., 2004). A second panel of laypersons, monolingual in the target language and similar to the target population, then review the translation with a coordinator to ensure appropriate understanding of items (Hagell et al., 2010; Swaine-Verdier et al., 2004). Lay panels typically find more appropriate and intelligible wording than the bilingual panel, and can highlight unexpected connotations (Swaine-Verdier et al., 2004). Hagell et al. (2010) compared translations using the Dual Panel approach with the FB approach and found similar psychometric properties between the two versions; however, there were more missing data from the FB translation and monolingual laypersons preferred the Dual Panel version.

Other quality assurance strategies include bilingual equivalence assessment or expert review. Bilingual individuals can judge the equivalence of items between versions on various dimensions, either in a panel (Sartorius & Kuyken, 1994) or individually, using rating scales (Acquadro et al., 1996; Squires et al., 2013; Stewart & Napoles-Springer, 2000; Ware et al., 1995; Ware et al., 1996). Alternatively, fully bilingual groups can complete the assessment in each language, comparing results using a test-retest analysis (Chapman & Carter, 1979). Another
possibility is to review the new tool with experts in the field of ASD or developmental disabilities from the target language and culture (Sousa & Rojjanasrirat, 2011).

Importantly, teams should keep detailed documentation, including difficulties with translation, alternative wording choices, feedback from panels, and rationales for any revisions made throughout the process, especially those resulting in significant deviations from the source version (Acquadro et al., 1996; Swaine-Verdier et al., 2004). Documentation will improve transparency of the process, promote the importance of the various quality checking procedures, provide helpful information for later validation studies, and provide pertinent information to any future research teams who attempt translation efforts on the same instrument (Beaton et al., 2000; Maneesriwongul & Dixon, 2004; Swaine-Verdier et al., 2004).

**Reaching a Final Consensus**

A final consensus meeting with the entire adaptation team is deemed one of the most critical steps in reproduction (Beaton et al., 2000; Wild et al., 2005). A multidisciplinary team is most often recommended, including a methodologist, experts in the area of interest, all translators, the source instrument developers (Beaton et al., 2000; Guillemin, 1993), and possibly a layperson monolingual in the target language, such as a monolingual caregiver (Sousa & Rojjanasrirat, 2011). The team reviews all parts of the instrument and all data collected during the reproduction and quality testing process, considering linguistic and construct equivalence (Beaton et al., 2000; Guillemin, 1993; Sousa & Rojjanasrirat, 2011). Re-translation of individual items may be necessary if problems arise (Beaton et al., 2000; Bracken & Barona, 1991; Guillemin, 1993; Sousa & Rojjanasrirat, 2011). Using consensus rather than compromise, the team makes revisions where necessary and develops a consolidated pre-final version that will be used for pre-testing (Beaton et al., 2000; Guillemin, 1993; Ware et al., 1996; Wild et al., 2005).
Step 2: Pre-Testing

Pre-testing is a vital step to evaluate the quality and readability of a translated instrument in the target population following reproduction and quality testing and prior to undergoing full psychometric analyses. Pre-testing involves administration of the adapted instrument to a small sample of target participants to ensure that no obvious, easily reparable problems exist with the instrument’s wording or administration mode before expending the considerable resources necessary for full validation studies with larger samples. There are a variety of ways to pre-test the translation. Qualitative data are most often collected, though some authors perform preliminary statistical analyses using quantitative data (Bracken & Barona, 1991; Maneesriwongul & Dixon, 2004).

Pre-testing allows for assessment of face-validity (Acquadro et al., 1996), and examination of linguistic and construct equivalence (Guillemin, 1993; Sousa & Rojjanasrirat, 2011), and identification of any unnecessarily complex or formal language (Acquadro et al., 1996). The sample should adequately represent the diversity within the target population (Acquadro et al., 1996; Bracken & Barona, 1991; Ware et al., 1996; Wild et al., 2005). For developmental screening and assessment tools, the sample should consist of caregivers of children in the target age range, with a variety of education levels, and representing a variety of regions in the target geographic area. Adaptation teams may choose to include some caregivers whose children are known to have developmental difficulties. The literature is not consistent with recommendations for sample size, ranging from 5 (Wild et al., 2005) to 50 or more (Ware et al., 1996). Focus groups may be used, though individual cognitive interviews are recommended (Acquadro et al., 1996). During a cognitive interview, participants complete items as intended and are asked probing questions about each item, including their understanding of the item’s
meaning, how they come up with answers to the items, and the appropriateness of response options (Beaton et al., 2000; Guillemin, 1993; Stewart & Napoles-Springer, 2000; Wild et al., 2005). Many authors suggest documenting participant reactions such as facial cues, resistance to respond, or nervous laughter, because these reactions could indicate inappropriate wording or administration method (Bracken & Barona, 1991). Researchers should also note possible acquiescence response bias and socially desirable responding in parent-report instruments, particularly if a stigma around child developmental disabilities exists within the culture. Less often, feedback surveys are completed following self-administration of the instrument (Reichenheim & Moraes, 2007; Sousa & Rojjanasrirat, 2011), common with large pre-testing samples (Ware et al., 1996).

Bilingual pre-testing samples can provide unique information regarding the equivalence of instruments, though fully bilingual laypersons may be difficult to recruit (Guillemin, 1993; Ware et al., 1996). Methods include equivalence ratings (Guillemin, 1993; Squires et al., 2013) or test-retest or split-half analyses (Acquadro et al., 1996; Sartorius & Kuyken, 1994; Sousa & Rojjanasrirat, 2011; Wang et al., 2006). Items with low correlations on such equivalence tests, or low ranking items on equivalence ratings, are revised.

Data collected during the pre-testing phase should be used to revise problem items, instructions, response options, or administration methods so that linguistic, construct, technical, and in turn, psychometric equivalences are maintained (Soto et al., 2015; Sousa & Rojjanasrirat, 2011). Similar to the reproduction phase, the pre-testing phase should be well documented, including rationales for any changes made (Beaton et al., 2000; Maneesriwongul & Dixon, 2004). Finally, individual items may be sent back into the reproduction phase if significant issues are identified.
**Step 3: Psychometric Testing**

Although ignored in many guidelines in the literature (Squires et al., 2013; Wild et al., 2005), the re-testing of psychometric properties of an instrument may be the most important step in the process of translating and adapting a screening tool. Many consider it a gross misconception to assume that the psychometric properties of an instrument are retained when translated and adapted for a new population (Beaton et al., 2000; McKenna & Doward, 2005). Without psychometric data, the reliability and validity of a given adapted screening measure are unknown.

Parent-report tools are used in developmental screening tools as well as more comprehensive assessments. Developmental screening tools are used clinically to identify children who exhibit concerning behaviors so they may be referred for a full evaluation and developmental support. Comprehensive assessments are used to provide information for evaluation teams in order to determine a differential diagnosis. Consequently, assessment tools should have better diagnostic validity than screening tools. When examining the properties of a new screening or assessment tool, researchers should prioritize analyses that test whether the tool consistently collects appropriate data and differentiates between those who do or do not have a given developmental disorder.

**Recruitment**

There are multiple recruitment strategies used to test psychometric properties of screening and diagnostic instruments. In case control studies, two separate populations are recruited: those more likely to have the disorder and those less likely to have the disorder (Campbell et al., 2015). Although this strategy will likely be fruitful in identifying participants with the disorder, diagnostic accuracy will be somewhat exaggerated compared to clinical
practice, as statistical differences between participant groups are likely to be pronounced (Campbell et al., 2015). Cross-sectional studies, however, recruit only from populations likely to have the disorder, providing a more valid measure of diagnostic accuracy for assessment tools in a clinical setting, where all patients present with developmental concerns (Campbell et al., 2015). Alternatively, participants may be recruited from the general population. This may provide a more accurate representation of a screening tool intended for general population screening. A less robust method involves a known-groups analysis, where participants are recruited who are already known to either have the developmental condition of interest or be typically developing (Ware et al., 1995; Ware et al., 1996). While few strict guidelines are provided, sample sizes should be large enough for the analyses of interest, which will depend on the number of items in the tool, the number of response options, and the variability of responses (Beaton et al., 2000; Sousa & Rojjanasrirat, 2011). In general population samples, the required sample size will also depend on the prevalence of the condition. For ASD, the estimated prevalence is 1-2% of the population (Baio et al., 2018), whereas communication disorders affect approximately 10% of children (National Academies of Sciences, Engineering, and Medicine, 2016). Consequently, studies of ASD-specific tools would require larger samples from the general population, as compared to communication disorder-specific tools, in order to identify enough cases for psychometric analysis.

**Methodology**

Methodology for the assessment of psychometric properties is dependent on the intended statistical analyses. For most tests, a large sample of participants will complete the adapted tool and analyses will follow. Participants may complete an additional assessment or provide additional information in validity studies. Tests of diagnostic accuracy involve the administration
of a gold standard index test to be used as a definitive factor that categorizes participants as having the condition in question or not having it (Campbell et al., 2015). Diagnostic accuracy of the instrument is determined by comparing those correctly and incorrectly identified by the instrument (Campbell et al., 2015). To reduce bias, researchers can alternate the order of administration, use appropriate time delay between examinations, blind examiners, recruit a diverse, yet representative population compared to the target population (Campbell et al., 2015; Kim et al., 2015), and use an index test that is valid and reliable in the target culture and language (Habbema et al., 2009).

**Statistical Analyses**

To examine a tool’s performance with the new population, analyses begin with examination of data quality, including data completeness, variability of scores, and response bias. Large portions of missing data on a particular item may indicate a confusing, offensive, or inapplicable item, whereas large portions of missing data on many items in a section may indicate problems with instructions, layout, or administration methods (Ware et al., 1995; Ware & Gandek, 1998; Ware et al., 1996). When comparing findings to results of psychometric testing from the original instrument, participants should select from the same range of response options on each item, and their scale scores should span the same range of scores possible (Stewart & Napoles-Springer, 2000). Similarly, participants should have similar proportions of ceiling or floor level scores as in the source instrument (Ware & Gandek, 1998). In the case of screening tools, there should be large numbers of individuals who score at low risk when the tool is used in a low-risk sample. Measures of acquiescence response bias, social desirable responding, and preferences for extreme response categories (Stewart & Napoles-Springer, 2000) may be particularly relevant for constructs with high stigma, which may be the case with scales related to
child developmental difficulties. Finally, scale and item means and standard deviations should be similar, or at least follow a similar pattern of rank order, in both the target and source versions of the instrument (Bracken & Barona, 1991; Ware & Gandek, 1998; Ware et al., 1996). For example, many parent-report tools include items that ask about milestones acquired in typical development as well as behaviors not present in typical development. In both language versions, means from a low-risk sample should indicate frequent endorsement of items describing behaviors present in typical development, especially as children’s ages increase. Similarly, means from a low-risk sample should indicate less frequent endorsement of items describing behaviors generally not present in typical development.

The most frequent type of reliability analysis recommended in the literature is internal consistency (Bracken & Barona, 1991; Gjersing et al., 2010). Also commonly recommended are measures of test-retest reliability (Stewart & Napoles-Springer, 2000; Ware & Gandek, 1998; Ware et al., 1996). Inter-rater reliability is less frequently described (Acquadro et al., 1996), yet it may provide especially useful information if the screening tool is intended for use with varying informants, such as parents and childcare providers. Reliability analyses provide information about the tool’s consistency.

Arguably, the most important property of a screening or assessment tool is its diagnostic accuracy. For any tool, this begins with the scoring process, which requires reexamination when translated and adapted for use with a new population (Acquadro et al., 1996; Stewart & Napoles-Springer, 2000; Ware et al., 1996). Summated rating scale assumptions (Stewart & Napoles-Springer, 2000; Ware et al., 1995; Ware & Gandek, 1998; Ware et al., 1996) or weighting of items should be reexamined (Guillemin, 1993; Stewart & Napoles-Springer, 2000). Similarly, cut-off scores cannot be assumed to be maintained with a new population (Habbema et al., 2009;
Stewart & Napoles-Springer, 2000). Cut-off scores, which determine which individuals are considered at high risk according to the instrument, must be chosen strategically, in order to maximize sensitivity and specificity and to minimize errors (i.e. false positives and false negatives; Kim et al., 2015). This is often performed using an area under the curve analysis of the receiver-operating characteristic curve (Campbell et al., 2015; Habbema et al., 2009).

Given appropriate scoring methods, the tool’s sensitivity, specificity, positive predictive values, and negative predictive values should be measured (Campbell et al., 2015; Habbema et al., 2009; Kim et al., 2015; Stewart & Napoles-Springer, 2000). It is helpful for authors to provide frequencies of true and false positives and negatives so that the reader can calculate these and other variables related to diagnostic accuracy (Kim et al., 2015). Tests of convergent and discriminant validity can also provide information about what an instrument is measuring (Beaton et al., 2000; Hagell et al., 2010; Sousa & Rojjanasrirat, 2011). This includes examining correlations with external variables or other measures that are thought be associated or not with the given construct (Anderson, Mcfarlane, Naughton, & Shumaker, 1996; Stewart & Napoles-Springer, 2000; Ware et al., 1996).

*Equivalence testing* determines if the two versions of the tool, the source and new versions, are measuring the same information. Evidence of equivalence between versions is less important than diagnostic accuracy if the tool is planned for clinical use rather than use in cross-cultural research, where data collected by the different versions of the tool will be pooled or compared. However, if the versions are not found to be equivalent in their measurement, it may be that each is measuring a separate latent construct, which may or may not be equally able to predict the presence of a developmental disability.
Equivalence analyses should show that the translated instrument is equivalent to the original instrument in all observable respects, where the distribution of item and scale scores depends on actual values of latent variables and not on any other characteristic, a concept referred to as measurement invariance (Küçükdeveci et al., 2004; Ware & Gandek, 1998). Classical analysis of measurement invariance involves a series of multiple group factor analyses with increasing constraints placed on the parameters until all parameters are fixed. If the final model continues to fit the data from both versions of the instrument, then the two instruments show measurement invariance. An alternative approach, based in item response theory and focused at the item level, is known as differential item functioning (DIF; Anderson et al., 1996; Küçükdeveci, Sahin, Ataman, Griffiths, & Tennant, 2004). If high DIF is found for a particular item, participants appear to be responding to that item differently across versions. The item should then be reviewed and possibly revised (Anderson et al., 1996; Küçükdeveci et al., 2004). Of note, analyses of measurement invariance and DIF require considerable variability within the data and a large covariance matrix. Depending on the length and factor structure of the tool, an extremely large sample size may be needed, which may make these analyses logistically difficult.

At any level of equivalence analysis, if there are differences found between scales, the research team must decide if these differences are likely to be true population differences or differences in the instrument itself (Reichenheim & Moraes, 2007; Ware & Gandek, 1998; Ware et al., 1996). In the case of the latter, the adapted instrument may need to be revised (by returning items to the reproduction and pre-testing phases), depending on its intended purposes. As previously indicated, lack of measurement invariance is less important if diagnostic accuracy is maintained for an instrument designed solely for clinical use. As in each of the other steps,
transparent documentation of psychometric analyses should be sufficient to allow the reader to examine the reliability and validity of the new version (Acquadro et al., 1996).

Once acceptable psychometric properties have been achieved for the tool’s intended purpose, the new instrument can be used in the target population. In an effort to maximize transparency, the entire translation and adaptation process should be described with as much detailed information as possible in peer-reviewed sources (Acquadro et al., 1996; Maneesriwongul & Dixon, 2004). However, due to the complexity and length of the process, such information may be published in multiple articles (Sousa & Rojjanasrirat, 2011).

**A Quality Appraisal Tool**

Using these guidelines presented in the literature, we propose a quality appraisal tool (Table 3.1). This checklist has been designed to systematically analyze the instrument translation and cultural adaptation process of any one assessment. It could be used in a clinical setting by those who seek to compare the quality of the translation of a particular instrument or set of instruments. Empirically, it could be used by researchers to systematically analyze the translation methods of a set of instruments for a systematic evidence map. Clinicians and researchers alike who wish to translate and culturally adapt an instrument themselves could use the quality appraisal tool as a way to guide their own process. The appraisal tool does not have a quantitative scoring system, nor is there consensus on what level of quality would indicate acceptable level of rigor in methodology. However, it provides a set of minimum steps thought necessary to ensure appropriate translation of self-report developmental screening and assessment tools.
Table 3.1 Quality Appraisal Tool for Translated and Adapted Caregiver-Report Developmental Instruments

1. Was more than one translator used? Yes / No
   - Did translators have varying backgrounds? Yes / No / NA
   Description:

2. Was at least one sound quality assurance method used in the reproduction phase? (i.e. pre-planning, cognitive testing, equivalence testing, expert panel, back-translation with consensus committee review, dual panel) Yes / No
   Description:

3. Was pre-testing completed with an appropriate sample of members of the target population? Yes / No
   Description:

4. Were appropriate revisions made using data from quality assurance and pre-testing steps? Yes / No
   Description:

5. Was psychometric testing completed on the new version? Yes / No
   - Was the sample sufficiently large and diverse to perform statistical analyses and adequately represent the target population? Yes / No / NA
   Description:

   - Was appropriate blinding and time delay used when applicable? Yes / No / NA
   Description:

   - Was the gold standard reference test culturally and linguistically appropriate (validated on the target population)? Yes / No / NA
   Description:

   - Were scoring procedures and cut-off scores reexamined and found to be appropriate? Yes / No / NA
   Description:

   - Were diagnostic accuracy data reported and acceptable (true/false negatives, true/false positives)? Yes / No / NA
   Description:

   - Was data quality acceptable? (e.g.; ranges of scores, ceiling/floor rates, etc.) Yes / No / NA
   Description:

   - Were appropriate reliability and validity tests completed with results reported? Yes / No / NA
   Description:

   - Were appropriate item and scale equivalence tests completed with results reported? Yes / No / NA
   Description:
Final Considerations

Parent-report developmental screening and assessment tools are widely used to identify children showing signs of ASD or other developmental disabilities. Because of their utility, these tools are being translated into multiple languages and used internationally. Reliable and valid measures are a critical need for improving our global capacity to serve individuals with ASD and other developmental disabilities. Yet traditional forward-back translation methodology is not sufficient to translate and culturally adapt these measures such that their reliability and validity is retained in the new target population. In this article, we have described the various dimensions of equivalence most relevant when translating and culturally adapting self-report developmental screening and assessment tools, and provided a description of the basic stages in the process, along with various strategies that can be used in each stage. Due to the scarcity of studies comparing translation methodologies, it is not possible to recommend any one specific set of strategies based on empirical evidence. Clearly, more research is needed in this area to determine which strategies are the most efficient and effective in translating and culturally adapting tools so that the new tool functions appropriately in the new target population. For now, we hope that shedding light on the translation and cultural adaptation process will encourage research teams within the fields of ASD and developmental disabilities to increase the rigor in their translation methods and describe their methodology in greater detail in research reports.

Conflict of Interest: Both authors declare that they have no conflicts of interest. This paper was prepared as a portion of a doctoral dissertation, as provided by U.S. Department of Education Doctoral Leadership Grant, H325D130041.
REFERENCES


instruments: language, setting, time and statistical considerations. *BMC Medical Research Methodology*, 10(13), 1–10.


CHAPTER 4: RIGOROUS TRANSLATION AND CULTURAL ADAPTATION OF AN AUTISM SCREENING TOOL: FIRST YEARS INVENTORY AS A CASE STUDY

Behavioral symptoms of autism spectrum disorder (ASD) have been documented in children as early as 9 to 12 months of age (Baranek, 1999; Dietz et al., 2006; Turner-Brown et al., 2013; Zwaigenbaum et al., 2005). Identifying children with early behavioral symptoms of ASD can allow us to intervene when neurological processes are still developing and the full impact of the disorder has not yet been realized (Dawson, 2008; Dawson et al., 2010). Despite documented presence of early symptoms, the average age of diagnosis continues to be over 4 years in the United States (Baio, Wiggins, & Christensen, 2018). Average age of diagnosis is higher in non-English speaking countries, such as Spain and France, where average age of diagnosis is reported to be as late as 7 and 10 years, respectively (Chamak et al., 2011; Morales-Hidalgo et al., 2018).

Screening for ASD

Currently, the most feasible way to identify children with ASD symptoms at a young age is to screen children universally for early behavioral markers common in ASD. This is typically accomplished in a parent-report format (Thabtah & Peebles, 2019). Used appropriately, many parent-report screeners support early identification of young children who will be diagnosed with ASD (Hampton & Strand, 2015; Robins et al., 2014; Turner-Brown et al., 2013; Wiggins et al., 2007). Ideally, concurrent or predictive diagnostic accuracy of these screening tools should be as high as possible, within the constraints of the parent-report format. The goal is to maximize true
positives and negatives while minimizing false positives and negatives to ensure that (a) we do not miss children who do in fact have ASD, thus delaying diagnosis and early intervention, and (b) we minimize undue family stress and unnecessary expenditure of resources that are both associated with false-positive screens.

**Screening in Diverse Populations**

To address our vast global cultural and linguistic diversity, many ASD screening tools that were developed in English in the US, UK, or Australia are being translated into other languages and used in other world regions. The most frequently used translation approach for these measures is a traditional “forward-back” approach (El-Behadli et al., 2015; Soto et al., 2015). In this approach, one translator translates the source text into the new language. That forward translation then goes through back-translation, where a second translator translates the new text back into the original source language. Finally, the two same language versions are compared for equivalence (original source text and back-translated text). Some items may be re-translated and back-translated again as a result of this comparison step. The instrument is ready for use once this comparison step is completed.

Unfortunately, several recent studies of forward-back translated developmental screening tools appear to suggest that this methodology may not produce translated versions that meet high psychometric standards. The *Modified Checklist for Autism in Toddlers* (M-CHAT; Robins et al., 2001), along with the revised version (M-CHAT-R/F; Robins et al., 2014), is an ASD screening tool used widely internationally. Many forward-back translations of this tool have resulted in unexpectedly high risk scores and screen positive rates, high false positive rates, and poor item-discrimination rates as compared to the original English version (Brennan et al., 2016; DuBay & Watson, 2019a; Guo et al., 2019; Guthrie et al., 2019; Kara et al., 2014; Kimple et al., 2014;
Mohamed et al., 2016; Seung et al., 2015; Windham et al., 2014). Forward-back translations of
the Ages and Stages Questionnaire (Bricker & Squires, 1999), a broad-band developmental
parent-report screening tool, have also resulted in poor sensitivity (Guiberson & Rodriguez,
2010), high screen positive rates (Juneja et al., 2012), and differential item functioning (El-
Behadli et al., 2015). Similarly, the forward-back Hebrew translation of the First Year Inventory
v. 2.0 (FYIv2.0; Baranek et al., 2003) also resulted in a high risk score mean and differential
distribution of responses as compared to the original English version (Ben-Sasson & Carter,
2012).

This pattern of poor psychometric results from forward-back translations may be due to
underlying dimensions of equivalence that are not maintained by the forward-back approach.
There are a range of dimensions of equivalence that can exist between any two versions of an
instrument, including linguistic, construct, and technical equivalence (Beaton et al., 2000; DuBay
& Watson, 2019b). Linguistic equivalence refers to a translation’s words and phrases and how
well these match the original version, with special consideration taken for multiple meanings of
words, word connotations, and idioms (Beaton et al., 2000; Guillemin, 1993). Construct
equivalence refers to the foundational meaning of the concepts and experiences described in the
instrument, and whether those are fully and appropriately represented in the new language and
cultural context (Guillemin, 1993; Squires et al., 2013; Stewart & Napoles-Springer, 2000).
Technical equivalence refers to the data collection methods, including administration format,
print layout, and complexity of language, and whether these need to be adapted to collect the
same data in the new cultural context (Flaherty et al., 1988; Guillemin, 1993; Squires et al.,
2013). The traditional forward-back approach does not include clear standards for cultural or
linguistic considerations as they relate to these various dimensions of equivalence.
**Rigorous Cultural Adaptation Approach**

An alternative cultural adaptation approach to translation has been recommended to maintain equivalence across each of these dimensions (Beaton et al., 2000; Guillemin, 1993). Since these recommendations were first published, multiple studies across fields have provided additional steps to increase the rigor in the adaptation process (Acquadro et al., 2008; Gjersing et al., 2010; Hagell et al., 2010; Sousa & Rojjanasrirat, 2011; Squires et al., 2013; Swaine-Verdier et al., 2004; Wild et al., 2005). Although examples are still limited, this more rigorous approach has been increasingly used in the field of ASD in recent years (Canal-Bedia et al., 2011; Cuesta-Gomez et al., 2016; Magán-Maganto et al., 2018; McClure et al., 2018).

DuBay and Watson (2019b) provided a set of guidelines for translation and cultural adaptation specifically for parent-report tools for ASD or other developmental disorders. This more rigorous approach involves a forward translation by a team of translators from a range of backgrounds who are trained in dimensions of equivalence. There should be at least one quality assurance method to ensure that the translation is of high quality prior to moving to later steps, such as back-translation, dual panel, expert review, preplanning, etc. Then a consensus team of the translators, coordinators, original instrument developers, and others reviews all of the data collected thus far and decides on a pre-final version of the instrument. The pre-final version goes through a pre-testing round where qualitative data are collected from members of the target population to ensure that the text is interpreted as intended. After any additional revisions as needed, the final version is administered to a large community sample to determine population-specific norms and relevant psychometric properties (Beaton et al., 2000; DuBay & Watson, 2019b; Gjersing et al., 2010; Nelson McDermott & Palchanes, 1994; Sousa & Rojjanasrirat, 2011; Wild et al., 2005). Importantly, many guidelines call for detailed documentation of this
process in publications to increase transparency and inform future studies of similar instruments (Acquadro et al., 1996; Maneesriwongul & Dixon, 2004; Swaine-Verdier et al., 2004).

**First Years Inventory**

This article uses the *First Years Inventory v. 3.1* (FYIv3.1; Baranek et al., 2013) as a case study of the application of this more rigorous translation and cultural adaptation methodology. The FYIv3.1 is a research version of a parent-report ASD screening tool that was developed in English in the US. The FYI was originally developed and normed on 12-month-old children (FYIv2.0; Baranek et al., 2003; Reznick et al., 2007), and showed sensitivity of .44 and specificity of .99 for ASD in a community sample (Turner-Brown et al., 2013). The FYIv3.1 was expanded to a broader age range and currently contains 69 items over two forms (form A, form B), with 48 items per form and 27 overlapping items. The FYIv3.1 uses a two-domain structure, including a social-communication domain and a sensory-regulatory functions domain. Items from each domain are interspersed throughout the instrument. Preliminary data suggest the FYIv3.1 has upper bounds of a sensitivity of .4 and specificity of .98 for ASD in a clinical sample of 8 to 16-month olds (J. Sideris, personal communication, Feb. 18, 2020). Using data collected on the FYIv3.1, the FYI authors aim to ultimately create one or more shorter versions with validity and clinical utility in identifying infants and young toddlers at elevated likelihood of later ASD diagnoses across the age range of 8-16 months.

**Purpose**

This project aimed to translate and culturally adapt the FYIv3.1 (title, instructions, response options, and all 69 items) for a target population of US-based Spanish speakers with Latin American backgrounds, using the rigorous approach described in DuBay and Watson (2019b). The US is home to a large heterogeneity of Latin American cultures as well as dialects
of Spanish. Therefore, our goal was to develop an instrument that would be clearly intelligible to a range of Latin American cultures and dialects of Spanish. We incorporated input from a range of regions of Latin America, with a focus on Mexican dialects, as approximately 65% of US-based individuals with “Latino/Hispanic” ethnicity report Mexican heritage (Motel & Patten, 2012). This translation was in preparation for collection of normative data on infants and young toddlers from Spanish-speaking families in the US, comparable to the data collected on the English version of FYIv3.1.

The purpose of this paper is to document a rigorous cultural adaptation process using the FYIv3.1 as a test case, highlighting features of the instrument that led to the greatest challenges in translation or that required the most cultural adaptation in order to maintain linguistic, construct, and technical equivalence. We aim to both provide transparency in our process of translating of the FYIv3.1 and to inform future translation and adaptation projects for similar tools.

Methods

Participants

Twenty-five Spanish-speaking caregivers participated in this study, including twenty-one mothers, one father, two grandmothers, and one aunt, all primary caregivers. Caregivers ranged in age from 22 to 60 years (average = 33.2 years). Their children ranged in age from 8 months to 18 months (average = 11.8 months). All caregivers were functionally monolingual in Spanish, with the exception of one who was born in the US and raised bilingually. All others were born in other countries of Latin America. Similar to the Spanish-speaking population in the US, most participants were born in North or Central America. About half were from Mexico, 3 from Honduras, 2 from Guatemala, and 1 from EL Salvador. The remaining participants were from
northern South America (4 from Venezuela, 1 from Colombia, and 1 from Peru). Of the 14 who provided information on household income, 50% reported $20,000-39,000USD annual income, with 36% under $20,000 and 14% between $40,000-$59,000. Participants reported a range of education levels, with 30% reporting less than 8th grade level of education and the rest relatively evenly distributed between some high school, high school degree, some college or Associates degree, Bachelor’s degree, and Master’s degree. Participants had been living in the US for 10 years on average, with one third of participants having emigrated less than 5 years prior to data collection. Eleven of the participants’ children were girls and 14 were boys. Most (79%) of the children had older siblings (average = 2.1), while 21% were first-born.

Participants were recruited from local community organizations serving Latin American families (e.g. bilingual parent-child reading groups, food pantries, and giving closets). Interested families who met inclusion criteria (native Spanish-speaker, primary caregiver for a child within the age range, self-reported Latino or Hispanic ethnicity) were invited to participate in an interview at the place and time of their choosing. Participants were compensated for their time with a $30USD gift card.

Measures

First Years Inventory v. 3.1 (FYIv3.1; Baranek et al., 2013)

This ASD screening questionnaire consists of a series of questions that ask parents or other caregivers how frequently their child engages in specific behaviors. For example, “When you point towards something interesting, does your child look at it?” For each item, the five response options include “always,” “sometimes,” or “never,” with two intermediate unlabeled options between “always” and “sometimes” and between “sometimes” and “never.” The preliminary FYIv3.1 scoring algorithm converts responses for each item to risk scores. Risk
scores are weighted based on frequency probabilities in the normative sample for a corresponding age group, then averaged into single scores for each domain, one for social communication and one for sensory-regulatory functions. Children screen positive on the FYIv3.1 when their risk scores fall above the cutoff threshold.

Demographics questionnaire

Prior to the administration of the FYIv3.1, participants completed a demographic questionnaire requesting their age, gender, race/ethnicity, relationship to the child, number of years living in the US, country of origin, native language, household income, and education level. Participants also provided the age, gender, and birth order of their child.

Procedure

The FYIv3.1 was translated and culturally adapted using a rigorous approach (described by DuBay & Watson, 2019b) consisting of a forward translation phase with reproduction, quality assurance checks, and consensus meeting, and a pre-testing phase with cognitive interviews.

1. Forward Translation

1a. Reproduction. To begin the translation process, three bilingual translators were trained in the aforementioned dimensions of equivalence. Translators were all native Spanish speakers and reported Latino/Hispanic ethnicity. They were native to Puerto Rico, and the Guanajuato and Jalisco regions of Mexico, respectively, raised in the local language and culture, and also fluent in English and residing in the US. One translator was an expert in occupational therapy in the area of ASD. One translator was an expert in early childhood education. The final translator was a mother of a child in the target age range. Each translator translated the title, instructions, response options, and items of the FYIv3.1 independently. These translations were
collaboratively synthesized to a single version during a synthesis meeting with the translators and the lead investigator (a bilingual speech-language pathologist, ASD expert, and first author).

1b. Quality Assurance. Two quality assurance methods were used to check the quality of the synthesis version, including a preliminary pre-testing round and back-translation. In the preliminary pre-testing round, the synthesis version was administered to five members of the target population: Latin American, Spanish-speaking caregivers of children in the target age range, via cognitive interviewing strategies. (See pre-testing section for details of cognitive interviewing methods and strategies). The lead investigator administered all preliminary pre-testing interviews. Participants were administered between 37 and 52 items, according to their preference. There were 27 “core” items that are included on both Form A and Form B of the FYIv3.1. These 27 items were administered to each of the participants for this preliminary pre-testing round. The remaining items were each tested by an average of 1.9 participants during this round.

Preliminary pre-testing resulted in recommended revisions to the instructions and 36 items. Two bilingual native English speakers who were trained in dimensions of equivalence then independently back-translated this edited version to English to check for accuracy and appropriateness of the translation. One translator had content expertise as a graduate student in speech-language pathology with a social, educational, and professional background in multiple Central American Spanish dialects. The second translator was a mother of a child in the target age range, with a cultural and linguistic background in Colombian Spanish through family ties. The first author completed comparisons between the source text and back translations.

1c. Consensus Meeting. An adaptation committee of 10 members, including original FYI creators, all translators and back translators, bilingual experts in child development and
developmental disabilities, and the lead investigator met to review preliminary cognitive interview results as well as results from back-translation comparisons. Native Spanish-speaking attendees, all of whom were living within the US, represented a range of cultures and regional dialects from multiple Latin American countries (Mexico, Puerto Rico, Colombia). The committee agreed on a pre-final Spanish version using consensus instead of compromise (Beaton et al., 2000).

2. Pre-Testing

This pre-final version was pre-tested with twenty target population members using cognitive interviewing techniques (Willis, 2011). For this full pre-testing round, the expert back-translator served as a cognitive interviewer, along with a second graduate student in speech-language pathology who was a native Spanish speaker from the Coahuila region of Mexico. Interviewers were trained on cognitive interviewing techniques, provided with an interview protocol, listened to sample interview recordings, and co-administered one interview with the first author before administering interviews independently.

During the interview, participants first gave informed consent and completed the demographic questionnaire. Interviewers explained the cognitive interview format, administered one “practice item,” then proceeded with the FYIv3.1 title, instructions, and items. For both pre-testing rounds, interviewers administered the Spanish FYIv3.1 using a combination of verbal probing and think-aloud strategies (Willis, 2011). Specifically, participants were asked to read each item aloud, rephrase the item, then orally narrate their decision-making process as they chose from response options. Additional pre-determined follow-up questions were asked to target specific concepts or challenges noted in previous translation steps. Whenever misinterpretations were noted, interviewers asked follow-up questions specific to these
responses. Interviewers then attempted to convey the intended item meaning and asked for suggestions for improving intelligibility. This interview format revealed participants’ comprehension of items, ability to recall target information, and decision-making and response processes as they read test items.

Items that presented no challenges during preliminary pre-testing (n = 18) were administered to five more participants during the full phase of pre-testing. Items that had presented moderate challenges (n = 14) were administered to thirteen more participants during the full phase of pre-testing. Items that had presented significant challenges (n = 37) were administered to all participants in the full phase of pre-testing. Thus, during this full pre-testing phase, each item was administered to as few as 5 or as many as 20 participants. Four items continued to result in significant challenges during our second round of pre-testing with the first half of the participants. Modifications to these items were made prior to administering them to the remaining participants to examine if new modifications would improve intelligibility.

Following the pre-testing phase, the adaptation committee re-convened to review results and make final revisions. The lead investigator compiled a detailed record of all recommended wording for all content (title, instructions, response options, and items) along with rationales for adaptations and back translations for ease of review by non-Spanish-speaking team members.

**Data analysis**

Qualitative data from both pre-testing phases was analyzed using informal methodology, including individual interview review and compilation of results across interviews (Willis, 2011). Interviewers reviewed audio recordings and notes from cognitive interviews to create individualized transcript summaries for each participant. Summaries included the participant’s rephrase and thought process in response to items as well as the response option selected.
Particular detail was recorded where participants showed unexpected or problematic reactions to items and any level of unintended interpretation. Discrepancies in linguistic, construct, or technical equivalence between each item’s original intent and its meaning as understood by participants were recorded, along with specific suggestions for improvement of intelligibility. Results across all participants from each unit of content were then reviewed separately. Where problems were identified from more than two participants, these results were discussed in consensus meetings and alternative translations were recommended. For the purpose of informing future translation and cultural adaptation studies, all misinterpretations and unexpected participant reactions were additionally coded for type and organized into categories. Rather than individualized results for each item, these categories are presented in the results section.

Itemized data from all pre-testing participants were examined for a preliminary analyses of data variability by comparing patterns of responses to previously collected normative data from English-speaking participants.

Results

By implementing this translation and cultural adaptation process, we identified a total of 128 distinct challenges in translating the wording or phrasing of the original FYIv3.1. Of the 69 FYIv3.1 items, 57 required considerable edits or adaptations to achieve an acceptable level of understanding among our pre-testing participants. Most (81%) of the challenges encountered were identified during pre-testing rounds.

Results are organized into two sections: low-challenge areas, or content that presented little to no challenges, and high-challenge areas, or content that presented more challenges. Areas were considered “high-challenge” when (a) translators encountered difficulty in translation that
required discussion of several alternatives with the entire translation team, or (b) more than two pre-testing participants showed responses indicating unintended interpretations as they rephrased content or discussed their thought processes while selecting a response option. Characteristics of both types of content are described along with frequencies of challenges encountered within the FYIv3.1. High-challenge areas are organized into thematic categories. In addition, we have provided at least one detailed example for each of the high-challenge areas to clearly illustrate the multiple layers of challenges identified by this process. Table 4.1 contains a summary of the patterns encountered.

**Low-Challenge Areas**

There were 12 items that did not present significant challenges in translation. These items contained 13.3 words on average, whereas high-challenge items contained an average of 17.7 words (t = 2.8, p < .01). Low-challenge items typically described simple concepts and lacked complex syntax or language structures. An example of a low-challenge item was, “Does your child have difficulty falling asleep?” The three forward translators each independently produced an identical translation of this item, the wording appeared equivalent in the same-language comparison, and the item did not produce any unintended interpretations in any round of pre-testing.
Table 4.1. Patterns of challenges encountered during the translation and cultural adaptation process.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
<th>Threats to Validity</th>
<th>Solution</th>
</tr>
</thead>
</table>
| Response options                      | “Never” “Never” “Sometimes” “Always”                                    | • Increased likelihood of respondents leaving items blank or answering on the wrong end of the scale  
  • Reduction in variability of data                              | • Add “yes/no” at the beginning of response options  
  • Label all response options                                     |
| Complex grammar/complex actions       | “Does your child seem unaware of a moderately painful experience like bumping his/her head?” | • Increased likelihood of respondents misinterpreting items  
  • Increases error in data                                           | • Simplified syntax and/or further description                         |
| Hesitation to report lack of typical milestone acquisition | Items containing well-known milestones, such as pointing. | • Increased likelihood of respondents reporting more skills than their child currently exhibits  
  • Deflates risk score                                                 | • Add information to instructions                                      |
| Reversed positive response bias       | Items describing atypical behaviors, especially when they contain only neutral language. | • Increased likelihood of respondents reporting more red flags than their child actually exhibits  
  • Inflates risk score                                                | • Add description  
  • Add information to instructions                                      |
| Reverse-answer                        | Items describing atypical behaviors, especially when they contain a negative term such as “difficult.” | • Increased likelihood of respondents answering in reverse  
  • Inflates risk score                                                | • Bold/underline key words                                           |
| Emotional States/Degree of Intensity  | “very distressed” “particularly interested”                              | • Increased likelihood of respondents misinterpreting items  
  • Increases error in data                                            | • Identify terms that result in most accurate interpretation              |
| Descriptions of actions               | “stick out your tongue” “shake your head” “look with his/her head sideways” | • Increased likelihood of respondents misinterpreting concepts  
  • Increases error in data                                            | • Identify terms, descriptions, or examples that result in most accurate interpretation  
  • Use culturally sensitive actions                                    |
| Spanish language (Noun/Pronoun use)   | Usted vs Tú Third person vs second person formal conjugation            | • Increased likelihood of respondents misinterpreting concepts due to pronoun confusion  
  • Increases error in data                                            | • Repeat names of nouns/pronouns within text of items                  |
High-Challenge Areas

A large majority of content presented challenges in translation. These challenges included the title, instructions, answer choices, and 83% of items. The vast majority of challenges were identified during one of the pre-testing rounds, though some challenges were identified either during the forward translation synthesis discussion, back-translation same-language comparison, or consensus meetings. These identified difficulties ranged in type and level of impact on risk score outcomes. The majority of the challenges encountered fell into one of the following categories, which we will use to organize the description of our results: 1) Complex syntax or complex actions, 2) Response options, 3) Positive response bias, 4) Emotional states/Degrees of intensity, 5) Descriptions of actions, 6) Spanish language challenges. In addition, there were numerous miscellaneous challenges which we will describe as well.

Complex Grammar or Complex Actions

While short, simple sentences or simple actions did not pose challenges in translation, content with complex syntax or content describing complex actions often proved difficult to translate. The instructions for the FYIv3.1 contain long sentences with multiple phrases and embedded clauses, making these difficult to translate.

One example of an item containing complex language is the following: “If you say to your child ‘Where is ___?’ and you name a familiar person or object without pointing or showing, will your child look at the person or object named?” This item contains 30 words with multiple phrases and embedded clauses. It describes an action with multiple steps and qualifiers. Our first relatively direct translation proved difficult for target participants during pre-testing rounds (misinterpretations among 50% of pre-testers). We modified the translation by changing the initial verb to “pregunta/ask,” combining the first and second clauses, and simplifying the
punctuation. These modifications resulted in more accurate interpretations among target population members (only 17% misinterpreted).

Another particularly challenging item read “Does your child seem unaware of a moderately painful experience like bumping his/her head?” The term “unaware of” does not have a simple direct translation to Spanish, and the phrase “no darse cuenta de/not to notice” was chosen as the closest translation. However, this syntax required participants to mentally perform a double negative; i.e. if their child did notice painful experiences, they should respond “never.” This mental switch proved significantly difficult for participants, 50% of whom misinterpreted the item or failed to respond appropriately. In order to address this challenge, the syntax of this item was simplified and the child action was changed to “ignorar/ignore.”

**Response Options**

During pre-testing rounds, we encountered two difficulties in the way participants chose response options. Our first synthesis translation used a direct translation of the five response option terms (nunca/never, a veces/sometimes, siempre/always), and maintained the lack of labels for the two intermediate options.

The first notable challenge here was that respondents frequently did not choose from the full range of options. When presented with three labeled options, they chose labeled response options 92.8% of the time, and unlabeled response options 7.2% of the time. Several participants would answer “almost never” aloud in response to some items, then select “never” on paper. When directly asked, participants reported either that the three labeled options were sufficient or that they did not know when to use the unlabeled responses. Comparatively, participants taking the English FYIv3.1 for a normative study selected labeled response options 65.9% of the time, and unlabeled response options 34.1% of the time (Chen, Davis, & Watson et al., 2018). Adding
labels to the intermediate options in the translation (casi nunca/almost never; casi siempre/almost always) improved the variability in responses, where intermediate responses, now labeled, were chosen 24.1% of the time (see Table 4.2). A 2x5 chi-square analysis showed a significant difference in the frequency of selection for each response option between those who were presented with three labeled response options or five labeled response options ($X^2 = 71.3; p < .001$).

Table 4.2. Frequency of responses for each response option selected by participants who were presented with either 3 or 5 labeled response options.

<table>
<thead>
<tr>
<th></th>
<th>3 response options labeled</th>
<th>5 response options labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never (Almost never)</td>
<td>Sometimes (Almost always)</td>
</tr>
<tr>
<td>Never</td>
<td>27.6%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Sometimes</td>
<td>18.4%</td>
<td>10.8%</td>
</tr>
</tbody>
</table>

The second response option challenge we encountered was related to syntax. All FYIv3.1 items are written as yes/no questions (see examples above). There is some clarification in the instructions that prompts caregivers to report frequency-based responses; however, pre-testing participants had significant difficulty making a switch from their first “yes” or “no” reaction to the frequency-based scale. Participants would often rephrase an item well, confidently describe their child’s behavior, then express substantial confusion over which response option to choose. To address this challenge, we added “yes” or “no” to the beginning of each response option as a transition for respondents to move from the syntactic yes/no question to the frequency-based options (i.e. “no, nunca; no, casi nunca; si, a veces; si, casi siempre; si, siempre”).

**Positive Response Bias**

Participants’ reactions and responses to items seemed to be influenced in part by a desire to report typical development in their child. That is, participants chose responses not only by
considering their child’s current behaviors, but also by considering which response option they perceived would indicate typical development. Evidence of this positive response bias appeared in two patterns, including a hesitancy to report the lack of typical behaviors and general difficulty with items describing atypical behaviors.

**Hesitation to report low frequency of developmental skills.** When reading items, participants often considered whether the behavior described was developmentally appropriate for their child’s age. If they perceived their child to be too young for a certain behavior, they sometimes followed one of two patterns. First, participants modified the meaning of the item so that it described a similar, but less developmentally advanced skill, then responded to the modified question. For example, one mother stated that her 15-month-old child was not yet pointing with his finger, but that he was able to point “with his eyes,” selecting “always” as her response. A less common reaction involved answering in anticipation of the child gaining that skill. Another mother whose 8-month-old child was not pointing commented that he would “probably start doing that soon,” and selected “sometimes.”

**Difficulty with items describing atypical behaviors.** Participants also exhibited a positive response bias with items describing atypical behaviors, though they often continued to show a bias toward the more frequent side of the scale, i.e. a reversed positive response bias. A large number of participants assumed that all items in the FYIv3.1 would describe behaviors that are present in typical development; however, many items describe behaviors that are not typical in development. These are often related to sensory or play behaviors, for which a high frequency of occurrence would indicate “red-flags.” Many participants showed difficulty rephrasing these items as compared to items describing typical behaviors. When rephrasing, they frequently re-conceptualized the item so that it described a behavior that would be considered positive or less
atypical. For example, one item asks whether the child plays with parts of toys rather than the whole toy. Many participants rephrased this item as asking whether their child was able to focus on and examine toys, or if they were curious about new toys, then answered on the more frequent side of the scale. This reversed positive response bias pattern was evident on 52.4% of all items describing atypical behaviors. Among these items, participants exhibited this re-conceptualization an average of 24.3% of the time. Items resulting in fewer re-conceptualizations contained words with negative connotations, such as “distressed,” “difficult,” or “have trouble” (participants re-conceptualized these 19.1% of the time). Items resulting in more re-conceptualization contained words with more neutral connotations, such as “prefer to,” “spend a lot of time,” or “interested in” (participants re-conceptualized these 28.8% of the time).

Answering items in reverse was another trend noted among items describing atypical behaviors. For 8 items (38.1% of all items describing atypical behaviors), participants rephrased items accurately and described their child’s behavior confidently, then selected an answer that would indicate the opposite of their description. For example, one item asks whether the child’s facial expressions are difficult to understand. Thirty percent of the time, participants stated that their child’s facial expressions were not difficult to understand and then selected a response option on the more frequent side of the scale, indicating frequent difficulty. Similar to items likely to result in lower levels of reversed positive response bias, five of these eight items contained words with negative connotations, such as “refuse,” “have trouble,” or “difficulty.” While the remaining three items contain only neutral terms, they each have clear semantic alternatives. Participants reverse-answered 12.1% of the time on average among these 8 items. Three of these eight items also resulted in reversed positive response bias. Interestingly, participants often appropriately interpreted and responded to eight other items describing
atypical behaviors. These better-understood items typically used simple syntax, described simple concepts or actions, and/or contained terms with clear negative connotations.

**Emotional States/Degrees of Intensity**

We encountered considerable semantic challenges when translating items that included descriptions of emotional states or degrees of intensity. Many early symptoms of autism relate to sensory reactions that caregivers may perceive as changes in emotional states. For example, the FYIv3.1 specifically inquires whether children appear “distressed,” “interested,” or “concerned” in various scenarios or situations. Additionally, many items specify a degree of intensity, such as “very,” “a small part,” or “easy.” Some items include both emotional states and degrees of intensity, such as “…seem particularly interested in….” Results from pre-testing revealed many unintended interpretations of these types of terms. One item begins “Does your child get very distressed when he/she…?” The most direct translation of “distressed” is “angustiarse,” but this word is uncommon and was not initially understood by all preliminary pre-testing participants. Our second draft version included the verb phrase “se molesta mucho/get very upset.” However, participants who were presented with this version frequently perceived “se molesta mucho” to be less intense of a reaction than our original intent of the phrase “very distressed.” One participant remarked “Nunca se molesta o se enoja, pero se asusta/He never gets upset or mad, but he gets scared.” This participant then responded “never” to this item. To correct for this unintended connotation of “molestarse,” the original verb “angustiarse/get distressed” was used, in addition to “o se asusta mucho/or get very scared.” This provided a synonym for the more precise but less common verb to aid in accurate interpretation.
Descriptions of actions

A quarter of all challenges encountered related to descriptions of actions, where several rounds of edits were required in order for participants to accurately and consistently interpret the intended action. This occurred frequently for items describing play with toys, many of which describe atypical behaviors. For example, one item asks, “Does your child look at toys in unusual ways (for example, stare, look with his/her head sideways, or look from the corner of his/her eyes)?” Our initial, fairly direct translation frequently resulted in re-conceptualization of this behavior to a more typical behavior. Specifically, one mother commented that “Quiere saber si le interesan/It wants to know if he thinks they’re interesting.” Further, the direct translation of “from the corner of his/her eyes” was poorly understood by many participants and required modification. To emphasize the nature of the actions, we added a synonym for “unusual,” re-ordered the examples so that the most unusual example appeared first, and added “in a strange way” to the “stare” example.

We also encountered several culturally-specific challenges among items describing actions. Two items ask about turn-taking skills, both of which use “peek-a-boo” as an example. Two of our forward translators suggested “cucú” or “cucú tras” as equivalent games in their regional culture. Because not all regions of Latin America use this term, our initial synthesis version included the phrase “…juegan en dónde estás (cucú)/play where are you (cucú)…” in addition to another turn-taking example (rolling a ball back and forth). During our initial round of pre-testing, our Peruvian participant described the term “cucú” as referring to an evil mythical character, similar to the “bogeyman” in North American English. In order to avoid this unintended reaction, we identified another phrase to describe an equivalent experience, “dónde está el bebé?/where is the baby?” For individuals from regions who would still not be familiar
with the game, we added a description of the experience to go along with the phrase. Results from pre-testing with this version were much improved, both among those familiar with the game “dónde está el bebé?” and those unfamiliar with it.

The concept of pretend play, referenced in two FYIv3.1 items, was also difficult to translate. The items include the verb “pretend” and provide examples, such as pretending to cook. We attempted multiple translations using the verb “pretender” that were frequently poorly understood. While some participants did understand the intended concept, many others understood the meaning of “pretender” to be “to intend” or “to try.” For example, one participant rephrased, “Como que él lo hace, lo intenta hacer/Like he does it, he tries to do it.” This false cognate confused the intent of the item for many participants. Further, after clarifying the concept to participants who misinterpreted the item, many explained that they understood this concept as imitation of daily activities, e.g., “Yo digo de imitar, imitar acciones del diario vivir./I call it imitating, imitating daily life actions.” However, the term “imitar/to imitate” was already being used in other items that specifically inquire about immediate imitation skills. Instead, we used the term “jueguitos imaginarios/little imaginary games” to describe the concept. We then used “pretende” for the examples, with a syntax that would allow for a caregiver to interpret the verb as “pretend” or “try” and still understand the intent of the item.

One of the items that inquires about immediate imitation skills presented another culturally-related challenge. The item uses “stick out your tongue, clap your hands, or shake your head” as examples of imitating adult behaviors. During preliminary pre-testing, participants became visibly uncomfortable after reading our fairly literal translation and stated that it would be rude for them or their child to protrude their tongue, including during play. Because of the negative cultural connotation of this action, this example was changed to an action involving the
same body parts but without a negative connotation: “abrir y cerrar la boc/ open and close their mouth.” Further, participants frequently understood “sacudir la cabeza/shake your head” as a rapid, repetitive movement, like that of shaking a bottle, rather than the intended “no” gesture that it refers to in English. Therefore, this verb was changed to “decir no con la cabeza/say no with their head” which is the phrase used to describe rotating the head back and forth. During later pre-testing rounds, these actions were both consistently understood as the intended actions, and they appropriately exemplified the conceptual meaning of the item.

**Spanish Language Challenges**

Most Spanish language challenges that we encountered related to use of pronouns. Spanish has two levels of second-person pronouns according to formality, each with separate verb conjugation. Forward translators chose to use the formal version (“usted”) to refer to the caregiver so as to avoid offending any Latin American cultures that prefer this version. In Spanish, pronouns are often left unstated; however, the verb conjugations for “usted” mirror the verb conjugations for third person nouns/pronouns. Therefore, it was necessary to repeat pronouns that may have ordinarily gone unstated in order to ensure understanding. Similarly, third person pronouns are equivalent for objective nouns in Spanish, whether referring to objects or people. For example, one item reads: “If you show your child a new way to play with a toy (for example, turn it upside down or put it on your head), will your child try that new action with other toys?” If we simply used the pronoun “it” in the examples, some participants interpreted this pronoun as referring to the child (i.e. turning the child upside down or putting the child on the adult’s head). It was necessary to restate the noun (“the toy”) in the examples rather than using a pronoun.
*Miscellaneous Semantic/Syntactic Challenges*

A plurality of challenges (28.1%) consisted of semantic and syntactic difficulties that did not specifically fall into one of the previously described categories. We describe two examples of these below.

One short item probes whether the child shows objects to the adult. Participants misinterpreted a fairly direct translation 75% of the time. The vast majority of these misinterpretations involved conceptualizing “mostrar/show” as any communicative function, including crying to get help or dancing to draw an adult’s attention. In order to convey the intended conceptual meaning, we added an example to the end of this item. “Por ejemplo, ¿le acerca un objeto para mostrárselo?/For example, does s/he bring an object to you to show it to you?” Despite the addition of the example, this adaptation resulted in similar misinterpretations 33% of the time. However, this was considered a significant improvement and the final version retained the example. Examples or additional descriptions were also added to three other items in order for pre-testing participants to consistently interpret items accurately.

The title of the FYIv3.1, First Years Inventory, was originally chosen in part for the aesthetics of the acronym it produced in English (G. Baranek, personal communication, Dec. 3, 2018). When translated directly, *Inventario de los Primeros Años*, the resulting acronym (IPA) does not contain any meaning in Spanish. Additionally, the term “inventario” produced unintended interpretations related to material consumption among pre-testing participants. We pre-tested *Cuestionario de los Primeros Años*/First Years Questionnaire, which resulted in more accurate understanding.
Discussion

This article describes the process of a rigorous translation and adaptation methodology using the FYIv3.1 as a test case. We encountered numerous challenges that negatively impacted respondent interpretation of items; these required significant deviation from the wording of the original text in order to maintain linguistic, construct, and technical equivalence between translations. We anticipate that the results and details provided herein will help guide future teams undergoing this process with similar instruments.

Challenges Identified

Complexity

A small percentage (19%) of items presented little to no challenges in the translation process. These low-challenge items typically contained few words and simple syntax, whereas longer items with more complex syntax often caused confusion and required simplification or additional descriptions. Indeed, many authors note that increased language complexity makes translation more difficult (Acquadro et al., 1996; Bracken & Barona, 1991).

Response Options

Adaptations were required to maintain technical equivalence between versions with respect to the response options. A direct translation of the response options presented two threats to technical equivalence of the two versions. First, participants exhibited difficulty moving from the “yes/no” question to the frequency-based response, leading to blank or inaccurate responses. Second, participants failed to choose unlabeled response options, effectively reducing the scale from a five-point scale to a three-point scale and significantly reducing the variability in the data. Such a trend would likely negatively impact the FYIv3.1’s predictive validity. Multiple authors note the need to ensure technical equivalence between response scales across two cultures,
including an analysis of the frequency of endorsement for each option (Acquadro et al., 1996; Stewart & Napoles-Springer, 2000; Ware et al., 1996).

**Positive Response Bias**

Some participants were hesitant to report when their child had not achieved a milestone in typical development, especially if they felt their child was younger than the typical acquisition age for the target skill. In a study comparing observations to parent-interview responses, Blacher et al. (2014) found that Latino caregivers similarly over-reported achievement of milestones. While these patterns were relatively rare in this sample, the impact of these patterns was concerning enough to address through modification of the instructions.

Participants assumed that all items described typical behaviors and re-conceptualized over half of the items describing atypical behaviors so that they sounded more typical. This reversed positive response bias was more frequent when items contained neutral language (e.g. “interested in”) as opposed to negative language (e.g. “have trouble”). Among an Israeli population with a Hebrew FYIv2.0 translation, Ben-Sasson and Carter (2012) reported a significantly higher endorsement rate for several of these items (e.g. body feels loose, playing with parts of toys) as compared to normative data from the English version. In fact, most itemized differences between the Israeli and US data occurred within the sensory-regulatory domain, which contains many of these items describing atypical behaviors with neutral language. Ben-Sasson & Carter (2012) concluded that caregivers may not interpret these items as describing concerning behaviors, which was evident in our data. Similarly, studies of forward-back Spanish translations of the M-CHAT, a similar ASD screening tool, have also reported higher than expected endorsement rates of items describing atypical behaviors (Guthrie et al., 2019; Kimple et al., 2014; Windham et al., 2014).
Less frequently, caregivers reverse-answered items describing atypical behaviors, such that they accidentally answered on the more frequent side of the scale. This may have been an artifact of the cognitive interview format. As participants discussed the item aloud, they may have unintentionally reversed their understanding of the item before responding on paper.

“Reversed items” are items that require a reversal in scoring to be consistent with the direction of the rest of the items in a scale, which is the case with the items describing atypical behaviors in the FYI. Reversed items are often found to be problematic in measurement research (Suárez-Alvarez et al., 2018; van Sonderen et al., 2013; Weijters et al., 2013). We modified the font on key terms within items most likely to be reverse-answered, added information to the instructions, and provided additional descriptions to reduce these various positive response bias patterns.

**Emotion/Intensity**

Items that contained terms related to emotional states or degrees of intensity required multiple rounds of pre-testing and edits to achieve our intended interpretation among a majority of participants. Concepts related to emotions are delineated differently by lexicons across languages and cultures (Jackson et al., 2019), complicating translation. Guillemin (1993) and Ware et al. (1996) caution that descriptions of social and emotional concepts often require adaptation in order to maintain construct equivalence.

**Actions**

Items containing descriptions of actions often required adaptation to achieve appropriate understanding. These challenges presented themselves in many items involving gestures and play, and introduced the potential for threats to linguistic as well as construct equivalence. Due to the nature of early symptoms of ASD, many FYIv3.1 items describe small but significant
details of behaviors and thus require multiple verb phrases and descriptions, which complicate syntax. In addition, verb phrases often vary by dialect, where slightly different wording would result in differential interpretation across participants, either linguistically or conceptually. For example, when being asked about their child’s imitation skills, if participants interpret “sticking out your tongue” as an inappropriate behavior, they would be likely to select a less frequent response for this item, increasing their risk score. This item would in effect measure the construct of their child’s behavioral compliance rather than their imitation skills.

Similar to our challenges noted in items describing pretend play, difficulty with this concept was also apparent in the Spanish translation of the M-CHAT-R/F (DuBay & Watson, 2019a). Additionally, there is some literature to suggest that pretend play is less common among children from Latin American cultures (Farver & Howes, 1993; Göncü et al., 2000), which may indicate a cultural difference in recognizing and promoting this behavior among children.

Similar to “pretend play,” “peek-a-boo” does not have a direct Spanish translation. Our study’s translators, experts, and participants identified similar infant games and a range of possible terms, though no universally understood term was identified. The terms “cucú-tras,” and “ojitos” are used in other Spanish translations of similar developmental screening tools (Bricker & Squires, 1999; Canal-Bedia et al., 2011). However “cucú” received poor pre-testing responses in our study and our team of experts from a range of regions were not all familiar with ojitos. Albores-Gallo et al.’s (2012) Spanish translation of the M-CHAT for Mexico reported that no term existed for the concept of “peek-a-boo” in Mexican culture and chose to simply provide a description.
**Spanish-Language Challenges**

We identified some linguistic challenges related to the nature of the Spanish language. Specifically, pronoun use and verb conjugations in Spanish often made it necessary to restate nouns rather than using a pronoun, which added redundancy to the translation but clarified the subjects and objects of sentences.

**Miscellaneous Challenges**

While many challenges encountered fell into one of these categories mentioned above, 28% of the challenges did not. These miscellaneous semantic and syntactic problems were either identified during the initial forward translation or led to misinterpretations in pre-testing rounds, requiring changes and adaptations to maintain linguistic and conceptual equivalence.

**Impact on Psychometric Properties**

If left unaddressed, some of the challenges presented here would falsely inflate the risk score, while others would falsely deflate it. For example, a positive response bias to items describing atypical behaviors would inflate the risk score, while hesitation to report the lack of typical milestones would deflate the risk score. Many other challenges contributed to general confusion or misinterpretation among the participants, the psychometric impacts of which are harder to predict. All caregivers encounter all of these types of items when completing the questionnaire. Therefore, if the challenges had been left unaddressed, each risk score would have been influenced by the sum impact of all of these patterns. Their combined effect would inevitably contribute to overall error in the data, both within and across respondents, at the expense of sensitivity, specificity, positive predictive value, and negative predictive value. Theoretically, re-norming the items such that local norms reflect some of these patterns could mitigate unaddressed linguistic or cultural differences between versions. However, most patterns
led to general confusion among participants, which could sway responses either positively or negatively across participants, making statistical correction impossible.

**Steps in the Translation and Adaptation Process**

Given the field’s continued use of the forward-back approach, it is worth noting the impact of each step in this adaptation process, including which contributed most to the identification or potential prevention of misunderstandings and misinterpretations.

**Forward Translation**

With the use of multiple translators, threats to linguistic and construct equivalence were identified during this phase. Independent translations often differed considerably across translators for the same source content. When synthesizing their translations, the team rarely chose any one translator’s independent translation, but rather combined words and phrases from multiple translators to create simple and accurate translations. This process allowed for prevention of idiosyncratic language patterns from any one translator (Guillemin, 1993; Wild et al., 2005), though it is difficult to quantify how many challenges may have been prevented. This process also facilitated consideration of multiple dialects during the synthesis meeting and provided multiple alternatives for each item when challenges were identified later in the process.

**Back Translation**

Back translators accurately identified eight errors or problems in the forward translation, all of which related to linguistic equivalence. Three of these problems had already been identified in preliminary pre-testing. Of the remaining five errors, one did not alter construct equivalence (i.e. waving “hello” instead of waving “goodbye” as an example of a gesture), and the rest represented minor construct differences (e.g. “cup” instead of “sippy cup”). However, back translators themselves made errors in translation that would have inaccurately identified 19
additional items as problematic. These errors ranged from minor grammatical errors to altering the conceptual meaning of items. Other studies have also noted the likelihood of back translation errors resulting in a false impression of problems in the forward translation (Behr, 2017; Colina et al., 2017). More importantly, the vast majority of challenges encountered in this study, some of which represented significant threats to construct equivalence, were missed by back translation. For example, the back translation of one item resulted in the exact wording of the original instrument, leaving the impression of a perfect translation. However, no participants were able to interpret this item accurately during preliminary pre-testing. This pattern of a false positive impression of the forward translation has been documented in other studies (Behr, 2017), and was prolific throughout our instrument.

**Consensus Meetings**

This step addressed linguistic, construct, and technical equivalence. During consensus meetings, multiple options for edits could be discussed, speakers of multiple dialects could be consulted, and concepts could be visually represented (i.e. acted out) to ensure that chosen language matched conceptual intent. Original FYI developers had to clarify the intent of some items to ensure accurate translation. Preplanning, which consists of formally clarifying item intent and providing additional information for translators prior to the start of the forward translation (Acquadro et al., 1996; Behr & Scholz, 2011), would have prevented the need for this clarification during the consensus meetings; however, we did not include this as a part of the translation and adaptation process. Of note, back translation was inevitably interwoven throughout the consensus meetings to facilitate full understanding and participation for all attendees, as some attendees did not have Spanish fluency.
Pre-testing

Pre-testing proved to be the most fruitful step in the process, identifying the vast majority of challenges (81%), spanning linguistic, construct, and technical equivalence. As participants shared their reactions aloud, we were able to directly examine how target population members would understand and respond to the questionnaire in a natural setting (Beaton et al., 2000; Guillemin, 1993; Stewart & Napoles-Springer, 2000). Due to the nature of the ongoing cognitive interview format, we were able to test multiple rounds of edits and discuss alternatives with individual participants.

Limitations

We were unable to represent all Latin American Spanish dialects or cultures among our translators and participants. However, we did have representation from a range of countries, which reflected the demographics of Latin Americans living in the US (Motel & Patten, 2012). In addition, the cognitive interview format does differ from the intended administration format for parent-report tools (i.e. independent completion). We therefore may have encountered a slightly different set of challenges than those encountered through independent completion alone.

Conclusions

This work presents a case study of the process of translation and cultural adaptation using a specific ASD screening tool. We have illustrated the need for considering linguistic, construct, and technical equivalence in the translation process and for increasing rigor within a cultural adaptation approach. Our results show the clear importance of pre-testing as a means of directly assessing the understanding of target population members, and we urge future translation teams to utilize pre-testing methods in the translation process. Our results also illuminated the potential pitfalls of back translation methods, including both false positive and false negative impressions.
of the translation. We recommend that future translation teams avoid relying on back-translation as the sole means for quality assurance.

We have also identified specific areas of the FYIv3.1 that were in most need of cultural adaptation for a Spanish-speaking Latin American population. Specifically, response options, instructions, pronouns, and items describing atypical behaviors, emotional states, degrees of intensity, and actions all presented significant challenges in the adaptation process. Special attention should be given to these areas as potentially problematic in translations and adaptations of other parent-report tools containing similar constructs. Further, this work illustrates the importance of implementing this adaptation process fully, as a plurality of the problems identified did not fall within any of the aforementioned areas. They represented a range of miscellaneous threats to linguistic equivalence that would have contributed to general confusion and misinterpretations had they not been addressed.

**Future Directions**

For the newly translated and culturally adapted Spanish version of the FYIv3.1, normative data on response distributions to each item are now being collected on a large sample of target population members. Future work will assess diagnostic outcomes when children reach an age in which valid ASD diagnoses can be made. We hope to evaluate the predictive value of different items at different ages, and then generate and validate a shorter Spanish version of the FYI for use in research and clinical settings.

More broadly, our findings in the area of back-translation and other translation procedures suggest the need for an empirical comparison between the traditional forward-back translation approach and this more rigorous translation and cultural adaptation approach through psychometric analysis. Future research could also examine how findings from this
implementation may differ for other language and cultural groups, as well as how this process may differ for tools intended to be administered by professionals instead of parents. Finally, this work also serves to promote transparency in the translation process for research measures in this field. Authors using translated instruments should provide clear methods for their translation process so that readers are able to adequately judge their validity.
REFERENCES


DuBay, M., & Watson, L.R. (2019a). Does the M-CHAT-R/F work the same in Spanish as it does in English? Poster presented at The American Speech-Language-Hearing Association Convention, Orlando, FL.

Disorders, 62, 55–65.


Robins, D., Fein, D., Barton, M., & Green, J. (2001). The Modified Checklist for Autism in


