THE GRADUATE COLLEGE OF

THE UNIVERSITY OF OKLAHOMA

PSYCHOMETRIC PROPERTIES OF THE SPANISH-TRANSLATED TRANSITION ASSESSMENT AND GOAL GENERATOR (TAGG)

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

BELKIS CHOISEUL-PRASLIN

Norman, Oklahoma

2021

PSYCHOMETRIC PROPERTIES OF THE SPANISH-TRANSLATED

TRANSITION ASSESSMENT AND GOAL GENERATOR (TAGG)

A DISSERTATION APPROVED FOR THE

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Kendra Williams-Diehm, Chair

Dr. John Jones

Dr. Maeghan Hennessey

Dr. Howard Michael Crowson

Dr. Rachel Carr

Dedication

Para mis abuelos;

Jorge y Doris Choiseul-Praslin
y
Dolores y Rosaura Membreño

Acknowledgements

A mis padres; Byron Sr. y Janeth. Cualquier meta, logro o reconocimiento que he recibido o reciba en el futuro es solamente posible por sus sacrificios y ejemplos. Ustedes dos son mis héroes e inspiración. Agradezco su amor y apoyo en todos mis metas. Ser su hija es mi mayor alegría y espero hacerlos orgullosos.

To my brother and best friend; Byron Jr. Thank you for always picking up the phone, for being there unconditionally every time I needed you, and for advising and guiding me on all matters big and small. I appreciate your candor and encouragement more than you know.

To my family; Jessica, Bryanna, Sophia, Byron III, and my Choiseul-Praslin and Membreño tio’s, tia’s, and primos. I have carried you all in my heart every moment I’ve been away. I love you all so much and hope to make you proud.

To my dearest friends; Sheena, Marilyn, Nicole, Grace, Genevra, and Stephanie. I admire you all far more than I can put into words. Thank you for the constant encouragement, ridiculously wonderful trips, and laughs. Your friendship means the world to me.

To my ZC friends and family; Tracy, Malarie, Mindy, Heather, Joshua, and Melissa. I am eternally lucky to have learned from and worked next to you for the better part of four years. You have shown me the many ways one can be smart, strong, professional, observant, and decisive. I consider you all my inspiration and am fortunate enough to call you my friends. I would be remiss if I did not acknowledge Drs. Martin and McConnell – our time working together may have been short, but your impact in my life and work have been significant. I could not have completed this study without your work and dedication to the field, thank you.

Finally, to my committee. First, Dr. Kendra William-Diehm, my doctoral advisor - from suffering long meetings during my first year just to help me get involved in organizations to

supporting every wild idea I've ever had since - you have been the most integral part of this academic journey. I truly appreciate your guidance and confidence in my abilities and cannot thank you enough. Drs. Maegan Hennessey and Mike Crowson, I am always in awe of your knowledge and skills. I would be completely lost without your guidance in this study and thank you both for your time and attention in helping me finish. Drs. John Jones and Nicole Carr, thank you for offering kind words of encouragement when needed. Lessons learned (and there were many) in your courses have informed much of the work done for this study.

Table of Contents

List of Tables

List of Figures

Abstract

Use of valid and reliable transition assessments is crucial to the transition planning process in the individualized education program (IEP). However, few transition assessments are linguistically inclusive and even less have sufficient validity evidence for translated versions. As it stands, there are no transition assessments that accurately measure the strengths and needs of English language learners/students with disabilities (ELSWDs) in the U.S. This is especially concerning as students from linguistically minoritized backgrounds are among the fastest growing school-age populations, including ELSWDs. Thus, the purpose of this study was to evaluate the psychometric properties of the Spanish-translated student version of the Transition Assessment and Goal Generator (TAGG). The study originally aimed to establish measurement equivalence/invariance (ME/I) evidence for the translated TAGG-S but when ME/I tests could not proceed indicating the TAGG factor structure does not hold for Spanish-testing students, alternative factor structures were explored through a series of exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) tests using three different statistical software programs. Use of multiple software programs allowed for unique model specifications and comparisons. After testing numerous models, a final six-factor model fit the data best. While an alternative factor structure was identified for the Spanish-translated TAGG-S, further testing is needed to ensure the structure holds. A formal evaluation of the translated items in the TAGG-S is also needed to ensure the assessment is appropriate to ELSWD test-takers. Findings from this study highlight the importance of validating translated assessments, particularly when assessment results are used to make educational decisions for students.

Keywords: transition assessments, ELSWD, assessment validation

Chapter 1

Introduction

The Individuals with Disabilities Education Act (IDEA, 2004) mandates transition planning for students with disabilities beginning, at least, by age sixteen. The directive to provide transition planning services directly corresponds to a primary purpose of IDEA which emphasizes that students must be given a free, appropriate, public education (FAPE) that ultimately prepares them for further *education*, *employment*, and *independent living* (300.1). Prior to the latest reauthorization of IDEA (2004), the President's Commission on Excellence in Special Education cited the failure of federal policies and programs to oversee a smooth transition to adult life as the overarching barrier preventing students with disabilities from succeeding once they left the school system (U.S. Department of Education [DOE], 2002). Consequently, the emphasis on further education, employment, and independent living became the central tenants of the transition plan in a student's individualized education program (IEP). Transition plans consider the long- and short-term goals of the individual student when planning for their post-school life as well as the services and supports needed to help them reach their goals. Research-identified best practices recommend that the goals and services written into the IEP's transition plan must be informed by age-appropriate transition assessment results (Deardorff, 2020; Petcu et al., 2014; Prince et al., 2014). Transition assessments are therefore viewed as an integral part of the transition planning process (IRIS, 2020; Sitlington & Clark, 2007).

The Division on Career Development and Transition (DCDT) of the Council for Exceptional Children (CEC) defines transition assessment as an ongoing process whose goal it is to facilitate the attainment of the student's post-secondary goals (Neubert & Leconte, 2013).

While this and other definitions exist, clarity surrounding the types of transition assessments and how many are required per year is needed. According to Prince et al. (2013; 2014), to avoid the complex and often legal aftermath of poor transition planning, students with disabilities must be assessed annually using at least two transition-related assessments, with one of those being designated as "formal." The terms "formal" and "informal" are often used across many school-based education-related disciplines to describe whether or not an assessment is data-based (Weaver, 2020). The National Technical Assistance Center on Transition (NTACT; 2016) provides some guidance to educators in distinguishing between formal and informal transition assessments: formal assessments are standardized instruments that include descriptions of the norming process, reliability, validity, and recommended use. Informal assessments are essentially the inverse, lacking a formal norming process and information on reliability and validity. While technically correct, the concepts of validity and reliability and their importance to transition assessments are vastly oversimplified in the NTACT transition assessment guide. These generalized definitions are echoed in similar national transition resources for educators (e.g., IRIS modules and Transition coalition modules). A simplified understanding of validity and reliability may be acceptable for daily classroom practices, but they are too broad for the high-stakes IEP. To provide additional clarity, one should refer to professional standards and guidelines for the development and administration of tests, *The Standards for Educational and Psychological Testing*, (hereafter referred to as the *Standards*; American Education Research Association [AERA] et al., 2014) which define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11) and reliability as "the consistency of scores across replication of a testing procedure" (p. 33). Based on these definitions, a transition assessment is deemed formal when there is acceptable associated evidence of validity and reliability to support its continued use. Validity and reliability have been at the forefront of educational and psychological measurement since the 1950's (APA, 1954; AERA et al., 2014). Theoretical frameworks and decades of related research (e.g., Cronbach, 1971; Kane, 2006; Messick, 1989) have informed the current version of the *Standards*. Among its pages, the *Standards* provide guidance for ensuring aspects crucial to establishing validity and reliability are considered and implemented when developing or administering a test. Transition assessments created within the last decade often cite the *Standards* as evidence of fidelity in their user manuals or publications establishing validity evidence (see: Martin, 2013; Martin et al., 2015; Shogren et al., 2017). The concepts of validity and reliability are further defined in chapter 2.

The 2014 *Standards* emphasizes the need to address and ensure fairness to all test takers in any given assessment. Given this consideration, fairness is seen as a fundamental validity issue. As an attribute of validity, fairness serves as the catalyst to re-evaluate current assessments that have previously been deemed to produce valid and reliable results for test takers. When the topic of fairness in testing is at hand, there are two larger concepts to consider: procedural fairness and substantive fairness (Kane, 2010). Procedural fairness is the requirement for fairness and validity in testing allotted to all individuals. One may refer to the American with Disabilities Act (ADA) for an example of procedural fairness in everyday life. Substantive fairness requires that the score interpretation and any test-based decision be reasonable and appropriate for all test takers. As the *Endrew F.* (2017) special education court case revealed, there is a need to meet the procedural requirement, but, unless the substantive requirement is met, the premise of fairness is not met. This concept holds true in educational testing as well. The attention paid to evaluating the fairness of a test is particularly important when considering the testing needs of students with disabilities from culturally and linguistically diverse (CLD) backgrounds, non-native English speakers, and English language learners (ELLs). Providing a translated version of a test for ELL students with disabilities (ELSWD) meets the procedural requirement as it ensures an individual can test in their native or preferred language. To meet the substantive requirement, the translated assessment must be validated to ensure the intended score interpretation and any decisions made resulting from test scores are accurate. If these considerations for procedural and substantive fairness are not taken into account, significant threats to the assessment's validity for ELSWD test takers likely exist.

Problem Statement

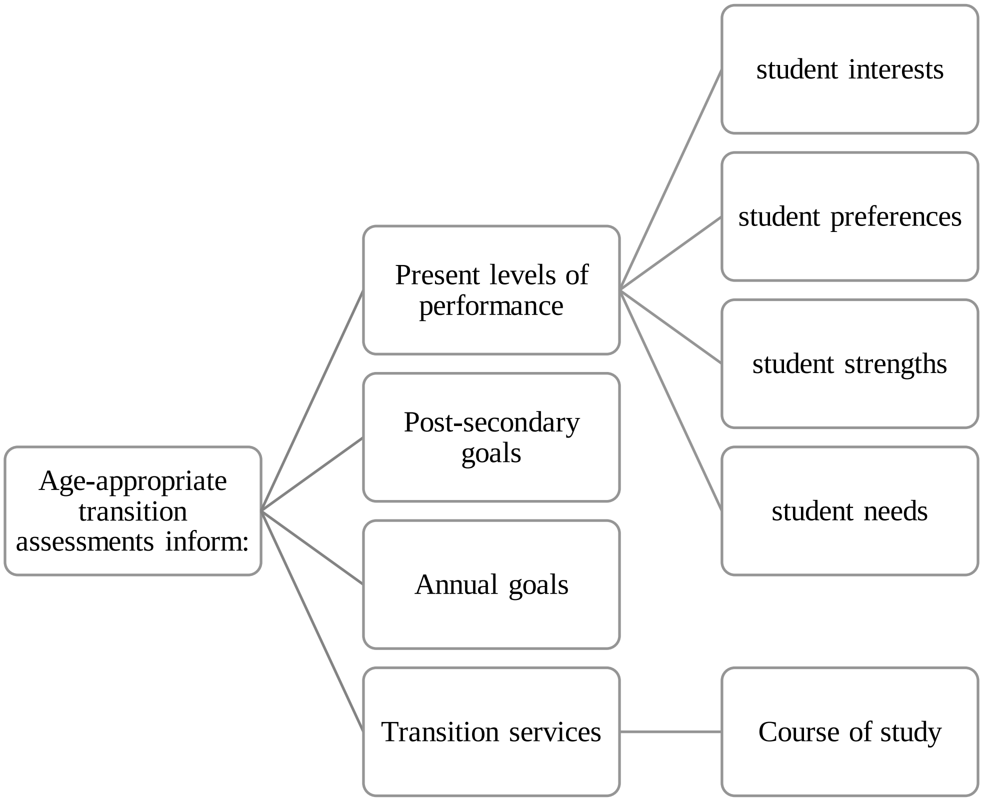
The fairness and associated validity evidence specific to ELSWDs is particularly important given their known post-secondary outcomes. As a whole, students with disabilities experience poor outcomes in the areas of further education, employment, and independent living when compared to their same-aged peers without disabilities (Banks, 2014; Blackorby & Wagner, 1996; Carter et al., 2012; Flexer et al., 2011; Grigal et al., 2011; Mithaug et al., 1985; Newman et al., 2009, 2011; Prince et al., 2018; Test et al., 2009). When the added factors of linguistically minoritized and ELL backgrounds are included, these outcomes worsen (Leake & Black, 2005; Oswald et al., 2002; Skiba et al., 2005; Trainor et al., 2016; Wagner et al., 2007; Zhang & Benz, 2006). Given the growing diversity observed and projected in U.S. student populations (National Center for Education Statistics [NCES], 2018), including those from linguistic minoritized backgrounds (Artiles et al., 2005), it is crucial that translated transition assessments are evaluated and validity is established for the translated versions.

While there are many transition assessments in use, when looking at comprehensive (testing more than one transition area) transition assessments, one will find a total of eight assessments, four of which are translated into a language other than English. To date, there are only two translated transition assessments with associated evidence of validity and reliability: the Spanish-translated Self-Determination Inventory (SDI) - student report (SR; Mumbard´o-Adam et al., 2018) and the Chinese-translated Self-Directed Search (SDS) form-R, 4th edition (Yang et al., 2005). However, the sources of evidence for both may not be appropriate to ELSWDs in U.S. schools as they were conducted outside of the U.S. and, in the latter's case, on an older version of the assessment which is no longer in use. To an extent, translated assessments address the consequential validity (i.e., the positive or negative social consequences of a particular test) of educational decision making for ELSWDs (Glen, 2020a). This is particularly important when considering the role transition assessments play in the IEP process. As is shown in Figure 1, they directly and indirectly contribute to every single key component of the transition plan in the IEP. When educators select a formal translated transition assessment for ELSWDs whose primary or preferred testing language is not English, any choice can be construed as faulty as none are validated for ELSWDs in the U.S. This faulty choice can trigger an evaluation of the appropriateness or accuracy of the transition plan in the IEP, leading to further negative social and legal consequences (Prince et al., 2013, 2014).

From the researcher end, assessment creators who are translating or have translated a test can negate issues of consequential validity by ensuring the measurement equivalence or invariance (ME/I) of its various translated versions (Zumbo, 2003). Establishing ME/I is particularly important when the goal is to ensure that comparisons across groups of participants, as are often seen in cross-cultural studies, are both meaningful and valid (Lee, 2018). Boer et al. (2018) assert that cross-cultural research is often driven by assumptions of differences and studies in this field aim at detecting those differences across samples from different cultural groups. However, they also note that score differences in these studies can be significantly misinterpreted if comparability across groups is lacking. Without tests of ME/I, we cannot assume the constructs an assessment aims to measure are accurately assessed across groups (Chan, 2011). As is evident by the lack of ME/I research in translated special education transition assessments, researchers often ignore invariance issues and compare latent factor means across groups even though the psychometric basis for doing so does not necessarily hold (Van De Schoot, 2015). Hence, a major need exists to add ME/I validity evidence to translated transition assessments specifically for ELSWDs in the U.S. Research in this area should assess if translated tests exhibit ME/I for all of its intended users (Zumbo, 2003).

Figure 1

Age-Appropriate Transition Assessments Inform Key Parts of the IEPs Transition Plan.



Note. Age-appropriate transition assessments are needed to provide students with an individualized and explicit plan of action. Transition plans that do not properly utilize both formal and informal transition assessments cannot meet Inidcator-13 requirements and violate provisions of IDEA.

Purpose of the Study

For this study, I evaluated the psychometric properties of the Spanish-translated student version of the Transition Assessment and Goal Generator (TAGG). The TAGG is a comprehensive transition assessment for secondary aged students (ages 14-22, grades 9-12) developed by Martin et al. (2015) at the University of Oklahoma. The TAGG was chosen as the assessment under inspection in this study as it ranks amongst the most popular transition assessments used in the U.S. (Martin, 2013), has over 50,000 completed assessments in its database, and has established Spanish translations. The 34 test items in the TAGG derive from research-identified non-academic student behaviors associated with positive post-school outcomes (Martin et al., 2015; McConnell et al., 2012). The TAGG consists of a professional (TAGG-P), family (TAGG-F), and student (TAGG-S) version; the TAGG-S and TAGG-F are translated into Spanish and Chinese. The TAGG is designed to assess the individual ability of students with mild/moderate disabilities across eight constructs. Below are summaries of the TAGG constructs as defined in the user manual.

- **Strengths and limitations** – the student’s ability to identify their owns strengths and weaknesses, academic and non-academic.
- **Disability awareness** – the student is aware of their specific disability and is able to explain it to others. The student knows what supports they need and can seek information to better understand their disability.

- **Persistence** – the student believes in their ability to face and overcome adverse situations. They adapt to the situation using lessons they have learned to keep making progress.
- **Interacting with others** – the student effectively interacts with individuals across school and community settings.
- **Goal setting and attainment** – the student is able to break large goals into smaller achievable tasks, adjusting as needed.
- **Employment** – the extent to which a student has had a paid job and aspires to have a job that matches their interests.
- **Student involvement in the IEP** – the student is actively involved in in their IEP meetings and can describe their current performance levels and post-secondary goals.
- **Support community** – the student can recognize people who provide positive support and only uses supports when needed.

When evaluating the TAGG one will find substantial validity evidence (Burnes et al., 2018; El-Kazimi, 2012; Hennessey et al., 2018; Martin, 2013; Martin et al., 2015) but none for its translations. As students from CLD Spanish-language backgrounds, including ELSWDs are among the largest projected student population in the U.S., this study focused on the Spanish-translated TAGG-S. Therefore, the purpose of this study was to evaluate the psychometric properties of the Spanish-translated TAGG-S. To that end, I sought to provide evidence of ME/I across groups (English-language and Spanish-language test takers), utilizing a structural equation modeling (SEM) framework, and confirm reliability of the TAGG's internal structure. Though establishing validity evidence can be done in multiple ways, ME/I procedures were chosen as the primary method in this study to first ensure the Spanish-translated TAGG-S is an appropriate assessment for the overall group of Spanish-language TAGG-S users. Studies seeking to explore score differences by characteristics of the group (e.g., by disability category, age, gender, etc.) or analysis of the professional and family responses patterns for the group can be completed after establishing the translated assessment is an appropriate tool.

Research Questions

The following research questions guided this study:

1. Are the constructs measured in the TAGG functioning equivalently across English-language and Spanish-language test-taker groups?
    a. Are the measurement parameters (factor loading, intercepts, and error variances) equivalent across groups, based on the weak, strong, and strict tests of equivalence?
    b. If measurement invariance is found, are there mean differences in latent factors?
2. Does the internal structure of the overall and subscale scores of the Spanish-translated TAGG-S meet acceptable standards of reliability as determined by Cronbach's alpha?

Significance of the Study

Though this is largely technical work, the significance of this study's findings is two-fold: first, adding validity evidence for the Spanish-translated TAGG-S directly addresses an area of need in special education and transition assessment research. No other translated transition assessment has standalone evidence of measurement invariance, making this study novel to the field. As previously stated, ensuring translated transition assessments not only exist but are also validated is of utmost need. Second, establishing ME/I validity evidence allows special educators to accurately assess the fast-growing linguistically diverse student population, including ELSWDs. This means educators and schools can not only avoid legal issues resulting from faulty transition plans but will be better equipped to create meaningful transition plans and identify appropriate transition services to support a sect of the SWD population who typically experience poor post-school outcomes.

Chapter 2

Literature Review

Assessments serve an important purpose in education; they allow for identification of needs and strengths and guide the education services students receive. Educators tasked with assessing students must be able to select appropriate assessments, correctly interpret and report results, and build educational services from the results. Selecting appropriate assessments for students can be daunting for those not familiar with the concepts of validity and reliability in educational testing or the importance of selecting validated assessments for students from linguistically diverse backgrounds. This holds true for the field of special education where assessments play a crucial role in the legally binding educational services listed in the student's individualized education program (IEP). To aid educators in selecting valid, reliable, and linguistically inclusive assessments, an understanding of validity and reliability must be established.

Several frameworks have been developed to support the use of educational and psychological assessments. These frameworks range from theoretical models dedicated to defining measures of validity (e.g., Cronbach, 1971; Kane, 2006; Messick, 1989) to professional standards and guidelines for the creation and administration of assessments (e.g., American Psychological Association [APA] et al., 1954; AERA et al., 2014; ITC, 2019). Reliability indices are often included in these frameworks to determine acceptable levels of various assessment properties. Together, validity and reliability are crucial and complex components of any educational assessment.

The definitions of validity and reliability are multifaceted and can vary depending on the context of the field, experimental design, or measurement procedures (Frey, 2018). For example, in the special education transition field, validity is often considered a property of an assessment (i.e., “this assessment has ample validity evidence”), but the extent to which an assessment is deemed valid lies in how the test scores are interpreted and reported (Popham, 2008; Messick, 1989). The oversimplification of validity used in special education may be due, in part, to the use of outdated definitions and lack of access to current educational measurement and evaluation research. The research-to-practice gap has been an area of concern for the field of special education for over 20 years (Carnine, 1997; Farley-Ripple et al., 2018; Greenwood & Abbott, 2001). Similarly, reliability is measured by a number of mathematical calculations, but the interpretation of the output is left to arbitrarily agreed upon levels of acceptability (Cortina, 1993). Thus, it is important to understand the concepts of validity and reliability through a historical lens to provide a modern evaluation of educational assessments.

Validity

The *Standards* define validity as “the degree to which accumulated evidence and theory support a specific interpretation of test scores of a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence for each interpretation is needed” (AERA et al., 2014, p. 225). Given this definition, the *Standards* (2014) further assert that validity should be the most fundamental consideration when developing and evaluating tests. Though this definition exists, there are concerns about its universality (Cizek, 2012; Sireci, 2009; Tiffin-Richards & Anand Pant, 2017). Throughout the history of psychological and educational testing, validity has been defined and redefined as a unitary concept and as a set of distinct elements, types, aspects, or categories (Sireci, 2009; Tiffin-Richards & Anand Pant, 2017).

Our understanding of validity has evolved exponentially since the first descriptions seen in literature in 1915 (Lissitz, 2009). The ongoing reconceptualization of validity began in 1954 when the APA published the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* in which they proposed four types of validity evidence: construct, content, concurrent, and predictive. One year later, Cronbach and Meehl (1955) adopted the term "criterion-oriented validity" to unify concurrent and predictive validity. According to their unification theory, criterion-oriented validity occurs when the investigator is primarily interested in some criterion which they wish to predict, construct validity is involved whenever a test is to be interpreted on a measure that is not operationally defined, and content validity is established by showing that test items are in-line with the universe the investigator is interested in (Cronbach & Meehl, 1955). This unitary approach was reflected in the updated *Standards* (AERA, 1966).

Messick (1989), a proponent of the unitary approach, further redefined construct validity and proposed a new framework wherein validation "embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated" (p.14). In this work, Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other models of assessment" (p.14). Citing numerous scholars, Messick provided guidelines and categories of evidence to be considered in validity inquiry including convergent and discriminant evidence, construct underrepresentation, and construct irrelevance variance. Although considered a critical document and pivotal moment in validity research, Messick's work was steeped in epistemological deliberations which made it difficult for those without the appropriate background knowledge to access and apply the material (Moss et al., 2006). For example, in this text, Messick included a taxonomy of research strategies which delves into five inquiry systems developed by Churchman (1971) and Singer (1959). While explanation of each was provided, further clarification was needed on the practical applications of the inquiry systems as they relate to instrument development and use. In the same year, Messick published a condensed article simplifying the larger points in his original text. In this synthesized version, he defined construct validity "not as a property of the test, or assessment…but rather of the meaning of the test scores" (Messick, 1989, p.741). This revised definition identified a four-pronged classification of the components of validity (test interpretation, test use, evidential bias, and consequential bias) and the six components of validity criteria for all educational and psychological measurement (content, substantive, structural, generalizability, external, and consequential) which formed the basis for addressing validation procedures in educational and psychological testing (Tiffin-Richards & Anand Pant, 2017). The 1999 version of the *Standards* drew on Messick's position which included test properties that can be quantified as well as those that cannot.

Although his views on construct validity were generally in line with Messick's, Kane (2006) argued for a pragmatic approach to validation. He defined validation as "the process of evaluating the plausibility of proposed interpretation and uses" and validity as "the extent to which the evidence supports or refutes the proposed interpretations and uses" (p.17). Kane (2006) posited that any given test score may have multiple legitimate interpretations which can then be used to make decisions about and for the test taker. In such realities, each interpretation requires its own validation for continued use. Kane's proposed framework includes a symbiotic relationship between an interpretative argument and a validity argument. Essentially, the interpretative argument concerns itself with the inferences and assumptions that conclusions are drawn from. The validity argument relies on logical and empirical evidence. Much like Messick's explanation of constructive realism, Kane proposed including test properties that can and cannot be quantified. In Kane's (2006) framework, the validity argument provides an evaluation of the interpretive argument and then an evaluation of each inference, using evidence such as expert judgment, empirical studies, the results of previous research, and value judgments. The framework also included a division of validation into two stages: development (during creation of a test) and appraisal (after a test is developed). The appraisal stage is intended to be much more critical than the development stage (Moss et al., 2006). Along with Cronbach and Messick, Kane's influence on the evolution of assessment validity can be seen throughout the most recent version of the *Standards* (2014).

Sources of Validity Evidence

The current (2014) version of the *Standards* claims assessment validation involves gathering specific evidence to support the proposed uses and interpretations. Below are brief summaries of each of these sources of validity evidence as described in the *Standards*.

- **Evidence Based on Test Content -** content validity allows for distinctive claims about *what* the assessment is designed to measure. The content of an assessment must align to the construct(s) it was created to assess. The process of alignment includes content themes, wording, and formatting of assessment items. The way in which tests are administered and scored play an important role in content-based evidence. Developers must evaluate if constructs assessed are appropriate with the content knowledge of the test-takers. Construct underrepresentation or irrelevance may serve as disadvantages and skew results for test-takers. Without evidence of content validity, inferences on performance cannot be made. However, there is not a designated statistical test to prove content evidence. Content evidence is generally established through a panel of experts who deem whether or not the instrument measures the construct well.

- **Evidence Based on Response Process -** in some assessments, the test-taker's response process to tested items helps establish validity evidence. Assumptions about the cognitive processes test-takers undergo throughout an assessment may support the interpretation of constructs. To do this, developers should question test-takers from various groups and study their test-taking patterns and individual responses. Evidence gathered from the response processes of test-takers can answer questions about differences in interpretation of a specific construct or the entire assessment. Response processes of the test administrators and scorers should also be evaluated when appropriate. It is important to note not all assessments require response process evidence. Like content evidence, there is not a specific statistical test designed to measure response process validity. Regardless, response process can be measured in a multitude of ways (e.g., observations, interviews, feedback, software tracking) and can include rigorous statistical procedures.
- **Evidence Based on Internal Structure -** evidence based on internal structure pertains to "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16). In short, scores on an assessment need to be consistent with the construct(s) assessed. Rios and Wells (2014) further explain there are three aspects of internal structure: dimensionality, measurement invariance, and reliability. Each of these aspects can be statistically evaluated. Results from such analysis helps developers establish internal structure evidence.
    - **Dimensionality -** dimensionality is concerned with determining whether inter-relations among the assessment items support the intended test scores. If so, these test scores can be used draw inferences about the test-takers and the test itself. To

quantify dimensionality, researchers can employ confirmatory factor analysis (CFA) statistical procedures. While there are other methods to determine dimensionality, CFA's are among the most widely used (Rios & Wells, 2014).

- **Measurement equivalence/invariance (ME/I) -** the meaning of invariance is "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, p. 117). As such, ME/I is concerned with determining whether assessment item characteristics, like degree of difficulty, are comparable across identifiable subgroups (e.g., race, sex, age). ME/I testing has gained in popularity since the 1990s (Rutkowski & Svetina, 2014) and despite advocacy for using ME/I to first establish comparability across groups before concluding where differences among groups lie (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; van de Vijver & Leung, 1997), only about 17% of studies citing ME/I focus on cross-cultural comparisons (i.e., invariance testing for two or more culturally different groups; Boer et al., 2018). Comparisons across groups require assurances of measurement comparability (Boer et al. 2018); otherwise, conclusions derived in studies could be considered insufficient, at best.
  - Developers need to ensure assessments are fair and free of biases for all intended test-takers, particularly if the assessment has been translated and is promoted for use in a translated language. Therefore, bias must be extensively considered when assessing ME/I across cultural groups. Bias refers to the systematic errors in measurement that threaten the validity of the study (van de Vijver & Leung, 1997, 2000; van de Vijver & Tanzer,

2004). There are three types of biases distinguished in van de Vijver & Tanzer's (2004) taxonomy: construct bias (the underlying construct measured is not the same across cultures), method bias (which includes three branches: sample bias, instrument bias, and administration bias), and item bias (occurs when an item has a different meaning across cultures). Lack of consideration for fairness and bias in this regard has led to some assessments bring entirely discarded (Rios & Wells, 2014).

    - To test for ME/I across groups, researchers often turn to traditional methods within structural equation modeling (SEM) framework such as exploratory factor analysis, principal component analysis or the most commonly used in literature, multi-group confirmatory factor analysis (MGCFA; Boer et al., 2018; Bollen, 1989; Byrne, 1994; Muthen, 1989). Multidimensional scaling and differential item functioning (DIF) within the item-response theory (IRT) framework may also be used (Boer et al., 2018).

- **Evidence Based on Relations to Other Variables -** the intended interpretation of an assessment means the construct(s) it was built around should be related to other variables. These identified relations require separate analysis of the assessment scores to the external variables. External variables may include measures of criteria the assessment is expected to predict, or relations to other assessments using similar constructs. Researchers should also measure categorical variables (such as those collected in participant demographics) as the results may be relevant in evaluating group differences.

    - **Validity generalization -** developers and researchers alike are often concerned about the generalizability of their work (Martella et al., 2013). A pressing issue in educational testing is the degree to which validity evidence can be generalized to a new situation without further study of the validity evidence in said situation. When an assessment is used to predict similar performance in different times or places, test-criterion correlations may vary. However, through meta-analysis of published literature, these findings may be explained and add further evidence base to an assessment's validity. Generalization of an assessment's results across groups of test-takers is an important part of the validation process. Initial samples of test-takers should be demographically representative of the intended population use (Martella et al., 2013).
- **Evidence for Validity and Consequence of Testing -** the interpretation of assessment results has direct consequences for continued use of the assessment. Unintended results can sometimes have positive consequences and lead to interpretation or use beyond the intentions of the developers. But, unintended results can also signal larger issues with the assessment. Developers must evaluate if identifiable subgroups of test-takers respond differently than predicted.

Reliability

The concept of reliability in education measurement is much more positivist in nature than the validity arguments presented by Messick (1989) which outline some unquantifiable aspects of test development and evaluation. Not all, but some, sources of evidence can and should be objectively evaluated. This does not mean that reliability is a simple process of running data through a computer program and crossing your fingers that the results meet a predetermined minimum standard. Reliability, like validity, is multi-faceted and includes many aspects to consider when developing or appraising an assessment. An assessment with sound reliability supports the validity of the assessment, and conclusions about the intended interpretation of results can then be drawn.

At its crux, reliability is concerned with the accuracy of test scores (McCoach et al., 2013). Nunnally (1967) defined reliability as "the extent to which [measurements] are repeatable and that any random influence which tends to make measurements different from occasion to occasion is a source of measurement error" (p. 206). As Messick (1989) pointed out, it is incorrect to state a test is valid or reliable; rather, the results of a test are valid and/or reliable. An assessment developer intent on establishing validity must also consider the reliability of the assessment. With regard to reliability, assessing the repeatability of test scores lies within the accuracy and error. Accuracy and error are inversely related in reliability measures: the more errors there are in the measurement, the less reliable the scores are, and vice versa (McCoach et al., 2013). Below are brief summaries of five common approaches to testing the reliability of an assessment.

- **Internal Consistency -** internal consistency is at the forefront of reported reliability measures for most assessments (Cortina, 1993; Stanley, 1971). Internal consistency refers to the degree of interrelatedness among test items measuring a specified construct (Cortina, 1993; Green et al., 1977) and should be determined before a test can be employed further (Tavakol & Dennick, 2011). Internal consistency can be measured using coefficient α, also known as Cronbach's alpha (Cronbach, 1951). While there are limitations to the use of Cronbach's alpha (Crocker & Algina, 1987; Sijtsma, 2009), it is arguably the most commonly used statistic (Rios & Wells, 2014). Inter-item correlations, item-total correlations, or split-half reliabilities may also be used (Cortina, 1993). It is important to note, none of these statistical tests alone provide evidence that a test is accurately assessing an intended construct even though it is often reported as such. Cortina (1993) suggests using one of the internal consistency statistics to confirm a single construct structure after factor analytic procedures have been used to determine that the assessment is unidimensional.

- **Homogeneity -** homogeneity refers to the number of items and interrelationships of items to a construct. Homogeneity should be tested using exploratory or confirmatory procedures (Cortina, 1993). It is increasingly common to see both procedures reported in one article as exploratory and confirmatory procedures tend to go hand-in-hand. Evidence of these procedures includes factor eigenvalues, scree plot, (Pituch & Stevens, 2016) or parallel analysis (Vivek et al., 2017), and factor loadings of latent variables to the scores on the test items (Schumacker & Lomax, 2016).
- **Stability –** stability is best known in literature as test-retest reliability. Test-retest reliability refers to the agreement between the results of an assessment administered at two points in time (Phelan & Wren, 2006). This test is not appropriate for measures where change is likely or expected. Pearson's product moment correlation can be used to measure test-retest reliability. Test-retest is a common reliability measurement approach used in special education transition and related assessments (McConnell, 2012).
- **Inter-rater –** inter-rater refers to the extent to which an assessment produces the same results when administered by multiple people (Phelan & Wren, 2006). Having multiple raters is useful in decreasing the influence of the individual bias of each rater especially when there is a human element to the assessment (e.g., observation, human scoring, etc.)

which may contribute to scoring error. Specific inter-rater procedures are often seen in single-case methodology and require explicit instructions and definitions to reduce variability in score reports.

- **Parallel form –** parallel form is obtained by administering different versions of an assessment to the same group of individuals (Phelan & Wren, 2006). The different versions of the assessment must contain items that probe the same construct. The scores from the two versions can be correlated in order to evaluate the consistency of the results across versions. Parallel form reliability is generally evaluated through Cohen's (1988) standards when multiple forms have been developed with the intention of measuring the same constructs (AERA et al., 2014).

Many more types of reliability measures exist which may be more appropriate in establishing reliability of an assessment. Lissitz and Samuelsen (2007) list a series of questions and related potential sources of evidence (statistical tests) which can be used to establish reliability.

Overview of Special Education

Special Education in the United States is currently governed by IDEA (2004). Among the many requirements in the legislation is the directive to select assessment or evaluation materials to assess a child's disability, needs, and/or progress in a manner that is not discriminatory on a racial or cultural basis and the measures of which are valid and reliable (IDEA, 300.304). However, definitions of what constitutes valid and reliable assessments are not provided in the legislation nor are the *Standards* or other similar text mentioned to further elaborate on validity requirements in special education. In fact, the Office of Special Education and rehabilitative Services (OSERS) official transition guide simply lists transition assessments as plural and needed on an annual basis (OSERS, 2017). This omission leaves room for interpretation and error which, for a discipline rooted in litigation, can have significant consequences (Gershwin Mueller, 2015; Parker & Cross, 2020). To better understand the need for appropriate tools in special education, one must also understand the history of special education.

One of the stated purposes of the IDEA (2004) is to "ensure that all children with disabilities have available to them a free appropriate public education [FAPE] that emphasizes special education and related services designed to meet their unique needs and prepare them for further education, employment, and independent living" (300.1). Driven by over 45 years of advocacy and education-related lawsuits (Yell et al., 2019a), the premise of FAPE was formally introduced in 1975 when P.L. 94-142 or the Education for All Handicapped Children Act (EAHCA) was signed into law. As a new iteration of the original Elementary and Secondary Education Act (ESEA, 1965) and the following Education of the Handicapped Act (EHA, 1970), EAHCA laid the critical groundwork for what would become IDEA.

IDEA itself would undergo three amendments (1990, 1997, 2004), each targeting the improvement of special education practices and the in-school and post-school outcomes of students with disabilities. Though each round of amendments served an important purpose in the history and implementation of IDEA and special education practices have certainly improved over time (Madaus & Shaw, 2006), concerns regarding student outcomes in the areas of further education, employment, and independent living have also been documented over decades (Banks, 2014; Blackorby & Wagner, 1996; Mithaug et al., 1985). As a whole, students with disabilities experience significantly lower post-school outcomes in nearly all areas of adult life when compared to their same-aged peers without disabilities (Carter et al., 2012; Flexer et al., 2011; Grigal et al., 2011; Newman et al., 2009, 2011; Prince et al., 2018; Test et al., 2009).

Outcomes of students with disabilities from culturally linguistically diverse backgrounds, non-native English Speakers, and English Language Learners

In the U.S., the number of students from culturally and linguistically diverse (CLD) backgrounds is projected to surpass the number of non-CLD students by the year 2025 (NCES, 2018). The growing diversity includes an increased number of non-native English speakers or English Language (EL) learners (Artiles et al., 2005). Approximately 23% of all public-school students speak languages other than English at home (Zeigler & Camarota, 2018) with 14.3% of these students receiving EL services in school (NCES, 2020). Currently, students from American Indian/Native American, Black/African American, and Hispanic/Latino/a/X backgrounds represent the largest number of students served under IDEA (NCES, 2019). Similarly, 9% of students who receive special education services are dually identified as ELs (NCEO, 2011). Socio-demographically, most EL students with disabilities (ELSWD) come from Spanish-speaking, Hispanic Latino/a/X backgrounds (Aud et al., 2011).

Both groups of students, CLD and EL, experience poor post-school outcomes when compared to their non-CLD (Leake & Black, 2005) and non-EL peers (Trainor et al., 2016). Findings from the National Longitudinal Transition study-2 (NLTS-2) show students with disabilities from minoritized linguistic groups or CLD backgrounds experience significantly less success in reaching post-school education, employment, and independent living goals when compared to their same-aged non-CLD peers with and without disabilities (Wagner et al., 2007). Students with disabilities from CLD backgrounds are at higher risk of experiencing poor in-school performance, receiving lower wages in employment settings, having limited access to post-secondary education, and having limited access and opportunity for living independently (Zhang & Benz, 2006). As a whole, EL students face poor post-school outcomes, and when the added factor of disability is included, they are more likely to experience obstacles in meeting the individuals with disabilities education act (IDEA) benchmarks of post-school success (Trainor, 2016). Studies have found membership in a minoritized linguistic group may also increase the risk of poor post-school outcomes for students with disabilities (Oswald et al., 2002; Skiba et al., 2005).

Overview of Transition

IDEA (2004) addresses the increasingly concerning outcomes of students with disabilities (from all backgrounds) by mandating transition planning and the use of transition assessments to determine appropriate transition goals and services for students with disabilities in the areas of post-school education, employment, and independent living. Students served under IDEA are required to have an annually updated IEP and, when at least 16 years old, a transition plan. The concept of transition and transition planning is not a new one; Halpern's (1992) analogy of pouring old wine into new bottles fits best with the history of transition-focused efforts in the U.S. The federal government has expressed concerns on the vocational education of citizens as far back as the civil war (Dugger, 1965). The career education movement essentially began after U.S. Commissioner of Education, Sidney Marland, declared career education a top priority for the education of students in the country (Brolin, 1983). Part of this movement included the Career Education Implementation Incentive Act of 1976, which specifically mentioned individuals with disabilities and the Council for Exceptional Children's (CEC) formal endorsement of career education (Halpern, 1992). The Office of Special Education and Rehabilitative Services (OSERS) later released a position paper, proposing the Bridges Transition Model, which comprised of three types of transition services deemed necessary to facilitate transition from school to work (Halpern, 1992; Will, 1984). In this model, Will (1984) defined a new federal initiative called *transition*.

Will's model was later improved upon by Halpern (1985) wherein the emphasis of providing appropriate transition services while students are in high school was brought to the forefront. Halpern (1992) later defined transition as "a period of floundering that occurs for at least the first several years after leaving school as adolescents attempt to assume a variety of adult roles in their communities" (p. 203). The concept of students floundering remained the center focus of definitions that followed Halpern. Rowe et al. (2015) provides the most recent and arguably most comprehensive definition of secondary transition: "A transition program prepares students to move from secondary settings to adult life, utilizing comprehensive transition planning and education that creates individualized opportunities, services, and supports to help students achieve their post-school goals in education/training, employment, and independent living" (p. 11).

Legal Implications in Transition Planning

The transition plan in the IEP is subject to scrutiny as it outlines the assessment tools used to identify the student's areas of strengths and needs and specifies post-secondary education, employment, and independent living goals the student will work towards for the calendar year. The progress towards any goal listed in the IEP must be monitored, documented, and adjusted as needed. This process, though logical, has proven to be complex in practice, and when done without fidelity, it can lead to serious legal implications.

Prince et al. (2013) found four major trends in their review of 21 transition-related court cases, each of which connect to the need for appropriate, validated transition assessments. First, schools must provide adequate and individualized services to ensure FAPE. This includes properly assessing a student’s vocational skills and determining individual and immediate needs**.** Second, transition plans in the IEP must enable the student to receive FAPE by “increasing the likelihood of successful transition in post-secondary environments” (p. 289). The premise of FAPE is met when a student’s transition plan explicitly states individualized post-secondary goals which are determined by the interpretation of transition assessment results. Third, transition plans must be a comprehensive representation, determined in part by transition assessment results, of the student’s needs and be designed to ensure the successful transition from school to post-school life. Fourth, transition planning must include “the use of multiple transition assessments, updated transition goals when students’ interests changed, and [have] career goals that are consistent with the student’s academic abilities” (p. 289). One can argue that without the use of appropriate transition assessment tools, the entire transition plan is rendered invalid. Thus, Prince et al. (2013) advocate for the use of multiple (meaning more than one), age-appropriate transition assessments to avoid legal repercussions. They also recommend shifting from the use of informal assessments (assessments without validity evidence) to transition assessments that meaningfully contribute to the development of transition goals and have validity evidence (Prince et al., 2014).

Transition Assessments

Provisions in IDEA (2004) require individual states to submit data on indicators of performance for school-aged children with disabilities. The National Secondary Transition Technical Assistance Center (NSTTAC) and the Office of Special Education Programs (OSEP) developed the Indicator-13 (2007) checklist to gather and evaluate data regarding post-secondary transition planning in the IEP. Six (of eight) components of the checklist rely on the results of the student’s age-appropriate transition assessments (see Figure 1). Without appropriate transition assessments, the transition plan in the IEP cannot meet Indicator 13 requirements, meaning it is not compliant with IDEA requirements, resulting in a violation of FAPE.

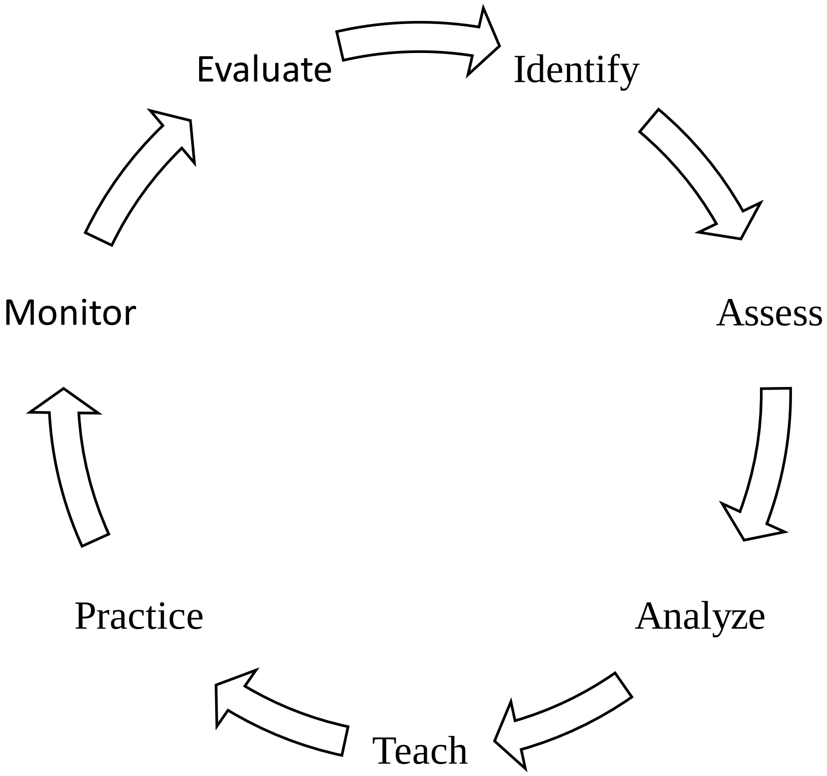
Transition assessments leave special educators with a means to guide students into their desired post-school lives using data (McConnell, 2012). Transition assessments are also needed to provide students with an individualized and explicit plan of action. In their 2013 position paper, the Division for Career Development and Transition (DCDT) define age appropriate transition assessment as:

> an ongoing process of collecting information on the youth's needs, strengths, preferences, and interests as they relate to measurable postsecondary goals and the annual goals that will help facilitate attainment of postsecondary goals. This process includes a careful match between the characteristics of the youth and the requirements of secondary environments and postsecondary environments along with recommendations for accommodations, services, supports, and technology to ensure the match. (Neubert & Leconte, p. 74-75)

In the same paper, the authors stress the importance of viewing transition assessments as a process rather than a mere requirement. The assessment process is cyclical in nature, starting with the selection of appropriate tools, continuing onto the interpretation of assessment results, creation of explicit goals, and identification of services to support the student in meeting their goals, providing needed services, monitoring progress made by the student, and repeating the process when the goals are met or require adjustment (Neubert & Leconte, 2013; see Figure 2 for a visual representation of the assessment process). This process should be student-centered and designed to emphasize the student's abilities (Neubert & Leconte, 2013; Sitlington et al., 1997, 2007; Test et al., 2006).

Figure 2

Transition Assessment Process



Note. The transition assessment cycle: (1) identify transition area of need, (2) assess students, (3) analyze results, create explicit goals and identify services needed to support the student in meeting their goals, (4) educators/stakeholders teach skills to address identified areas of need (determined from the transition assessment), (5) students practice the transition skills throughout the school year, (6) the student's progress is monitored and tracked, and (7) an evaluation of the progress is completed to determine if the goal is met or if the plan needs adjusting. This cycle should be repeated throughout the student's secondary career.

Transition Assessments for ELSWDs

With the growing diversity of students and the known outcomes of ELSWDs, it is obvious issues of in-school and post-school outcomes need to be directly addressed. However, little is actually known on the transition practices of ELSWDs (Trainor, 2016). Identifying and providing access to appropriate transition education and opportunities is a complicated but incredibly important endeavor (Povenmire-Kirk et al., 2010; Trainor, 2016). One way in which to embark on this endeavor is to ensure the creation, access, and validation of culturally appropriate and linguistically inclusive transition assessments.

An ongoing issue with transition assessments is the lack of consideration for students with disabilities from linguistically minoritized backgrounds and ELSWDs. While many transition assessments exist and are in regular use (Martin, 2013), few are available in languages other than English. Of those translated assessments, only two have associated evidence of validity in a language other than English (Mumbard´o-Adam et al., 2018; Yang et al., 2005), but neither study was conducted with students in the U.S. where the assessments are regularly used. The use of translated assessments is a complicated issue that raises many questions. For example, one may ask if the translation of test items is appropriate or if the original and translated versions of an assessment are equally measuring the intended constructs (Pitoniak et al., 2009). An important validation procedure for any assessment is the continual re-evaluation of its appropriateness and validity including in-depth examinations of its translations.

Formal transition assessments in use today were normed with linguistically mainstreamed students (English-language test takers), and while they include students from CLD backgrounds in their study samples, ELSWDs are not explicitly mentioned in any known transition assessment validation procedures. Transition assessments were not designed to support this group of students, and while they are a smaller group within the larger SWD population, their projected growth indicate a large oversight and potential problem for the field. To compensate for the lack of consideration and availability of language-inclusive transition assessments, Greene (2011) advocates for the increase of culturally responsive communication during the transition assessment process (e.g., learning and understanding the family's cultural beliefs about disability and transition, asking for and listening to family perspectives, and supporting family's hopes and aspirations for the child's future – even when they differ from the traditional ideas of transition in special education). While Greene's approach may be useful for daily education practices, it is prudent that action be taken from the research field as well. Transition assessments must be validated in languages other than English to provide accurate transition goals, plans, and services to students with disabilities from linguistically diverse backgrounds and ELSWDs.

Validity of Comprehensive Transition Assessments

To this point, validity and reliability have been discussed in relation to special education transition assessments. As well as the transition assessment considerations for ELSWD. A review of currently used transition assessments is needed to understand the availability of linguistically inclusive options and assess their accompanying validity evidence to determine their appropriateness when testing ELSWDs. While there are far more transition assessments in existence, below are descriptions of comprehensive transition assessments (i.e., assessments testing more than one transition area) that have varying degrees of validity and reliability evidence. A comparison of reviewed assessments can be seen in Table 1.

Table 1

Comparison of Comprehensive Valid and Reliable Transition Assessments

Assessment Name	Reference	Evidence of formality	Valid	Reliable	Translated into language	Valid in translated language	Assessment format	Intended Group	Assesses further education	Assesses employment	Assesses independent living	Assesses other area
Self-Directed Search (SDS) - Revisied (R) (5th edition)	Holland & Messer (2017)	n= 1739. Test-Retest reliability coefficient ranged from .82 to .96. Intercorrelations between scales, equivalence of prior editions, equivalence between print and online versions, convergent validity with other career interest measures.	X	X	Spanish		Paper-pencil and online	11-70 years old; students, parents, counselors, military, and businesses	X	X		
SDS-R (4th edition)	Yang et al. (2005)	n= 1602. Omnibus test found measurement invariance supported for males and females across regions and CFA results indicate subscale scores reflect constructs measured.	X	X	Chinese	X		Males and females in mainland China and Hong Kong.				
SDS - Easy (E) (4th edition)	Plake & Impara (2001)	n= 719. Internal consistency reliability coefficients range from .94 to .96 for the summary scales and .81 to .92 for the section scales.	X	X			Paper-pencil	Students and adults who can read to at least the 4th-grade level		X		

SDS - Career Planners (CP) (4th edition)	Plake & Impara (2001)	n= 101. Internal consistency ranged from .87 to .93 for the summary scores and .69 to .87 for the section scales. The section scale correlations were all greater than .80 and the summary scale correlations were greater than .94.		X			Paper-pencil	Professional-level employees and adults		X		
Transition Planning Inventory (TPI) 2	Kohler (1998)	n= 844 total. Test-retest reliability values across nine areas ranged from .70 to .98 across all tested users. For the total group of examinees, the coefficient alphas within each rater group ranged from .70 to .94. Creators cite content and criterion validity in the technical manual.	X	X	Spanish, Chinese, and Korean		Paper-pencil and online	Secondary-aged students with disabilities	X	X	X	
Transition Behavior Scale (TBS) 3	McCarney & Arthaud (2012)	n= 1,967 students. Internal consistency was .94 for the total score. Test-retest reliability yielded correlation coefficients exceeding .72 for each of the three subscales. Inter-rater reliability was .75 for the total score. Authors cite content and criterion validity. All correlations	X	X			Paper-pencil	Secondary-aged students with and without disabilities		X	X	X

		were significant at the p < .001 level.										
Casey Life Skills	Bressani & Downs (2002)	The CLS has a total internal consistency of .94α and a test-retest administration which ranged from .80 to .91α. Studies show significant differences in assessment results exist for the age of test takers. Older students are likely to score higher on the CLS than younger students.	X	X	Spanish		Online	Foster care and at-risk youth ages 14-21. CLS may be appropriate for all youth as well.	X	X	X	X
Self-Determination Inventory (SDI) - student report (SR)	Shogren et al. (2017)	n= 311 youth. CFA model fit for configural invariance was acceptable: χ2 (34) = 63.861, RMSEA = .075, CFI = .976, TLI = .960, and SRMR = .038. Reliability coefficients ranged from .46 to .87 for all tested subscales.	X	X	Spanish, American Sign Language videos	X	Online	All students with and without disabilities				X
Self-Determination Inventory (SDI) - student report (SR) - Spanish	Mumbard´o-Adam (2018)	n= 620 youth in Spain. Internal consistency ranged from .627 to .830. CFA model fit was acceptable: χ2 = 2.823, RMSEA = .06, CFI = ..986, TLI = .991, and SRMR = .03. Configural invariance was established.	X	X			Online	Spanish students with and without disabilities				X

Transition Assessment and Goal Generator (TAGG)	Hennessey et al. (2018)	n= 349 students, special educators, and family members. A CFA identified an eight factor structure: The fit statistics for the model were acceptable: χ2 = 1,043.08, df = 499; RMSEA = .058, CFI = .92, TLI = .91. CFAs were conducted two more times over subsequent two years with similar acceptable model fit. In all three phases, researchers assessed the internal consistency of the overall scale score and each subscale score for each version of the TAGG. Reliability coefficients ranged from .78 to .99 across all versions.	X	X	Spanish, Chinese, and American Sign Language videos		Online	Students with mild-moderate disabilities who plan to be competitively employed	X	X	

- **Self-Directed Search (SDS; Forms R, E, and CP)** – the SDS is a self-administered, self-scored, and self-interpreted vocational interest inventory developed by Holland et al. (1997). The SDS is used across many age groups and with persons of varied backgrounds and settings. Based on Holland's theory of vocational choice and using the Holland six work personalities (realistic, investigative, artistic, social, enterprising, and conventional), the SDS provides a three-letter code to users with matching occupations. Users can then explore within the identified preferred work environments. As Brown (2001) notes, the SDS is a well-researched and widely used system.
    - Three forms of the SDS are available for use: form R (5th edition), form E (4th edition), and form CP (4th edition). Form R is designed for high school and college students; Form E is designed for students and adults who can read to at least the fourth-grade level; Form CP is an alternate version of form R and intended for use by professional-level employees and adults. Brown (2001) informs that users should complete, score, and interpret results independently. Each form was tested using appropriate samples, and results for the norm group are provided in the assessment instruction manuals. Although there is ample validity evidence associated with the SDS, it is important to note the inventory's lack of predictive validity evidence. Additionally, the Form R is available in Spanish and is "designed specifically for Spanish-speaking individuals living in the U.S." but "the products used English norms only" (PAR, 2021) indicating that intention is not equal to validity.
    - Yang et al. (2006) examined the psychometric properties and measurement invariance of the Chinese-translated SDS-R 4th edition (made in 1994) with

    participants from mainland China and Hong Kong. Findings suggest the translated version demonstrated sound construct validity. However, this study is considered outdated now as the SDS-R 5th edition was released in 2013. The 4th edition has not been in use or continuously studied in over seven years. Additionally, findings particular to participants from mainland China and Hong Kong may not be generalizable to Chinese-testing participants in the U.S.

- **Transition Planning Inventory second edition (TPI-2)** – the TPI-2 is designed to identify a student's readiness to transition into a post-secondary setting. The TPI-2 consists of a student, school, and home version. The technical manual advises that more than one version must be completed per student. Clark and Patton (2006) found students consistently rated themselves higher across all items than did families or school personnel. Questions in the TPI-2 are nestled under three categories: learning, living, and working. These areas are directly related to the three areas included in the transition section of the IEP. The TPI-2 assessment is part of an entire transition curriculum and assessment package.
    - Though the TPI-2 claims to have ample evidence of validity, there is no evidence disconfirming racial or gender bias, and the sample lacks representation from Hispanic/Latino/a/X and Asian students (Massa-Carrol, 2018). Research on the TPI-2 is limited (Kohler, 1998; Rehfeldt et al., 2012), and the qualitative nature of the assessment suggests results are subjective to the grader. Massa-Carrol (2018) suggests a longitudinal evaluation of the impact of TPI-2 is needed to support claims of validity and effectiveness.

    - One study on Spanish-translated materials with the first edition of the TPI found the translation of items to be accurate and produce acceptable inter-item reliability (Stevens, 1983). A similar study has not been conducted with the TPI-2. Additionally, a new version of the assessment has very recently been released, TPI-3 (Patton & Clark, 2021), but norming information is not readily available for review.
- **Transition Behavior Scale third edition (TBS-3)** – the TBS-3 lists validity and reliability information in the front page of the assessment along with the administration instructions. The TBS-3 complete kit consists of a school and self-report version and is comprised of three domains: work-related, interpersonal relations, and social/community expectations. Questions in this assessment are focused on transition-related skills. The technical manual provides background information on concepts of transition and guidance on how to use results to support outcomes of students with disabilities.
    - Other than the information listed on the manual there is a significant lack of evidence supporting the continued use of the TBS-3 in education settings.
- **Ansell-Casey Life Skills Assessment or Casey Life Skills (CLS)** - the CLS is a self-report assessment originally designed for youth ages 16+ in foster care (Ansell et al., 2004). The CLS is one of the few assessments which specifically considered multiple cultural and societal factors of its users in development (Ansell et al., 2004). Even though the CLS was not created for special education populations, it is a widely used transition assessment (Martin, 2013). The CLS website offers 15 different versions for youth from various backgrounds and stages of life. For example, versions of the CLS exist for students who have disabilities, are homeless, American Indian, pregnant, and LGBTQ.

    - The original CLS version for foster-case youth has been studied numerous times and is widely considered an effective tool for at-risk youth. However, the 'Education Supports Assessment' version, which is designed to assess students in the areas of disability issues and support, IEP, 504 plan, and life after high school, does not have any stand-alone validity evidence. Validation of the Education Supports Assessment version is needed to support continued use in educational settings.
- **Self-Determination Inventory (SDI)** – the SDI is not considered a comprehensive assessment but included in this review for its validity findings relevant to cross-cultural ME/I. The SDI is used to solely evaluate the self-determination of individuals with and without disabilities. This assessment was created using the causal agent theory framework for understanding how people engage in self-caused, autonomous action (Mumardo-Adam et al., 2018). Self-determination is a known predictor of positive post-school outcomes (Test et al., 2009), and promoting related skills is recognized as a best practice in special education transition literature (Shogren et al., 2017). The SDI consists of three versions: student report (SDI:SR), parent/teacher report (SDI:PTR), and the adult report (SDI:AR). Presently, only the SDI-SR is validated and included samples of students with and without disabilities.
    - The SDI-SR is the only transition-related assessment with validity evidence for its Spanish-translated version (Mumbard´o-Adam et al., 2018). It is important to note the SDI Spanish validation study was conducted in Spain, where ideas of self-determination may be culturally different than those in America where the SDI was created and originally normed. Additionally, the study procedures show

standard SEM methods and a modified ME/I procedure for determining discriminant validity, which do not quite follow the recommended ME/I procedures for comparability (Boer et al., 2018) but can be deemed sufficient for continued use in Spain. Further studies are needed to determine the translated SDI-R's effectiveness for U.S. Spanish-testing populations and for ELSWDs.

The Transition Assessment and Goal Generator (TAGG)

The Transition Assessment and Goal Generator (TAGG) is designed to assess secondary aged students (grades 9-12) with IEPs who plan to be competitively employed and/or enroll in further education after graduation (Martin et al., 2015). The TAGG was chosen as the transition assessment under investigation for this study because of its availability and familiarity to the researcher (discussed further in chapter 3). The TAGG test items derive from research-identified non-academic student behaviors associated with positive post-school outcomes in further education and employment (McConnell et al., 2012). The TAGG provides users with a norm-referenced graphic profile, present levels of performance statement, lists of strengths and needs, and suggested transition goals which are compliant with IDEA standards (Martin et al., 2015).

The TAGG underwent an iterative development process following the *Standards* (Martin et al., 2015; McConnell et al., 2012; McConnell et al., 2020) and consists of professional (TAGG-P), family (TAGG-F), and student (TAGG-S) versions. It is designed to assess the individual ability of students with disabilities across eight constructs: *strengths and limitations, disability awareness, persistence, interacting with others, goal setting and attainment, employment, student involvement in the IEP, and support community.* These non-academic behaviors/constructs are often overlooked in transition planning but have been found to increase the likelihood of post-secondary success (McConnell et al., 2020).

TAGG Validity Evidence

Research shows the TAGG has ample evidence of validity and reliability (Burnes et al., 2018; El-Kazimi, 2012; Hennessey et al., 2018; Martin, 2013; Martin et al., 2015; McConnell et al., 2015; McConnell et al., 2020). There are some differences between gender and student disability categories (El-Kazimi, 2012; McConnell et al., 2015); however, the assessment manual asserts the TAGG is still appropriate for individuals of all genders with mild-moderate disabilities (Martin et al., 2015). The sources of validity evidence based on the *Standards* are summarized below:

- **Evidence based on test content** – as originally conceptualized by McConnell et al. (2012), the TAGG development team identified behaviors associated with positive post-secondary outcomes from which the test constructs are derived.
- **Evidence based on response process** – the TAGG development team observed 20 test administrations across four U.S. states. Feedback from observations were included in the following iterations of test items.
- **Evidence based on internal structure** – IRT tests confirmed the suitability of the response patterns across all subscales, providing evidence of internal structural validity. Hennessey et al. (2018) confirmed model fit for each construct and across all versions.
    - TAGG-P – the eight-construct professional version has acceptable model fit ($\chi2$ = 1,236.91, *df* = 499, RMSEA = .067, CFI =.969, TLI = .966). Test-retest findings show a large correlation between two administrators (.80).
    - TAGG-F – the eight-construct family version has acceptable model fit ($\chi2$ = 742.81, *df* = 499, RMSEA = .052, CFI = .970, TLI = .967). Test-retest findings show a large correlation between two administrators (.70).

    - TAGG-S - It is important to note the *strengths and limitations* and *support community* constructs collapse onto one another in the student version (Hennessey et al., 2018). The seven-construct student version has acceptable model fit ($\chi2$ = 885.34, *df* = 499, RMSEA = .054, CFI = .897, TLI = .884). Test-retest findings show a large correlation between two administrators (.70).

- **Evidence based on relations to other variables –** discriminant evidence was obtained by comparing TAGG scores to student GPA and percent of time the student spent in general education and no correlation was found (McConnell et al., 2015). McConnell et al. (2015) the impact of gender on TAGG scores and found no differences in the professional and student versions and only slight differences in family scores for females.
    - Predictive validity - refers to the degree to which scores on a given assessment successfully predict performance on a different outcome (Frey, 2018). Although predictive validity is not one of the five main sources of validity evidence included in the *Standards*, it is an important aspect of transition assessments used in special education settings. IDEA (2004) specifies the need to improve post-secondary outcomes for youth with disabilities. One method to do so is through the use of formal transition assessments with predictive validity. There are very few studies using educational assessments that establish predictive validity (Martin et al., 2008). To date, only the TAGG has supporting predictive validity evidence. Through a series of logistic regression analyses, Burnes et al. (2018) found constructs tested in the TAGG predicted post-secondary educational and post-secondary employment outcomes of students with disabilities.

- **Evidence based on consequences of testing –** the results from the TAGG can be used to identify transition skills and needs for individual students. The purpose of the assessment is to identify transition goals that help ensure the student has resources, opportunities to learn skills, and can participate in experiences that are known positive predictors employment and further education outcomes. One can conclude that positive consequential validity exists for the intended test takers (when taken in English).
- **Reliability across TAGG versions** – the total scores across all TAGG versions showed statistically significant medium-sized correlations (McConnell et al., 2015). Hennessey et al. (2018) confirmed internal consistency of the TAGG-P across constructs ranged from .69-.93, with overall consistency of .95; TAGG-F ranged from .60-.93, with overall consistency of .89, and TAGG-S ranged from .44-.82, with overall consistency of .89.

TAGG Language Validation

In addition to the original English version, the TAGG is also available in Spanish, Simplified Chinese, and Traditional Chinese. American Sign Language (ASL) videos and English and Spanish audio options are also available for each item assessed on all three versions. Despite being available in multiple languages, the TAGG has only been validated in English and not tested on Spanish- or Chinese-testing populations. The Spanish-translated version of the TAGG-S was selected for this study over the Chinese-translated version based on the frequency of languages spoken at home (other than English). There are over 41 million Spanish-speakers and fewer than 3.5 million Chinese-speakers in the U.S. (Statista, 2019). Though these numbers do not specify the testing language of ELSWD, they do imply a greater sample may be reached by narrowing the study to Spanish-language TAGG-S users. Preliminary data analysis of current Spanish family and student TAGG users (n = 34) show no significant differences between

Spanish-testing student and family users and English-testing student and family users (Choiseul-Praslin & Sinclair, 2021). However, a larger sample size is needed to determine validity of the Spanish-translated TAGG.

Given the known outcomes of ELSWDs who are primarily from Spanish-language backgrounds and their expected growth in coming years, it is prudent the Spanish-translations of the TAGG be extensively appraised. As it stands, the TAGG is promoted for use with Spanish-language test takers without sufficient evidence that the translations are appropriate or produce comparable results to the normed group (English-language test takers). To simultaneously meet the procedural transition assessment requirements in IDEA, provide substantive and meaningful transition assessment results for ELSWDs, and aid in supporting their post-school outcomes, the Spanish-translated version of the TAGG must be assessed for measurement invariance across groups. Confirming ME/I between English-language and Spanish-language test takers will further promote the use of the TAGG for ELSWDs. In doing so, researchers may find ELSWDs perform similarly or significantly differently than the normed population across the constructs. A significant difference in scores on the TAGG constructs may point to errors in the translation, cultural difference in how the construct is viewed by ELSWDs, or perhaps is an area of underperformance for ELSWDs as determined by literature of known outcomes and indicate where efforts to increase post-school outcomes should be focused. If measurement invariance is not found, then the Spanish-translated version of the TAGG must be formally changed and studied through specified procedures like those listed in the *Guidelines for Translating and Adapting Tests* (ITC, 2019).

Summary of Literature Review

Validity and reliability are crucial components of educational assessments. Assessments require detailed and intricate consideration during the development and appraisal stages. A fair amount of comprehensive special education transition assessments have evidence of validity and reliability, but the field of transition has not yet bridged the gap between researcher and practitioner understanding of validity in transition assessments. This lack of clarity could have dire consequences for students with disabilities, particular for ELSWDs, who are more likely to face undesirable outcomes after leaving high school. With regard to federal law and known post-school outcomes, transition assessment developers and researchers need to address procedural and substantive fairness in testing by validating translated transition assessments. Validating translated transition assessments through cross-cultural ME/I comparability procedures is an active step towards improving the post-school outcomes of ELSWDs.

Chapter 3

Methodology

To an extent, translated assessments address the consequential validity (i.e., the positive or negative social consequences of a particular test) of educational decision making for linguistically diverse students with disabilities or English Language learners/students with disabilities (ELSWDs; Glen, 2020a). However, in reviewing commonly used comprehensive transition assessments, which are available in languages other than English and have varying degrees of validity evidence, it becomes evidently clear that further validity evidence specific to the assessments' measurement equivalence/invariance (ME/I) across cultural groups is needed. Per the individuals with disabilities education act (IDEA, 2004) and Indicator 13 requirements, transition assessment results are used to inform the transition plan in the individualized education program (IEP) and are necessary when determining the individual student's abilities, needs, goals, and services. Currently, there are no known transition assessments with evidence of ME/I establishing comparability across groups of English-language and Spanish-language testing students, validating the assessment as an appropriate measurement tool for Spanish-language testing students. Thereby, it can technically be argued that any transition plan in the IEP citing the use of a translated transition assessment is not appropriate, resulting in a violation of the free appropriate publication education (FAPE) mandate in IDEA and deeming any decisions made based of the assessment results legally meaningless.

When examining the transition assessment and goal generator (TAGG), a special education transition assessment developed by Martin et al. (2015), one will find numerous sources of validity evidence (Burnes et al., 2018; El-Kazimi, 2012; Hennessey et al., 2018; Martin, 2013; Martin et al., 2015; McConnell et al., 2012; McConnell et al., 2015) but none for its translated versions which are available and advertised to users. The TAGG was chosen for this study as it ranks amongst the most popular transition assessments used in the U.S. (Martin, 2013), has over 50,000 completed assessments in its database, has established Spanish translations, and was convenient to the researcher. The Spanish-translated version of the TAGG was selected for this study over the Chinese-translated version for two reasons: first, students from culturally and linguistically diverse (CLD) Spanish-language backgrounds, including ELSWDs, are projected to surpass the non-CLD student population in the U.S. by 2025 (NCES, 2018), and ensuring an appropriate translated transition assessment is available for this group of students may prevent any IDEA-related legal complications. Second, the frequency of languages spoken at home show Spanish is the most common in the U.S. (other than English; Statista, 2019), and while this does not directly relate to the testing language of ELSWDs, it does imply a greater sample can be reached by narrowing the study focus to Spanish-language TAGG users. Finally, the student version of the TAGG (TAGG-S) was the sole version under investigation in this study as the results would inform next steps in research. Meaning that if the TAGG-S is not comparable across English-language and Spanish-language groups, the TAGG-P (professional version) and TAGG-F (family version) results would not require further evaluation as the constructs in the TAGG-S are not measured equivalently. Inversely, if there is evidence of ME/I, then follow-up studies examining the TAGG-P and TAGG-F results can be conducted. Therefore, the purpose of this study is to evaluate the psychometric properties of the Spanish-translated TAGG-S. To that end, I sought to provide evidence of ME/I across groups and confirm reliability of the internal structure.

Research Questions

The following research questions guided this study:

1. Are the constructs measured in the TAGG functioning equivalently across English-language and Spanish-language TAGG-S groups?
    a. Are the measured parameters (factor loading, intercepts, and error variances) equivalent across groups, based on the weak, strong, and strict tests of equivalence?
    b. If measurement invariance is found, are there mean differences in latent factors?
2. Does the internal structure of the overall and subscale scores of the Spanish-translated TAGG-S meet acceptable standards of reliability as determined by Cronbach's alpha?

Research Design

This study is an extension of Hennessey et al. (2018) which confirmed reliability and validity for all versions of the TAGG. While additional studies have added validity evidence to the TAGG, "there is a possibility that results using a test with a different population from that used in the original scale development will confound true construct differences with measurement differences" (Flora & Flake, 2017, p. 83). This marks an urgent need to evaluate the psychometric properties of the Spanish-translated TAGG-S. Studies seeking to add validity evidence by characteristics of the group (e.g., by disability category, age, gender, etc.) can be completed after first establishing the assessment is an appropriate tool (Flora & Flake, 2017). Though establishing validity evidence can be done in multiple ways, ME/I procedures were chosen as the primary method in this study to ensure the translated TAGG-S is an appropriate assessment for the Spanish-language TAGG-S user group. ME/I is particularly important to cross-cultural research, and seminal work in this field has demonstrated the extent to which measuring instruments and the constructs they intend to measure can vary across culture and even subgroups within the same culture (Hambleton et al., 2005; Tanzer; 1995; van de Vijver &

Leung, 1997). Researchers strongly advocate for conducting ME/I tests prior to reporting cross-group comparisons (Byrne, 1994; van de Vijver & Leung, 1997). First and foremost, comparisons across groups require assurances of measurement comparability (Boer et al., 2018; Byrne, 1998). Any conclusions made about the groups without meeting the first requirement are questionable and likely invalid. As Boer et al. (2018) point out, there is a systematic issue in the ME/I field as overgeneralizations of groups are concluded based off studies which did not first adequately address the assessment's measurement properties.

Correlational Research

This study will investigate the ME/I of the Spanish-translated TAGG-S through comparisons of English-language and Spanish-language TAGG-S groups. As group membership cannot be randomly assigned to the TAGG-S user, this study falls under the category of correlational research. Correlational research is non-experimental in nature and allows the researcher to measure and determine the extent to which factors under investigation covary (Price et al., 2017). Martella et al. (2013) identify three main attributes to correlational research: hypothesis, grouping, and data. Below, I address how this study meets the correlational research attributes.

- **Development of a hypothesis -** hypotheses in correlational research should be grounded on a theoretical framework and previous research. Prior to beginning the study, I hypothesized that: (a) the Spanish-translated TAGG-S sample would fit the known 8-factor structure, but the *strengths and limitations* and *support community* constructs would not collapse onto one another as they do in the original TAGG studies; (b) if differences existed between TAGG-S users, they would likely occur when strict measurement parameters were placed on the model, indicating partial equivalence and overall signifying ME/I for the Spanish-translated TAGG-S; and (c) the internal structure of the Spanish-translated TAGG-S would meet acceptable standards of reliability. This hypothesis was informed by years of TAGG-related research (see: Burnes et al., 2018; El-Kazimi, 2012; Hennessey et al., 2018; Martin, 2013; Martin et al., 2015), research on validity of other transition assessments (Bressani & Downs, 2002; Holland et al. 1997; Holland & Messer, 2017; Kohler, 1998; McCarney & Arthaud, 2012; Mumbard´o-Adam, 2018; Shogren et al., 2017) and research on validity of translated assessments outside of transition assessments (Daradkeh & Khader, 2008; Larsson et al., 2007; Todd et al., 2020; Yoo et al., 2005).

- **Selection of a homogenous group -** group membership requires an operational definition of membership. For practical reasons, methodological research in ME/I is often limited to two groups where one group serves as the reference population and the other is the focal population (Holland & Thayer, 1988; Rutkowski & Svetina, 2014). Participants for this study are defined in detail in the section below but can be generalized into two homogenous groups: the reference group (English-language TAGG-S users) and the focal group (Spanish-language TAGG-S users).
- **Collection and analysis of data -** a wide range of measurement tools can be used in correlational research including standardized assessments with evidence of validity and reliability. The TAGG will be the main source of data collection for this study and is a described further in the following section.

Instrumentation

The TAGG is a special education transition assessment designed to assess secondary aged students (grades 9-12, ages 14+) with IEPs who plan to be competitively employed and/or enroll in further education after graduation (Martin et al., 2015). There are 34 test-items in the TAGG which derive from research-identified non-academic student behaviors associated with positive post-school outcomes in further education and employment (McConnell et al., 2015). The TAGG was developed following the *Standards* (Martin et al., 2015; McConnell et al., 2012; McConnell et al., 2020) and consists of a professional (TAGG-P), family (TAGG-F), and student (TAGG-S) version. It is designed to assess the individual ability of students with disabilities across eight constructs: *strengths and limitations, disability awareness, persistence, interacting with others, goal setting and attainment, employment, student involvement in the IEP, and support community.* Research shows the TAGG has ample evidence of validity and reliability (Burnes et al., 2018; El-Kazimi, 2012; Hennessey et al., 2018; Martin, 2013; Martin et al., 2015).

Data Sources

Data in this study came from the TAGG database and were analyzed extant (see Appendix A for IRB approval of extant data use). Noting the lack of awareness and use of the Spanish-translated TAGG for students and family users, the Zarrow Center at the University of Oklahoma, where the TAGG is hosted, increased organizational recruitment efforts prior to the beginning of this study. The Zarrow Center conducts numerous trainings on the TAGG and presents at national and international conferences throughout the academic year, and they offered TAGG credits to special educators needing a transition assessment for Spanish-language users. They also shared the recruitment information with special educators via their TAGG and Zarrow Center social media accounts and email listservs. This recruitment was conducted outside the scope of this study. Though the impact of this recruitment strategy is unknown, any Spanish-completed TAGG-S results collected during the Zarrow Center limited-time offering were likely included in data analysis provided they met the inclusionary criteria listed below. Data was de-identified at the time of analysis; any identifying information gathered in the TAGG website at the time of assessment (e.g., name of user, location, etc.) remain confidential and cannot be traced back to any specific user.

Participants

Participant data were categorized in two groups: (a) the reference group - English-language TAGG-S and (b) the focal group - Spanish-language TAGG-S. To answer the research questions, a sufficiently large and equivalent sample size was needed for both groups. Though there is no singular minimum sample size recommended (Meade, 2005), structural equation modeling (SEM) studies typically require large samples to produce accurate results. However, there is disagreement in the literature as to what constitutes a large sample. For SEM-related procedures the general consensus appears to be greater than 200 (Field, 2009; Kline, 2010, 2016; Kyriazos, 2018; Schumacker & Lomax, 2015; Tabachnick & Fidell, 2013). However, the number of participants included in tests of ME/I are known to affect the power of tests (Putnick & Bornstein, 2016). Therefore, a sample size of at least 250-300 participants per group was determined appropriate for this study. This requirement superseded the minimum sample size needed for Cronbach's alpha (*n*= 30-100; Adam Bujang et al., 2018) test.

To select and analyze TAGG-S data for the reference and focal groups, I had to first download the entire TAGG data set (over 50,000 cases) and then apply filters for each group, beginning with the focal group. To select focal group samples, I sorted the data by TAGG-S responses and then by language the assessment was taken (selecting Spanish responses only). In doing so, I identified 334 Spanish-language TAGG-S responses and removed responses with missing data which left 315 samples for the focal group. For the reference group, I filtered the TAGG-S responses by language (selecting English responses only) and then matched each reference group sample to each individual focal group sample at three levels: gender, age, and disability category. Though not required for ME/I testing, matching the responses aided in ensuring the samples in the study were comparable. Below I have explicitly defined the intended focal and reference groups for this study:

- **Focal group:** students with mild/moderate disabilities who plan to be competitively employed and who complete the TAGG in Spanish.
    - Data analysis excluded completed TAGG-S scores that did not meet the language and disability level requirements. This is based off the TAGG's user manual which states the assessment is most appropriate for students with mild/moderate disabilities, ages 14-22, receiving special education services for academic support (Martin et al., 2015). The educator who initiates the student assessment indicates if the student meets these qualifications prior to administering the assessment, and data is recorded on the TAGG database.
        - Mild/moderate disability – the TAGG collects data related to student disability category and level (mild/moderate vs. severe/profound).
        - Ages 14-22 – encompasses ages of students who are likely receiving transition services as part of their special education services mandated by IDEA (Suk et al., 2020).
    - Competitively employed is defined as: work "performed on a full-time or part-time basis … and for which an individual is compensated at a rate that … meets local minimum wage" (Workforce Innovation and Opportunities Act [WIOA], 2014).

- Employment aspirations of the student are not automatically filtered by the demographic section of the TAGG assessment. I was not be able to control for users who take the TAGG and do not intend to meet this criteria. It is up to the discretion of the special educator who creates an assessment for the student to determine if the TAGG is an appropriate tool.

- **Reference group:** students with mild/moderate disabilities who plan to be competitively employed and who complete the TAGG in English.
  - Age, disability category and level, and employment considerations are the same as the target group.

Participant Demographics

After removing missing data and matching samples, a total of 630 samples were used in this study (315 for each group). Due to TAGG privacy guidelines, the only participant demographics available for review were: language test taken in, gender, age, and disability category. All data were analyzed after Zarrow Center recruitment efforts ended on February 15, 2021, and I could not assess where in the U.S. the test was taken. The majority of participants were female (*n* = 166, 52.7%), 16 years old (*n* = 87, 27.62%), and had a primary disability of Speech or Language Impairment (*n* =135, 42.86%). The differences in gender were slight (*n* = 17), and most respondents were between 15 and 18 years old (*n* =269). There was a significant skew in the data for primary disability category as over one-third of all samples came from students with a speech or language impairment.

Table 2

Participant/Sample Demographics

	Individual Groups		Combined	
	n	%	*n*	%
Gender				
Male	149	47.3	298	47.3
Female	166	52.7	332	52.7
Age				
13	3	0.95	6	0.95
14	18	5.71	36	5.71
15	66	20.95	132	20.95
16	87	27.62	174	27.62
17	70	22.22	140	22.22
18	46	14.6	92	14.6
19	16	5.08	32	5.08
20	8	2.54	16	2.54
21	1	0.32	2	0.32
Primary Disability				
Autism	16	5.08	32	5.08
Deaf-Blindness	0	0	0	0
Emotional Disturbance	26	8.25	52	8.25
Hearing Impairment (including	5	1.59	10	1.59

deafness)				
Intellectual Disability	21	6.67	42	6.67
Multiple Disabilities	21	6.67	42	6.67
Orthopedic Impairment	2	0.63	4	0.63
Other Health Impairment	5	1.59	10	1.59
Specific Learning Disability	55	17.46	110	17.46
Speech or Language Impairment	135	42.86	270	42.86
Traumatic Brain Injury	21	6.67	42	6.67
Visual Impairment (including blindness)	8	2.54	16	2.54
Other	0	0	0	0

Note. *N* = 315 for each individual group (reference and focal) and *N* = 630 for both groups combined.

Data Analysis

ME/I can be tested within an SEM framework or an item-response theory framework (Putnick & Bornstein, 2016). While some researchers are working to integrate the two frameworks (i.e., Raju et al., 2002; Widaman, 2014), the SEM approach is more widely used (Putnick & Bornstein, 2016). Because SEM remains the primary analytic strategy capable of testing assumed equivalences (Byrne & van de Vijver, 2017), I chose to conduct this study within the SEM framework. To answer the research questions, ME/I procedures were completed systematically and as instructed in reference literature (e.g., Byrne, 2010, 2016). The purpose of ME/I testing for this study was to establish evidence of comparability across the reference (English-language TAGG-S) and focal (Spanish-language TAGG-S) groups.

Widaman and Reise (1997) specified a four-step model for testing ME/I: configural, weak factorial, strong factorial, and strict factorial. Aspects of this model have undergone changes in years since (i.e., Milfont & Fischer, 2010; Putnik & Bornstein, 2016; van de Schoot et al., 2015; Vadenberg, 2002), but the four steps remain much intact. Byrne (2010, 2016) asserts the ME/I testing strategy must begin with a confirmatory factor analysis (CFA) for the target and focal group prior to proceeding with the four-step model. Therefore, I began ME/I testing with a series of CFA tests.

Initial CFA Procedures

I first attempted the CFA tests using the Statistical Package for the Social Science (SPSS) software and the AMOS SEM extension but received errors in the output as the TAGG-S uses a three-point Likert scale in measurement. SPSS would not recognize responses for each TAGG item as interval scaled measurements and instead treated the items as categorical measurements. This prompted me to change software for further analyses. I chose to use R-Studio with the Lavaan package as it is more equipped to run categorical CFAs.

For both groups, I began with a first-order CFA using maximum likelihood (ML). The analysis showed some variances were negative which further affirmed that a categorical CFA was needed. Next, I specified the variables in the model as categorical using the ordered function on R-Studio. The ordered function in Lavaan automatically switches from ML to the weighted least squares means and variance adjusted (WLSMV) estimator and specifically uses diagonally weighted least squares (DWLS) to estimate model parameters which is fitting as it is specifically designed for ordinal/categorical data (Li, 2016). This approach still uses the full weighted matrix to compute robust standard errors and a mean- and variance-adjusted test statistics (Rosseel, 2012) and was deemed appropriate and comparable to the Hennessey et al. study which used a similar method for validation. Since variables in the data are treated as categorical, assessing model fit for a CFA using a DWLS estimation occurs differently than with a CFA using ML estimation. Although it is generally agreed that model fit cannot be judged exactly the same, there is not a standardized method for result interpretation either (Beauducel & Yorck Herzberg, 2006; Li, 2016; Newsom, 2018; Savalei, 2020; Xia & Yang, 2019). Savalei (2020) propose two solutions (i.e., modified equations for the chi-square fit index) for adjusting parameter estimates but they have not yet been replicated to support its sole use. Newsom (2018) recommends using values greater than or equal to 0.95 for Tucker-Lewis index (TLI) and comparative fit index (CFI) and values less than or equal to 0.05 for root mean square error of approximation (RMSEA). The TLI is an incremental fit index also known as the non-normed fit index (NNFI) which offers the key advantage of not being significantly affected by sample size (Cangur & Ercan, 2015). The CFI is also an incremental fit index that is directly based on the non-centrality measure (Kenny, 2020) and tests for the extent to which the tested model is superior to the alternative model which is established when the manifest covariance matrix is evaluated (Chen, 2007). The TLI and CFI indices are correlated (Kenny, 2020) as both compare the fit of a hypothesized model with that of a baseline model (Xia & Yang, 2019). On the contrary, the RMSEA is an absolute fit index that assesses how far a hypothesized model is from a perfect model (Xia & Yang, 2019). Shi and Maudeu-Olivares (2020) suggest using only the standardized root mean square residual (SRMR) with the cut off values aligned with that of ML models (values less than 0.08) since SRMR is robust enough to encompass DWLS models. The SRMR is an absolute measure of fit (Kenny, 2020) and is an index of the average of standardized residuals between the observed and the hypothesized covariance matrices (Chen, 2007). SRMR has also shown to provide more accurate confidence intervals and test of close fit than RMSEA alone (Maydeu-Olivares et al., 2018; Shi et al., 2020).

As recommendations for assessing model fit of categorically treated data is relatively new and not yet widely tested, I chose to use a combination of Newsom and Shi and Maudeu-Olivares recommendations to assess model fit for both groups (RMSEA, SRMR, CFI, and TLI), evaluating them separately and then jointly to determine overall model fit. Because the methods for determining fit of categorical CFAs are largely untested, I chose to implement a strict threshold approach to assessing model fit using a dichotomous 'good' or 'bad' terminology to denote whether an index met the threshold. This strict approach included guidelines for determining overall model fit using the terms 'close', 'acceptable', and 'poor'. Use of this strict approach allowed for me to definitively assess model fit and is further defined in chapter 4. Thus, using the combined recommended indices and the strict threshold approach, I found poor model fit for the target and focal groups. This meant I could not proceed with further ME/I testing. Instead, I continued the study with a series of exploratory factor analysis (EFA) tests to identify and assess alternative factor structures.

EFA Procedures

EFAs are used to discover the factor structure of a measure and to examine its internal reliability (Newsom, 2005). This is a recommended strategy when there is no hypothesis on the underlying factor structure of the measurement tool (Anglim, 2014) and suggested method if ME/I model fit is inadequate (Byrne, 1994; Byrne 2016). In this study, the EFA was used to identify a new factor structure for the Spanish-translated TAGG-S as the hypothesized one did not hold. In exploring new factor structures, I used three different software programs to conduct EFAs on the focal group data: SPSS, R-Studio, and FACTOR. Each program exhibits unique qualities for assessing exploratory model fit, and results across each were compared. I also used oblique rotations in each of the programs as the variables were assumed to correlate (based on previous TAGG studies).

- **SPSS** - the factor procedure ignores the measurement scale of variable items and treats the variables as if they are on an interval scale (IBM, 2020). However, categorical EFAs in SPSS may be vulnerable to errors as the item difficulty can be overestimated when the measurement scale is ignored and also small (Gorusch, 1983). Promax rotations were used for all SPSS analyses.
- **R-Studio** - data that are ordinal/categorical can be analyzed using a polychoric correlation matrix in R-Studio as opposed to the standard Pearson's correlation matrix used in SPSS. Polychoric correlation measures agreement across samples between ordinal/categorical variables (Glen, 2020b; Mangal, 2010). The Pysch and GPA Rotation packages in R-Studio were used to run this analyses. Oblimin rotations with ML were used for all R-Studio analyses.
- **FACTOR** - data that are analyzed using polychoric correlation matrices can also be analyzed using FACTOR. This program is capable of running EFAs using parallel analysis, a method of determining the number of factors to retain from factor analysis (Lorenzo-Seva & Ferrando, 2006). DWLS and promax rotations were used for all FACTOR analyses.

In each program, I first conducted a general EFA (unrotated) to assess how many factors the data were grouping under. Across the three programs, parallel analysis and scree plot examinations suggested possible fit to the data in six-, four-, and three-factor models. I then forced the six-, four-, and three-factor models in each program with oblique rotations and assessed model fit using conventional EFA standards. EFA results across the three programs were compared. The R-Studio and FACTOR programs produce fit indices for which a less stringent approach was used to determine model fit. The less stringent approach was deemed appropriate as the EFA tests are exploratory in nature and a looser interpretation of model fit can be considered. The less stringent approach is further defined in chapter 4. Finding the promising results from the EFAs, post hoc CFAs were conducted for the six- and four-factor models.

Post Hoc CFA Procedures

Using R-Studio with the Lavaan package, I conducted a series of post hoc categorical CFAs for the Spanish-translated TAGG-S in the same manner as the initial categorical CFA procedures (i.e., ordered function, DWLS estimator). Based on the EFA comparison across programs, I analyzed the six- and four-factor models from the R-Studio and FACTOR programs and the three-factor model from the FACTOR program. I then used the combination of Newsom (2018) and Shi and Maudeu-Olivares (2020) recommendations for assessing model fit (RMSEA, SRMR CFI, and TLI) with the strict threshold approach to determine if the tested models met acceptable fit to the data.

Chapter 4

Results

This study sought to evaluate the psychometric properties of the Spanish-translated Transition Assessment and Goal Generator – Student version (TAGG-S). Specifically, this study's original aim was to provide measurement equivalence/invariance (ME/I) validity evidence across English-language and Spanish-language TAGG-S groups (called reference and focal groups, respectively) and confirm reliability of the Spanish-translated TAGG-S internal structure.

Initial Confirmatory Factor Analysis (CFA) Results

The study procedures departed from the original purpose within the initial step of ME/I testing. The identified eight-factor structure of the TAGG-S was replicated for all initial CFAs to ensure the model structure would hold across the focal and reference groups. First, a maximum likelihood (ML) CFA was conducted on R-Studio using the Lavaan package (see Appendix B for all syntax used for this study in R-Studio). Results showed negative variances indicating errors within the data for both groups. The TAGG-S responses for both groups are on a three-point Likert scale, meaning the data were treated as categorical instead of interval/scaled. This meant a categorical CFA using the diagonally weighted least squares (DWLS) estimator was most appropriate for the first step in ME/I testing. The categorical CFA was an appropriate deviation and in alignment with previous TAGG validation studies (Hennessey et al., 2018).

A strict threshold approach was used to determine individual and overall model fit using the Newsom (2018) and Shi and Maudeu-Olivares (2020) recommended fit indices for categorical data: root mean square error of approximation (RMSEA), Tucker-Lewis index (TLI), comparative fit index (CFI), and the standardized root mean square residual (SRMR). In this strict approach, only RMSEA values equal to or less than 0.05, TLI values equal to or greater than 0.95, CFI values equal to or greater than 0.95, and SRMR values less than 0.08 were considered 'good' and deemed appropriate to move forward. All other values were reported as 'bad' or unacceptable for continued ME/I testing. In this study, overall model fit is denoted by the terms 'close', 'acceptable', and 'poor'. A model with three or more 'good' values were deemed to have an overall 'close' fit to the data, models with two 'good' values were deemed 'acceptable' fit to the data, and models with only one 'good' value were deemed 'poor' fit to the data. Using the recommended fit indices and researcher-created thresholds for assessing index and overall model fit, it was determined both group models exhibited poor overall fit to the data when conducting a categorical CFA. The self-imposed strict thresholds additionally showed the focal and reference group data were not appropriate for continued ME/I testing (see Table 3). The DWLS CFA model also produced errors within the intercepts (all recorded as 0) which indicate further issues with the model fit.

Table 3

Initial CFA Model Fit Results

Fit Index	Reference Group		Focal Group	
	ML Estimator	DWLS Estimator	ML Estimator	DWLS Estimator
RMSEA	0.055 (bad)	0.065 (bad)	0.061 (bad)	0.177 (good)
TLI	0.74 (bad)	0.875 (bad)	0.805 (bad)	0.856 (bad)
CFI	0.769	0.889	0.827	0.941

	(bad)	(bad)	(bad)	(bad)
SRMR	0.079 (good)	0.100 (bad)	0.089 (bad)	0.109 (bad)
Overall Model Fit	Poor	Poor	Poor	Poor

Note. ML CFA fit indices are included as comparison to the DWLS CFA fit indices. The researcher-created strict thresholds for determining index and model fit were applied to all initial CFA tests.

Exploratory Factor Analysis (EFA) Results

As the ME/I testing could not be continued, alternative factor structures were explored for the Spanish-translated TAGG-S through EFAs across three different statistical software programs: SPSS, R-Studio, and FACTOR. Though the programs were individually capable of producing EFA results for categorical data, they each possessed the ability to include additional parameters relevant to the study that were not solely found on one single program. R-Studio and FACTOR also report fit indices for EFA tests, excluding SRMR. Unlike the initial CFA tests, a less stringent approach for determining fit was used since the new goal for the study was to identify possible alternative structures. In the less-stringent approach to determining index fit, RMSEA values equal to or less than 0.08, TLI values equal to or greater than 0.90, and CFI values equal to or greater than 0.90 were deemed ‘good’. All other values were reported as ‘bad’. The terms used for overall model fit (close, acceptable, and poor) remained the same. An EFA model with three ‘good’ values were deemed to have an overall ‘close’ fit to the data, models with two ‘close’ values were deemed to have ‘acceptable’ fit to the data, and models with only one ‘good’ value were deemed to have ‘poor’ fit to the data.

In all three programs, Bartlett's test for sphericity indicated correlation adequacy and the Kaiser-Meyer-Olkin (KMO) tests indicated sampling adequacy (see Table 4). These tests signal the correlation matrix is significantly different from the identity matrix, meaning there are relationships among the variables and the data is factorable (Buchanan, 2020). The proposed alternative factor structures for the Spanish-translated TAGG-S were replicated across the programs and results were compared to determine which factor structure may be a better fit to the data than the known eight-factor TAGG-S structure. The English-language TAGG-S was not included in this analysis as the data may not have been representative of the entire English-language TAGG-S database and would require further CFA testing before proceeding with alternative factor structure exploration.

Table 4

Bartlett's test and KMO STS Results Across Three Programs

	SPSS	R-Studio	FACTOR
Bartlett's Test for Sphericity (sig.)	0.00	0.00	0.00
Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO MSA)	0.835	0.84	0.91

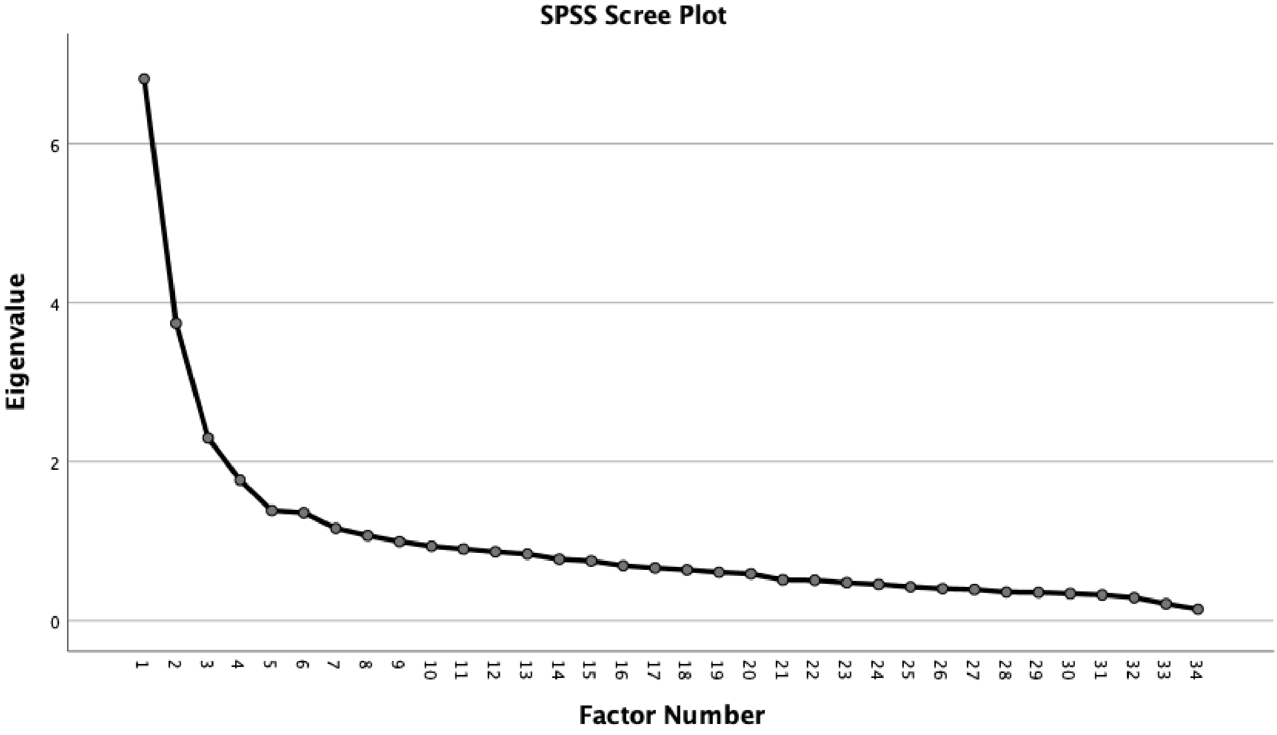
Note. Values < .05 for Bartlett's test for sphericity and values greater than .50 and close to 1.0 for KMO indicate favorable data for factor analysis.

EFA Results in SPSS

A series of EFAs were used to analyze the underlying factors in the Spanish-translated TAGG-S (hereafter referred to as STS data) using the SPSS program. The first EFA did not include any forced rotations so as to openly explore possible factors. The unrotated correlation matrix shows multiple correlations above .30 and the determinant is above Field's (2000) threshold (3.858E-6) which indicate the data is not likely subjected to multicollinearity. The default parallel analysis and scree plot (Figure 3) examination suggested six to eight possible factors; a more stringent parallel analysis (Patil et al., 2017) suggested a four-factor model may be more appropriate; and a three-factor model was also suggested based on the largest eigenvalues and percent of variance explained for the unrotated EFA. Rotated six-, four-, and three-factor models were evaluated further based on the initial unrotated EFA results.

Figure 3

SPSS: Scree Plot



Six-Factor Model in SPSS. Principal axis factoring with a promax rotation was used to identify a possible six-factor model to the STS data. After testing all 34 questions, 21 items were retained given a factor loading criteria of values greater than .40 (Pituch & Stevens, 2016). Table 5 shows items retained for the six-factor model in SPSS with factor loadings for each STS item.

Table 5

SPSS: Six-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1	-0.12	0.248	-0.018	0.31	0.223	0.116
2	-0.17	0.275	-0.053	0.28	0.245	0.003
3	-0.08	0.239	0.014	**0.492**	0.081	-0.014
4	-0.064	-0.005	-0.087	0.367	0.137	0.129
5	0.002	0.022	-0.029	**0.526**	0.145	-0.058
6	0.028	0.082	-0.02	**0.724**	-0.094	-0.024
7	0.038	-0.046	0.054	**0.488**	0.027	-0.02
8	-0.008	-0.058	0.167	**0.451**	0.017	-0.133
9	-0.045	-0.093	-0.01	0.043	**0.594**	0.035
10	0.003	-0.019	0.018	0.121	**0.613**	-0.103
11	0.399	-0.015	-0.032	0.101	0.338	-0.04
12	0.38	-0.112	-0.056	0.137	0.393	-0.049
13	0.34	-0.053	-0.005	-0.088	0.505	0.03
14	-0.06	0.162	0.08	-0.054	**0.342**	0.145
15	0.322	-0.093	0.129	0.293	-0.087	0.001
16	0.021	0.256	0.123	-0.128	0.378	0.181
17	0.382	0.113	0.003	-0.014	0.289	0.018
18	**0.728**	0.1	0.047	-0.075	-0.005	0.008
19	**0.771**	0.047	-0.037	-0.017	-0.051	0.035
20	**0.768**	0.076	-0.072	0.043	-0.024	-0.067

21	**0.706**	0.078	-0.003	-0.085	0.018	0.000
22	0.044	0.027	-0.206	-0.106	-0.015	0.114
23	-0.063	0.03	-0.059	0.092	-0.018	**0.801**
24	0.035	0.036	-0.019	-0.123	0.043	**0.959**
25	0.055	-0.186	0.18	0.219	0.059	0.156
26	0.122	-0.253	0.101	0.254	-0.037	0.226
27	-0.148	0.054	**0.8**	0.054	-0.031	0.097
28	-0.12	0.201	**0.78**	-0.062	0.042	-0.07
29	0.249	-0.104	**0.6**	-0.04	0.044	-0.056
30	0.084	-0.107	**0.698**	0.033	-0.003	-0.017
31	0.012	**0.734**	0.074	-0.069	0.103	-0.144
32	0.121	**0.705**	-0.063	0.008	-0.097	0.065
33	0.205	**0.702**	-0.031	0.059	-0.185	0.066
34	0.308	0.126	0.146	0.241	-0.27	0.081

Note. Items retained per factor are in bold font.

Four-Factor Model in SPSS. Principal axis factoring with a promax rotation was used to identify a possible four-factor model to the STS data. After testing all 34 questions, 28 items were retained given a factor loading criteria of values greater than .40. Table 6 shows items retained for the four-factor model in SPSS with factor loadings for each STS item.

Table 6

SPSS: Four-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3	Factor 4
1	-0.101	0.357	**0.446**	0.003

2	-0.139	0.285	**0.456**	-0.054
3	-0.077	0.187	**0.495**	0.106
4	-0.066	0.113	**0.407**	-0.017
5	-0.001	-0.049	**0.576**	0.071
6	-0.003	-0.022	**0.533**	0.173
7	0.02	-0.093	**0.434**	0.174
8	-0.025	-0.202	**0.413**	0.271
9	0.032	0.091	**0.457**	-0.098
10	0.086	0.028	**0.558**	-0.064
11	**0.437**	-0.002	0.325	-0.045
12	**0.419**	-0.087	0.394	-0.066
13	**0.401**	0.094	0.264	-0.087
14	-0.007	0.373	0.192	-0.001
15	0.291	-0.132	0.153	0.242
16	0.083	**0.5**	0.151	0.019
17	**0.426**	0.177	0.174	-0.032
18	**0.738**	0.081	-0.127	0.076
19	**0.763**	0.032	-0.11	0.023
20	**0.763**	-0.037	-0.02	-0.009
21	**0.715**	0.054	-0.11	0.019
22	0.044	0.122	-0.113	-0.221
23	-0.1	**0.689**	-0.007	0.007
24	-0.012	**0.807**	-0.126	-0.004

25	0.042	-0.016	0.186	0.24
26	0.089	-0.043	0.13	0.197
27	-0.152	0.178	-0.014	**0.806**
28	-0.092	0.174	-0.015	**0.702**
29	0.254	-0.112	-0.036	**0.592**
30	0.078	-0.085	-0.007	**0.71**
31	0.052	**0.517**	0.09	0.002
32	0.121	**0.631**	-0.03	-0.055
33	0.186	**0.6**	-0.055	0.009
34	0.259	0.1	-0.027	0.27

Note. Items retained per factor are in bold font.

Three-Factor Model in SPSS. Principal axis factoring with a promax rotation was used to identify a possible three-factor model to the STS data. After testing all 34 questions, 28 items were retained given a factor loading criteria of values greater than .40. Table 7 shows items retained for the three-factor model in SPSS with factor loadings for each STS item.

Table 7

SPSS: Three-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3
1	**0.644**	-0.004	0.042
2	**0.588**	-0.022	-0.009
3	**0.496**	0.055	0.169
4	0.377	0.059	0.034
5	0.308	0.198	0.161

6	0.3	0.17	0.257
7	0.168	0.169	0.25
8	0.046	0.127	0.353
9	0.382	0.184	-0.04
10	0.374	0.277	0.015
11	0.16	**0.556**	-0.004
12	0.125	**0.575**	-0.008
13	0.225	**0.489**	-0.066
14	**0.49**	0.000	-0.002
15	-0.086	0.345	0.284
16	**0.574**	0.054	0.002
17	0.242	**0.465**	-0.029
18	-0.086	**0.664**	0.049
19	-0.123	**0.709**	0.001
20	-0.131	**0.763**	-0.017
21	-0.096	**0.659**	-0.005
22	0.062	0.006	-0.258
23	**0.666**	-0.209	-0.041
24	**0.681**	-0.175	-0.074
25	0.076	0.088	0.281
26	0.012	0.124	0.228
27	0.112	-0.254	**0.819**
28	0.11	-0.187	**0.714**

29	-0.213	0.205	**0.622**
30	-0.161	0.023	**0.752**
31	**0.557**	0	-0.025
32	**0.581**	0.015	-0.107
33	**0.52**	0.069	-0.042
34	0.027	0.208	0.269

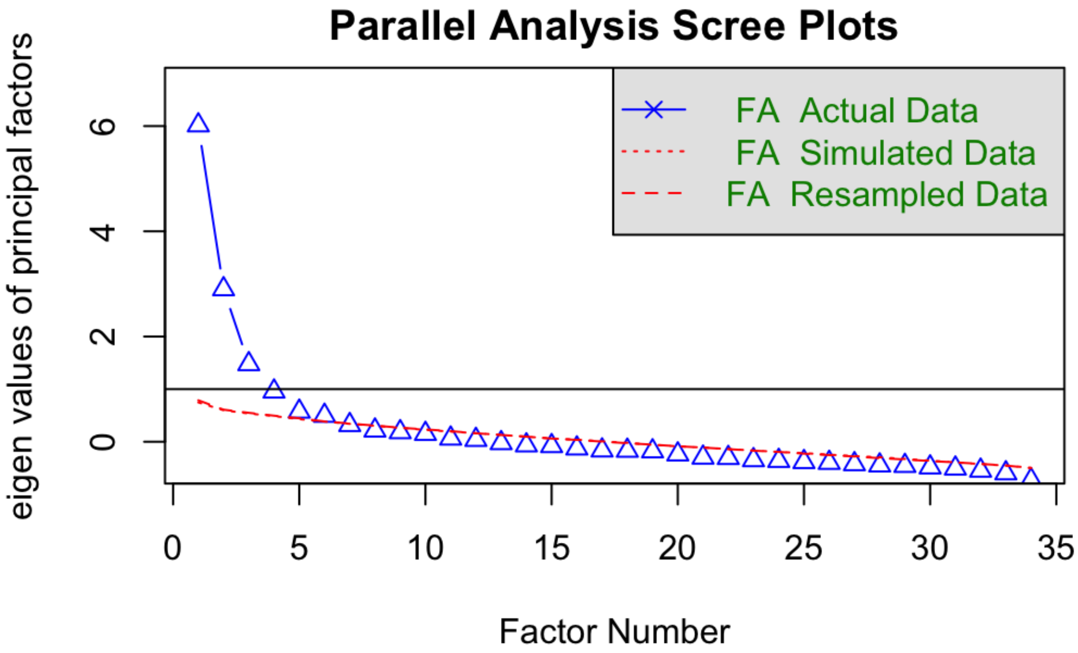
Note. Items retained per factor are in bold font.

EFA Results in R-Studio

A series of EFAs were used to analyze the underlying factors in the STS data using the R-Studio program. The R-Studio program allows for more specific factor analysis parameters when conducting categorical EFAs. Therefore, an ML estimation was used for all EFAs conducted on R as this technique would help refine the parameters of the distribution that best describe the data (Singla, 2018). Much like the SPSS procedures, the first EFA did not include any forced rotations. A parallel analysis and scree plot (Figure 4) examination suggested three- or four-factor models may fit the data best. In congruence with the SPSS EFA results, a six-factor model was also included for further testing.

Figure 4

R-Studio: Scree Plot



Six-Factor Model in R-Studio. An ML estimation was used with a direct oblimin rotation to identify a possible six-factor model to the STS data. After testing all 34 items, 18 items were retained given a factor loading criteria of values greater than .40 (same criteria used for SPSS results). Table 8 shows items retained for the six-factor model in R-Studio with factor loadings for each STS item. R-Studio also produces fit indices for EFA models. This model displayed close fit to the data: RMSEA indicated good fit at 0.047, 90% CI [0.034, 0.059] and good fit within the TLI (0.931) and the CFI (0.966) indices.

Table 8

R-Studio: Six-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1	-0.11	0.01	0.37	0.17	0.22	0.2

2	-0.17	-0.02	0.34	0.07	0.24	0.22
3	-0.07	0.06	**0.49**	-0.02	0.24	0.1
4	-0.06	-0.04	0.34	0.12	0.04	0.14
5	0.03	0.01	**0.52**	-0.01	0	0.15
6	0.07	0.04	**0.64**	0.04	0.03	-0.04
7	0.07	0.07	**0.49**	0	-0.06	0.02
8	0.03	0.18	**0.41**	-0.12	-0.05	0.03
9	-0.06	0	0.09	0.07	-0.04	**0.54**
10	0	0.03	0.17	0	-0.01	**0.56**
11	0.38	-0.01	0.11	0.01	-0.01	0.35
12	0.35	-0.04	0.16	-0.03	-0.09	0.39
13	0.3	0.01	-0.03	0	0.02	**0.48**
14	-0.08	0.1	-0.01	0.17	0.18	0.32
15	0.34	0.14	0.29	-0.04	-0.1	-0.07
16	-0.02	0.14	-0.08	0.2	0.27	0.38
17	0.34	0.03	0.03	0	0.14	0.3
18	**0.7**	0.07	-0.02	-0.02	0.07	0.01
19	**0.76**	-0.02	-0.01	0.03	0.01	-0.01
20	**0.76**	-0.05	0.07	-0.05	0.03	0
21	**0.69**	0	-0.05	0.02	0.03	0.04
22	0.02	-0.2	-0.09	0.11	0.03	-0.02
23	-0.06	-0.02	0.14	**0.81**	0.01	-0.04
24	0.03	0.02	-0.07	**1.01**	0	0.02

25	0.07	0.19	0.21	0.08	-0.13	0.07
26	0.12	0.13	0.18	0.15	-0.2	0.01
27	-0.09	**0.81**	0.07	0.08	0.03	-0.04
28	-0.08	**0.81**	-0.05	-0.04	0.14	0.05
29	0.3	**0.56**	0	-0.03	-0.16	0.03
30	0.15	**0.64**	0.07	0	-0.16	-0.01
31	-0.01	0.1	-0.01	-0.04	**0.65**	0.13
32	0.07	-0.02	0.06	0.09	**0.66**	-0.05
33	0.17	0	0.08	0.11	**0.62**	-0.1
34	0.31	0.17	0.19	0.02	0.11	-0.18

Note. Items retained per factor are in bold font.

Four-Factor Model in R-Studio. An ML estimation was used with a direct oblimin rotation to identify a possible four-factor model to the STS data. After testing all 34 items, 22 items were retained given a factor loading criteria of values greater than .40. Table 9 shows items retained for the four-factor model in R-Studio with factor loadings for each STS item. This model had moderately acceptable fit to the data: RMSEA indicated good fit at 0.059, 90% CI [0.051, 0.068], bad fit in TLI (0.882), and good fit within the CFI (0.924) index.

Table 9

R-Studio: Four-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3	Factor 4
1	**0.6**	-0.07	0.03	0.16
2	**0.61**	-0.13	-0.02	0.07
3	**0.62**	-0.05	0.12	-0.03

4	**0.41**	0	-0.01	0.07
5	**0.51**	0.1	0.07	-0.08
6	**0.47**	0.11	0.15	-0.03
7	0.35	0.12	0.16	-0.08
8	0.3	0.07	0.25	-0.19
9	**0.43**	0.08	-0.05	0.05
10	**0.52**	0.15	-0.02	-0.03
11	0.3	**0.47**	-0.03	-0.01
12	0.33	**0.46**	-0.05	-0.07
13	0.3	**0.41**	-0.06	0.01
14	0.34	-0.02	0.04	0.21
15	0.09	0.35	0.21	-0.1
16	0.37	0.04	0.06	0.28
17	0.3	**0.4**	-0.02	0.03
18	-0.03	**0.7**	0.08	0.02
19	-0.07	**0.76**	0	0.05
20	0.01	**0.76**	-0.03	-0.04
21	-0.05	**0.69**	0	0.04
22	-0.07	0.01	-0.21	0.13
23	0.09	-0.05	0.02	**0.8**
24	-0.03	0.05	0.01	**1**
25	0.13	0.11	0.23	0.02
26	0.03	0.16	0.18	0.08

27	0.05	-0.11	**0.82**	0.08
28	0.08	-0.09	**0.76**	0.02
29	-0.1	0.32	**0.57**	-0.06
30	-0.06	0.16	**0.67**	-0.04
31	**0.43**	-0.04	0.04	0.13
32	0.35	0	-0.02	0.26
33	0.31	0.09	0.01	0.26
34	0.05	0.26	0.23	0.03

Note. Items retained per factor are in bold font.

Three-Factor Model in R-Studio. An ML estimation was used with a direct oblimin rotation to identify a possible three-factor model to the STS data. After testing all 34 items, 26 items were retained given a factor loading criteria of values greater than .40. Table 10 shows items retained for the three-factor model in R-Studio with factor loadings for each STS item. This model had poor fit to the data: RMSEA indicated bad fit at 0.095, 90% CI [0.088, 0.103], bad fit in TLI (0.729), and good fit within the CFI (0.924) index.

Table 10

R-Studio: Three-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3
1	**0.64**	0.04	0.05
2	**0.58**	0.01	0.01
3	**0.5**	0.11	0.15
4	0.39	0.08	0.01
5	0.34	0.24	0.11

6	0.33	0.23	0.18
7	0.2	0.22	0.19
8	0.08	0.18	0.28
9	0.39	0.18	-0.03
10	0.38	0.28	0.02
11	0.2	**0.55**	-0.01
12	0.17	**0.56**	-0.03
13	0.25	**0.48**	-0.05
14	**0.47**	0.01	0.04
15	-0.03	0.39	0.23
16	**0.56**	0.07	0.06
17	0.27	**0.46**	-0.01
18	-0.04	**0.67**	0.06
19	-0.07	**0.71**	-0.01
20	-0.08	**0.76**	-0.03
21	-0.05	**0.66**	-0.01
22	0.05	-0.03	-0.23
23	**0.65**	-0.18	0
24	**0.66**	-0.15	-0.01
25	0.1	0.13	0.24
26	0.05	0.15	0.18
27	0.11	-0.13	**0.82**
28	0.09	-0.09	**0.76**

29	-0.18	0.3	**0.58**
30	-0.12	0.14	**0.68**
31	**0.52**	0.02	0.04
32	**0.56**	0.03	-0.04
33	**0.5**	0.1	-0.01
34	0.06	0.26	0.23

Note. Items retained per factor are in bold font.

EFA Results in FACTOR

A series of EFAs were used to analyze the underlying factors in the STS data using the FACTOR program. Like R-Studio, the FACTOR program allows for more specific factor analysis parameters when conducting categorical EFAs. Therefore, a DWLS estimation with a promax rotation was used for all EFAs conducted on FACTOR as this technique mirrors the techniques used in categorical CFAs. Polychoric correlations were also specified as the dispersion matrix for all FACTOR analyses. The first EFA used optimal procedures for determining the number of dimensions. Parallel analysis on FACTOR is based on minimum rank factor analysis (MRFA; Timmerman & Lorenzo-Seva, 2011) and results recommended a three-factor model when the 95 percentile was considered and a four-factor model when the mean was considered. In congruence with the SPSS EFA results, a six-factor model was also included for further testing.

Six-Factor Model in FACTOR. A DWLS estimation was used with promax rotation to identify a possible six-factor model to the STS data. After testing all 34 items, 30 items were retained given a factor loading criteria of values greater than .40 (same criteria used for SPSS and R results). Table 11 shows items retained for the six-factor model in FACTOR with factor loadings for each STS item. Like R-Studio, FACTOR also produces fit indices for EFA models. Using the researcher-created less-stringent thresholds for determining fit, this model was deemed to have close fit to the data: RMSEA indicated good fit at 0.00, 90% CI [0.00, 0.010]. Though the TLI (1.044) and CFI (1.029) indices exceed the thresholds and are ‘good’, they are also greater than 1 and possibly signal low correlations among variables, which may be due in part to the number of items retained. In this case, the TLI and CFI indices should be rounded to 1, which means they still meet the threshold for ‘good’ values.

Table 11

FACTOR: Six-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1	**0.47**	-0.081	0.021	-0.012	0.371	0.183
2	**0.44**	-0.182	-0.015	-0.156	0.375	0.215
3	0.33	0.098	-0.019	-0.054	**0.489**	0.017
4	0.192	0.264	-0.17	-0.076	0.356	0.096
5	0.004	-0.031	0.3013	0.032	**0.557**	0.171
6	0.091	0.064	0.006	0.005	**0.699**	-0.099
7	-0.042	0.033	0.094	0.046	**0.518**	0.047
8	-0.17	0.116	0.196	-0.013	**0.495**	0.027
9	0.075	-0.003	-0.013	0.005	0.044	**0.595**
10	0.019	-0.137	0.046	0.078	0.184	**0.603**
11	-0.008	0.024	-0.047	**0.467**	0.104	0.347
12	-0.124	0.044	-0.073	**0.445**	0.172	0.389
13	0.051	0.118	-0.049	0.392	-0.119	**0.495**

14	**0.442**	0.03	0.057	-0.047	-0.075	0.304
15	-0.146	0.124	0.173	0.318	0.304	-0.065
16	**0.553**	0.133	0.102	0.023	-0.206	0.365
17	0.19	0.1	-0.047	**0.449**	-0.029	0.255
18	0.082	-0.011	0.064	**0.805**	-0.081	0.017
19	0.003	0.067	-0.027	**0.819**	-0.033	-0.028
20	-0.043	-0.054	-0.064	**0.871**	0.097	-0.018
21	0.036	-0.022	0.033	**0.771**	-0.092	0.041
22	0.226	0.015	-0.302	0.12	-0.153	-0.011
23	**0.669**	0.188	-0.02	-0.249	0.049	0.016
24	**0.772**	0.207	-0.004	-0.099	-0.223	0.039
25	-0.021	**0.41**	0.181	0.022	0.141	0.079
26	-0.069	**0.779**	-0.023	0.065	0.074	-0.039
27	0.198	0.116	**0.78**	-0.167	0.032	-0.071
28	0.219	-0.063	**0.707**	-0.069	-0.006	0.011
29	-0.139	-0.059	**0.708**	0.269	-0.028	0.061
30	-0.102	-0.023	**0.842**	0.071	0.016	-0.001
31	**0.696**	-0.356	0.097	0.097	0.048	0.076
32	**0.809**	-0.129	-0.079	0.178	0.027	-0.139
33	**0.829**	-0.066	-0.061	0.275	0.063	-0.264
34	0.205	0.3	0.107	0.31	0.148	-0.3

Note. Items retained per factor are in bold font.

Four-Factor Model in FACTOR. A DWLS estimation was used with promax rotation to identify a possible four-factor model to the STS data. After testing all 34 items, 29 items were retained given a factor loading criteria of values greater than .40. Table 12 shows items retained for the four-factor model in FACTOR with factor loadings for each STS item. Items 1, 2, and 12 loaded onto two factors and were removed from any further analysis. This model had close fit to the data: RMSEA indicated good fit at 0.00, 90% CI [0.00, 0.010]. The TLI (1.014) and CFI (1.011) indices are greater than the threshold and possibly signal low correlations among variables, which may be due in part to the number of items retained. In this case, the TLI and CFI indices should be rounded to 1, which means they still meet the threshold for 'good' values.

Table 12

FACTOR: Four-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3	Factor 4
1	**0.449**	**0.463**	0.029	-0.115
2	**0.453**	**0.431**	-0.047	-0.155
3	**0.472**	0.293	0.107	-0.061
4	**0.458**	0.191	-0.035	-0.056
5	**0.614**	-0.02	0.068	0.035
6	**0.563**	0.012	0.187	0.02
7	**0.503**	-0.078	0.189	0.04
8	**0.491**	-0.195	0.308	-0.014
9	**0.478**	0.174	-0.149	0.083
10	**0.572**	0.097	-0.107	0.143
11	0.329	0.031	-0.083	**0.513**

12	**0.429**	-0.082	-0.103	**0.497**
13	0.267	0.145	-0.136	**0.472**
14	0.163	**0.506**	-0.014	-0.003
15	0.236	-0.179	0.29	0.312
16	0.112	**0.647**	0.028	0.088
17	0.16	0.23	-0.056	**0.492**
18	-0.109	0.071	0.078	**0.811**
19	-0.079	-0.017	0.033	**0.824**
20	0.012	-0.085	-0.018	**0.864**
21	-0.101	0.03	0.033	**0.78**
22	-0.148	0.225	-0.301	0.121
23	0.109	**0.683**	0.053	-0.237
24	-0.121	**0.814**	0.027	-0.074
25	0.283	0.015	0.312	0.062
26	0.227	-0.029	0.246	0.119
27	0.007	0.221	**0.812**	-0.164
28	-0.016	0.241	**0.663**	-0.068
29	-0.013	0.113	**0.654**	0.28
30	-0.001	0.078	**0.81**	0.076
31	-0.002	**0.676**	-0.003	0.076
32	-0.128	**0.761**	-0.047	0.145
33	-0.178	**0.758**	0.034	0.23
34	-0.035	0.154	0.319	0.289

Note. Items retained per factor are in bold font.

Three-Factor Model in FACTOR. A DWLS estimation was used with promax rotation to identify a possible three-factor model to the STS data. After testing all 34 items, 26 items were retained given a factor loading criteria of values greater than .40. Table 13 shows items retained for the three-factor model in FACTOR with factor loadings for each STS item. This model had close fit to the data: RMSEA indicated good fit at 0.021, 90% CI [0.00, 0.0300], good fit in TLI (0.989), and good fit within the CFI (0.991) index.

Table 13

FACTOR: Three-Factor Model Loadings from EFA

Item	Factor 1	Factor 2	Factor 3
1	0.066	-0.033	**0.755**
2	-0.011	-0.063	**0.74**
3	0.163	0.044	**0.589**
4	0.016	0.066	**0.495**
5	0.16	0.224	0.367
6	0.278	0.18	0.353
7	0.277	0.191	0.224
8	**0.41**	0.138	0.093
9	-0.105	0.224	**0.49**
10	-0.043	0.32	**0.464**
11	-0.056	**0.626**	0.198
12	-0.056	**0.656**	0.158
13	-0.128	**0.556**	0.28

14	-0.026	-0.009	**0.604**
15	0.346	0.389	-0.091
16	0	0.047	**0.693**
17	-0.065	**0.527**	0.282
18	0.042	**0.776**	-0.1
19	0.004	**0.812**	-0.163
20	-0.033	**0.893**	-0.169
21	-0.001	**0.755**	-0.126
22	-0.36	0.074	0.147
23	0.032	-0.289	**0.761**
24	-0.041	-0.208	**0.724**
25	0.369	0.126	0.155
26	0.293	0.176	0.077
27	**0.853**	-0.251	0.142
28	**0.687**	-0.152	0.155
29	**0.694**	0.244	-0.227
30	**0.864**	0.024	-0.182
31	-0.051	-0.001	**0.653**
32	-0.123	0.023	**0.652**
33	-0.047	0.088	**0.598**
34	0.316	0.241	0.058

Note. Items retained per factor are in bold font.

Comparison of Factor Models Across Programs

SPSS was the only program that did not report fit indices for the EFAs and served as a baseline in determining how many factor models to evaluate further using R and FACTOR. A comparison of the reported fit indices are shown in Table 14. Across the programs, R retained the least items (*M* = 20.3), followed by SPSS (*M* = 23.6). FACTOR retained the most items (*M* = 28.3) for each model. Factor loadings in FACTOR were significantly higher across programs and models. Factor loadings in FACTOR were significantly higher across programs and models. In the six-factor models, items 4, 15, and 34 were eliminated across all programs. In four-factor models, items 22, 25, 26, and 34 were eliminated across all programs. In three-factor models, items 5, 6, 7, 15, 22, 25, 26, 34 were eliminated across all programs. SPSS and R retained the same items onto the three-factor models, though the factor loadings were slightly lower in R likely due to the additional parameters placed onto the EFA.

Table 14

Comparison of Reported EFA Fit Indices

	R-Studio			FACTOR		
Fit Index	6-Factor	4-Factor	3-Factor	6-Factor	4-Factor	3-Factor
RMSEA	0.047 (good)	0.059 (good)	0.095 (bad)	0.000 (good)	0.000 (good)	0.021 (good)
TLI	0.931 (good)	0.882 (bad)	0.729 (bad)	1.044 (good)	1.014 (good)	0.989 (good)
CFI	0.966 (good)	0.924 (good)	0.924 (good)	1.029 (good)	1.011 (good)	0.991 (good)
Overall model fit	Close	Acceptable	Poor	Close	Close	Close

Note. SPSS does not report fit indices for EFAs. The researcher-created less-stringent thresholds for determining index and model fit were applied to all R-Studio and FACTOR EFA tests.

Post Hoc CFA Results

The less-stringent approach to determining index and model fit in the EFA tests allowed for wider consideration of alternative factor structures to the STS data. Should the strict approach been used, less models would have met the criteria for further testing. Since the four- and six-factor models from R and FACTOR and the three-factor models from FACTOR showed acceptable to close fit to the STS data in the EFA tests, a categorical CFA with a DWLS estimator was conducted for each (5 models total) using R-Studio and the Lavaan package. Model fit for these categorical CFAs were assessed using the strict cut-off threshold approach seen in the initial CFA tests. Since CFA tests are meant to confirm the exploratory findings, I chose to use the more defined process in determining fit and overall model fit. The less stringent approach to assessing model fit would have identified all models to have close or acceptable fit to the data. The strict approach was therefore chosen to narrow down which models were truly the best fit to the STS data.

In this strict approach, only RMSEA values equal to or less than 0.05, TLI values equal to or greater than 0.95, CFI values equal to or greater than 0.95, and SRMR values less than 0.08 were considered ‘good’ and deemed appropriate to move forward. All other values were reported as ‘bad’. A model with three or more ‘good’ values were deemed to have an overall ‘close’ fit to the data, models with two ‘good’ values were deemed ‘acceptable’ fit to the data, and models with only one ‘good’ value were deemed ‘poor’ fit to the data. A summary of the fit indices for tested models can be seen in Table 15; see Appendix B for images of the final factor models tested.

Table 15

Summary of Fit Indices for Models Tested in the Post Hoc CFAs

	R-Studio		FACTOR		
Fit Index	6-Factor	4-Factor	6-Factor	4-Factor	3-Factor
RMSEA	0.076 (bad)	0.086 (bad)	0.084 (bad)	0.086 (bad)	0.103 (bad)
TLI	0.986 (good)	0.955 (good)	0.936 (bad)	0.947 (good)	0.938 (bad)
CFI	0.989 (good)	0.96 (good)	0.943 (bad)	0.941 (bad)	0.945 (good)
SRMR	0.076 (good)	0.106 (bad)	0.108 (bad)	0.110 (bad)	0.124 (bad)
Overall Model Fit	Close	Acceptable	Poor	Poor	Poor

Note. DWLS CFA fit indices for post hoc tests. The researcher-created strict thresholds for determining index and model fit were applied to all post-hoc CFA tests.

CFA Results for the Six-Factor Models

The model fit for the six-factor model suggested by the R-Studio EFA results show close model fit, with an RMSEA of 0.076 (bad), TLI of 0.986 (good), CFI of 0.989 (good) and SRMR of 0.076 (good). The factor loadings for this model were all significant (see Table 16). However, the model also produced errors within the intercepts (all recorded as 0) which indicate possible issues with the model fit despite fit. Factor one includes four out of six items found within the *goal setting and attainment* construct; factor two includes all four items found within the *student involvement in the IEP* construct; factor three includes two out of four items found within the *employment* construct; factor four includes three out of four items found within the *support community* construct; factor five includes three out of four items found within the *disability awareness* construct; and factor six includes two out of five items found within the *persistence* construct.

The model fit for the six-factor model suggested by the FACTOR EFA results show poor overall model fit, with an RMSEA of 0.084 (bad), TLI of 0.936 (good), CFI of 0.943 (bad) and SRMR of 0.108 (bad). The factor loadings for this model were all significant (see Table 17). The R-six-factor model did fit the data better than the FACTOR-six-factor model.

Table 16

R-Studio: Six-Factor Model Loadings from Post Hoc CFA

	Item	Estimate	S.E.	*p-value*	Std.lv/Std.all
Factor 1	18	1.000	-	-	0.836
	19	0.991	0.064	0.00	0.828
	20	1.037	0.058	0.00	0.866
	21	0.897	0.066	0.00	0.749
Factor 2	27	1.000	-	-	0.897
	28	0.987	0.044	0.00	0.885
	29	0.827	0.047	0.00	0.742
	30	0.893	0.044	0.00	0.800
Factor 3	23	1.000	-	-	0.977
	24	0.984	0.056	0.00	0.961
Factor 4	31	1.000	-	-	0.782
	32	1.044	0.078	0.00	0.816

	33	0.985	0.075	0.00	0.770
Factor 5	5	1.000	-	-	0.699
	6	1.18	0.124	0.00	0.825
	7	0.871	0.123	0.00	0.609
Factor 6	9	1.000	-	-	0.679
	10	1.296	0.201	0.00	0.880

Table 17

FACTOR: Six-Factor Model Loadings from Post Hoc CFA

	Item	Estimate	S.E.	*p-value*	Std.lv/Std.all
Factor 1	1	1.000	-	-	0.804
	2	0.904	0.071	0.00	0.727
	14	0.693	0.064	0.00	0.557
	16	0.849	0.066	0.00	0.683
	23	1.16	0.053	0.00	0.933
	24	1.164	0.056	0.00	0.936
	31	0.83	0.059	0.00	0.667
	32	0.827	0.06	0.00	0.665
	33	0.794	0.063	0.00	0.639
Factor 2	25	1.000	-	-	0.716
	26	0.764	0.193	0.00	0.547
Factor 3	27	1.000	-	-	0.92
	28	0.94	0.048	0.00	0.865

	29	0.802	0.051	0.00	0.738
	30	0.863	0.046	0.00	0.794
Factor 4	11	1.000	-	-	0.719
	12	0.989	0.078	0.00	0.711
	17	0.908	0.084	0.00	0.653
	18	1.07	0.081	0.00	0.769
	19	1.053	0.079	0.00	0.757
	20	1.11	0.079	0.00	0.798
	21	0.971	0.081	0.00	0.698
Factor 5	3	1.000	-	-	0.745
	5	0.915	0.095	0.00	0.682
	6	0.988	0.086	0.00	0.736
	7	0.768	0.093	0.00	0.572
	8	0.652	0.104	0.00	0.486
Factor 6	9	1.000	-	-	0.619
	10	1.185	0.126	0.00	0.733
	13	1.106	0.135	0.00	0.684

CFA Results for the Four-Factor Models

The model fit for the four-factor model suggested by the R-Studio EFA results show acceptable model fit, with an RMSEA of 0.086 (bad), TLI of 0.955 (good), CFI of 0.960 (good) and SRMR of 0.106 (bad). The factor loadings for this model were all significant (see Table 18). Factor one includes three out of five items found within the *persistence* construct and five of six items from the *goal setting and attainment* construct; factor two includes all items found within the *strengths and limitations* construct, two of four items from the *disability awareness* construct, and two of five items from the *persistence* construct; factor three includes all items found within the *student involvement in the IEP* construct; factor four includes two of four items found within the *employment* construct.

The model fit for the four-factor model suggested by the FACTOR EFA results show poor overall model fit, with an RMSEA of 0.086 (bad), TLI of 0.947 (good), CFI of 0.941 (bad) and SRMR of 0.110 (bad). The factor loadings for this model were all significant (see Table 19).

Table 18

R: Four-Factor Model Factor Loadings from Post Hoc CFA

	Item	Estimate	S.E.	*p-value*	Std.lv/Std.all
Factor 1	11	1.000	-	-	0.694
	12	1.015	0.081	0.00	0.705
	13	0.912	0.087	0.00	0.634
	17	0.897	0.084	0.00	0.623
	18	1.11	0.086	0.00	0.771
	19	1.093	0.083	0.00	0.759
	20	1.174	0.084	0.00	0.816
	21	1.002	0.085	0.00	0.696
Factor 2	1	1.000	-	-	0.799
	2	0.907	0.075	0.00	0.725
	3	0.861	0.065	0.00	0.688
	4	0.617	0.076	0.00	0.493

	5	0.813	0.068	0.00	0.65
	6	0.831	0.067	0.00	0.664
	9	0.66	0.072	0.00	0.527
	10	0.809	0.062	0.00	0.646
Factor 3	27	1.000	-	-	0.903
	28	0.964	0.049	0.00	0.87
	29	0.842	0.05	0.00	0.76
	30	0.887	0.046	0.00	0.801
Factor 4	23	1.000	-	-	1.023
	24	0.897	0.064	0.00	0.918

Table 19

FACTOR: Four-Factor Model Factor Loadings from Post Hoc CFA

	Item	Estimate	S.E.	*p-value*	Std.lv/Std.all
Factor 1	3	1	-	-	0.703
	4	0.702	0.094	0.00	0.493
	5	0.904	0.097	0.00	0.635
	6	1.019	0.091	0.00	0.716
	7	0.801	0.095	0.00	0.563
	8	0.689	0.109	0.00	0.484
	9	0.78	0.086	0.00	0.549
	10	0.923	0.087	0.00	0.649
Factor 2	14	1	-	-	0.559

	16	1.269	0.135	0.00	0.709
	23	1.68	0.161	0.00	0.939
	24	1.703	0.162	0.00	0.952
	30	1.228	0.126	0.00	0.686
	32	1.261	0.129	0.00	0.705
	33	1.212	0.133	0.00	0.678
Factor 3	27	1	-	-	0.908
	28	0.965	0.047	0.00	0.877
	29	0.817	0.05	0.00	0.742
	30	0.87	0.046	0.00	0.791
Factor 4	11	1	-	-	0.672
	13	0.921	0.095	0.00	0.619
	17	0.981	0.097	0.00	0.659
	18	1.193	0.099	0.00	0.802
	19	1.14	0.095	0.00	0.766
	20	1.203	0.094	0.00	0.808
	21	1.049	0.096	0.00	0.705

CFA Results for the Three-Factor Model

The model fit for the sole three-factor model suggested by the FACTOR EFA results show overall poor model fit, with an RMSEA of 0.103 (bad), TLI of 0.938 (bad), CFI of 0.945 (good) and SRMR of 0.124 (bad). The factor loadings for this model were all significant (see Table 20).

Table 20

FACTOR: Three-Factor Model Factor Loadings from Post Hoc CFA

	Item	Estimate	S.E.	*p-value*	Std.lv/Std.all
Factor 1	1	1	-	-	0.767
	2	0.916	0.073	0.0	0.703
	3	0.77	0.064	0.0	0.591
	14	0.691	0.066	0.0	0.53
	16	0.866	0.067	0.0	0.664
	23	1.211	0.057	0.0	0.929
	24	1.239	0.061	0.0	0.95
	31	0.866	0.059	0.0	0.664
	32	0.887	0.058	0.0	0.68
	33	0.849	0.065	0.0	0.651
Factor 2	11	1	-	-	0.659
	12	1.008	0.09	0.0	0.664
	13	0.941	0.093	0.0	0.621
	17	0.952	0.094	0.0	0.627
	18	1.207	0.098	0.0	0.795
	19	1.176	0.095	0.0	0.775
	20	1.24	0.097	0.0	0.817
	21	1.089	0.096	0.0	0.718
Factor 3	27	1	-	-	0.91
	28	0.962	0.046	0.0	0.876

29	0.814	0.049	0.0	0.741
30	0.87	0.047	0.0	0.791

Summary of Results

The factor structure for the TAGG-S did not hold within the preliminary step of ME/I testing of the focal and reference groups. Therefore, further ME/I testing was not conducted. Instead, a series of EFA tests were carried out across three different programs to identify alternative factor structure for the STS data. Six-, four-, and three-factor models were tested using three separate statistical software programs. When assessing fit indices for the EFAs, the six- and four-factor models from R-Studio and FACTOR and one three-factor model from FACTOR displayed acceptable model fit and were submitted to post hoc CFA testing. Of the five post hoc models tested, only the six-factor model identified by the R-Studio EFA results showed overall close model fit. However, the four-factor model identified by R-Studio was found to have overall acceptable model fit. All other tested models were deemed to have poor fit to the data. Solely based off model fit, the six-factor model identified by R-Studio is the best fitting model to the data and should be considered the alternative factor structure to the Spanish-translated TAGG-S.

Chapter 5

Discussion

There are numerous Spanish-translated transition assessments for English Language learners/students with disabilities (ELSWD), but very limited research exists on the validation of any these translated assessments (Mumbard´o-Adam et al., 2018; Yang et al., 2005). For all students with disabilities, transition plans in the individualized education program (IEP) should be informed by age-appropriate transition assessment results (Deardorff, 2020; Petcu et al., 2014; Prince et al., 2014). Special education policy and practice recommendations strongly suggest the use of at least one formal/validated assessment for SWDs to avoid legal implications (Prince et al., 2013; 2014). While there are no known court cases regarding the use of translated assessments as appropriate tools for ELSWDs, the consequences of such a case could be detrimental to the field and the legitimacy of any transition plan written for a student who was assessed using translated and non-validated tool.

Measurement equivalence/invariance (ME/I) testing is designed to detect differences in assessment results across groups and can add validity evidence to translated assessments (Boer et al., 2018; Chan, 2011). After extensive research, I found no ME/I evidence reported for any formal comprehensive transition assessment (see Table 1). Confirmation of ME/I evidence for translated assessments would allow special educators to accurately assess the growing linguistically diverse student population in the U.S., including ELSWDs. For this study, I evaluated the psychometric properties of the Spanish-translated transition assessment and goal generator – student version (TAGG-S) with the intention of establishing ME/I validity evidence. In the sections below I summarize and interpret the study findings, address the limitations present in this study, and discuss the implications for the TAGG (Martin et al., 2015), research field, and special education transition practice.

Summary and Interpretation of Key Findings

ME/I procedures were proposed to answer research Questions 1, 1a, and 1b. Results from the preliminary step in ME/I testing (i.e., the initial confirmatory factor analysis or CFA) presented immediate concerns with the factor structure in the Spanish-translated TAGG-S (focal group) data. Because there are no clear or standardized guidelines to evaluating model fit for categorical and categorically-treated data, I created two approaches to assessing model fit: strict and less-stringent (defined in chapter 4). Using the strict threshold approach, I found the factor structure of the focal group sample did not hold, meaning constructs measured in the TAGG do not function equivalently across English-language and Spanish-language TAGG-S groups. Additionally, this finding meant that research Question 2 could not be answered. Although the reference group (English language TAGG-S data) CFA results may be considered acceptable under traditional maximum likelihood (ML) estimation of interval or continuous data, the diagonally weighted least squares (DWLS) estimation did not indicate acceptable model fit for when using the strict threshold approach and fit indices recommended by Newsom (2018) and Shi and Maudeu-Olivares (2020). This meant that I could not proceed with further ME/I testing. A possible explanation for the ill-fitting models was that the data could be more optimally described by an alternative number of factors (Byrne, 1994).

Since further ME/I testing was not possible, I set out to identify if an alternative factor structure existed for the Spanish-translated TAGG-S (also referred to as STS data) using exploratory factor analysis (EFA) tests. Reverting to the EFA stage in factor structure identification is a recommended strategy if ME/I model fit is inadequate (Byrne, 1994; 2010;

2016). EFA procedures provide a logical and statistically appropriate approach to assessing alternative factor structures (Byrne, 1994). Therefore, I performed a series of EFAs and post hoc CFA tests to determine if alternative factor structures fit the STS data better than the original eight-factor structure identified for the TAGG-S (see Hennessey et al., 2018). The reference group data was not subjected to further tests as the sample was matched by the focal group's characteristics (disability category, age, and gender) and is not necessarily representative of the entire English-language TAGG-S data.

The EFAs were conducted using three different statistical software programs (SPSS, R-Studio, and FACTOR) since the variables in the STS data were treated as categorical and required additional specifications. Each program used in this study is capable of including certain specifications and results for each tested structure were compared across programs. SPSS allows for clear unrotated and rotated models but does not use a polychoric correlation adjustment for categorical data. R-Studio can analyze data using a polychoric correlation matrix and an oblimin rotation, which allows factors to correlate. With these specifications, R-Studio can also use an ML estimation which determines the values for the parameters of a model and informs the likelihood that the model produced the data that were actually observed (Brooks-Bartlett, 2018). FACTOR can similarly analyze data using a polychoric correlation matrix and allows for a wide range of rotation options including oblimin and promax. Though oblimin and promax are both oblique rotations appropriate for this study, the promax rotation was chosen since it first conducts an orthogonal varimax rotation and then allows correlations between factors which reveal if any factors do not correlate (Russell, 2002). When a polychoric correlation is specified in FACTOR it does not allow for ML, so a DWLS estimation was used instead which was appropriate as it emulated the specifications used in the initial CFA.

Both the R-Studio and FACTOR programs included tests for parallel analysis within the first round of EFA tests (unrotated) and suggested four- and three-factor models were likely structures to the data. For SPSS results, a separate parallel analysis engine (Patil Vivek et al., 2017) was used in determining the eigenvalues cut off threshold and suggested a four-factor model would suffice. However, since there were no hypotheses formed on the nature of the alternative structure, a less stringent parallel analysis threshold (Kaiser, 1970) in conjunction with scree plot analysis was also used which suggested a six-factor model may also fit the data. A five-factor model was also suggested in this laxed parallel analysis but was eliminated as the majority of items were grouped under one factor and had low loadings.

The second round of EFA tests examined the item loadings of the six-, four-, and three-factor models through forced factor rotations. Results from the nine EFAs (three per program) were compared for model fit using the less-stringent approach to determine fit index thresholds. A cut-off value of .40 (Pituch & Stevens, 2016) was used when determining item loadings within each factor for all programs. Interestingly, the six-factor model across programs showed acceptable fit in this stage even though they were not recommended through the more stringent parallel analysis results. At this point, only the identification of an alternative structure was of concern and the naming and defining of factors typical to EFA/CFA tests was not done for this study as the items loaded differently across programs and required deeper consideration for why some items may correlate and other do not.

Typically, research that explores factor structure through EFAs include confirmation of the proposed factor structure through post hoc CFA tests (Kyriazos, 2018; Tabachnick & Fidell, 2003). Thus, in the final round of tests, post hoc CFAs were conducted on the six- and four-factor models proposed in the EFAs by R-Studio and FACTOR and the three-factor model proposed by FACTOR. These five proposed models were included in the post hoc tests since they all showed acceptable to close fit from the second round of EFA tests. The SPSS models were not included in the post hoc tests since they did not report preliminary fit statistics and offered limited control for model specification needed to assess if further testing was appropriate. The remaining models were analyzed in R-Studio using the same procedures as the initial CFAs and results showed poor model fit for the six-, four-, and three-factor models suggested by FACTOR. The four-factor model suggested by R-Studio showed acceptable fit and the six-factor model suggested by R-Studio met close. Based off model fit, the six-factor model suggested by R-Studio presents the best fit to the data. Interestingly, even though 16 items were not included in the final six-factor model, the retained 18 items were grouped under already existing TAGG constructs (see Table 21).

Table 21

TAGG Constructs and R: Six-Factor Model Items

TAGG constructs	R: Six-factor model	Item
Goal setting and attainment	Factor 1	18
		19
		20
		21
Student involvement in the IEP	Factor 2	27
		28
		29
		30
Employment	Factor 3	23

Support community	Factor 4	24
		31
		32
Disability awareness	Factor 5	33
		5
		6
Persistence	Factor 6	7
		9
		10

Note. Only 18 of the 27 items remained from the six factors identified in the R: six-factor model.

Overall, the results of my study show there are significant issues in the factor structure of the Spanish-translated TAGG-S. Through exploratory tests, I was able to identify one model which produced a better fit to the data over all others explored. The final six-factor model even exhibited items loading onto six of the already existing constructs of the TAGG-S. The final model did produce errors in the parameter estimates (all recorded as 0) and requires further investigation into its structure but stands out as the likely alternative factor structure of the Spanish-translated TAGG-S.

Implications

The most pressing implication of this study is that the Spanish-translated TAGG-S cannot be considered a valid and reliable assessment for Spanish-testing students with disabilities or ELSWDs. This means the Spanish-translated TAGG-S scores/results cannot be interpreted in the same manner as the English-language TAGG-S because each score interpretation requires its own validity evidence (AERA et al.,2014; Kane, 2006). It is recommended for the TAGG to include a disclaimer to users of the Spanish-translated version of the TAGG-S until the six-factor structure identified in this study is confirmed. Before the Zarrow Center began targeting the ELSWD student population, only 34 Spanish-translated TAGG-S assessments were completed between 2015 and 2021, meaning the demand for a Spanish-translated TAGG-S was not high despite the growing population (NCES, 2018). It is important to note that the English-language TAGG-S remains a formal, valid and reliable special education transition assessment. Considering the limited availability of validated formal comprehensive transition assessments, discarding of the English-language TAGG-S would be irresponsible to the many students it does accurately assess. Like most transition assessments, the TAGG is not considered a high-stakes assessment in the same way that college entrance exams or annual state tests are. This means educators of ELSWDs may supplement TAGG results with other assessments to inform the transition plan in the IEP. Though the IEP itself can be considered a high-stake document, explicit written information on *how* and *why* the TAGG was chosen as an age-appropriate transition assessment for a student can circumvent assessment selection concerns. The educator interpreting the results for a student who has completed the Spanish-translated TAGG-S should use evaluative judgement and knowledge of the student to determine if the TAGG-suggested strengths and needs and annual goals are appropriate for the student. This is a recommended practice for any version of the TAGG (Martin et al., 2015). In general, educators who make assessment selection decisions should use their professional judgement in selecting the TAGG for their students, noting that it is most appropriate for students with mild/moderate disabilities who plan to be competitively employed and test in English.

A thorough evaluation of the final six-factor model identified in this study is needed to determine if it does indeed hold and items within the model are representative of the construct and appropriate to the population of test takers. The findings from this study bring us full circle to the problems first identified in chapter 1 and discussed in chapter 2: there is a significant lack of valid and reliable translated transition assessments available for students who prefer or need to test in languages other than English. To compensate for the lack of validity evidence for language-inclusive transition assessments educators should heed Greene’s (2011) advice on increasing culturally responsive communication during the transition assessment process (e.g., learning and understanding the family’s cultural beliefs about disability and transition, asking for and listening to family perspectives, etc.). However, culturally responsive practices alone are not sufficient in addressing this area of need, assessment validation for translated assessments should be considered a top priority for all assessment developers.

This study also raises numerous concerns about assessments, including those used in special education and transition. If assessments are translated into other languages and assumed to produce valid scores without evidence of an iterative translation process and/or evidence of a separate validity studies with the intended testing population, then the translation is just that, a translation. Scores or results produced from these un-tested translated assessments cannot in good confidence inform about the test-taker’s strengths and needs, nor can (or should) those scores be used to make decisions for the test-taker. Although there has not been a high demand for a Spanish-translated TAGG-S in the past, there is a professional obligation to validate any translated assessments if advertised for use in another language (AERA et al., 2014). Additionally, a lack of information on the manner in which an assessment was translated forces one to question if there are errors in the actual translation of assessment items and if the assessment and its individual items are culturally relevant and appropriate for intended testing groups.

The study findings reflect the error the field of special education has made in assuming assessments scores validated in English generalize to other groups when translated. The continued assumption that validation can be generalized to translated assessments is statistically and ethically wrong. These issues are central to the topic of fairness and bias in testing and are discussed at length in the section below.

Fairness in Testing

Per the *Standards,* fairness in testing is an overriding foundational concern for any assessment and as Kane (2010) asserts, there is a need to meet procedural and substantive requirements of fairness in assessment development. Translated assessments technically meet the procedural requirement of fairness by providing a pathway in assessing linguistically minoritized students. On its own however, the procedural requirements for adapting assessments are not sufficient in serving the needs of ELSWDs. Translated transition assessments are not effective in transition planning if we do not also ensure the assessments are actually testing the intended constructs. Any conclusions drawn from a non-validated translated transition assessment is inherently inaccurate and should not be included in the legally binding IEP.

To meet the substantive requirement of fairness, any translated assessment must be properly validated to ensure the translation produces a version of the assessment that is comparable in content and difficulty and also yields reliable scores (AERA et al., 2014). This is not to say that responses should be the same across groups when properly translated and validated assessments are in use. Group differences are likely to occur as we cannot assume that all test takers share the same experiences and knowledge assessed. As one may imagine, meeting the substantive requirement is far more complex and time consuming yet remains a crucial step in ensuring fairness in testing. Literature in education testing and assessment validation emphasize different ways to approach the substantive requirement.

In education, the universal design for assessments is proposed as a means to minimize bias, thereby increasing fairness (AERA et al., 2014; Dolan et al., 2005; Geisinger, 1994; Ketterlin-Geller, 2005; Thompson et al., 2002; Russell et al., 2009). The concept of universal design for assessments mirrors the concept of universal design for learning (UDL; see Capp, 2017) used in education. Universal design calls for clarity and consideration of all test takers including subgroups. Following a universal design in assessment development (or translated assessment development) maximizes fairness and emphasizes the need to develop assessments that are usable for all intended test takers (Dolan et al., 2005; Geisinger, 1994; Ketterlin-Geller, 2005; Thompson et al., 2002; Russell et al., 2009).

Assessment validation literature suggests a more technical solution to meeting the substantive requirement. For example, under Xi's (2010) model, a fairness analysis would evaluate potential challenges of score interpretations. The resulting analysis and subsequent argument would identify potential threats to fairness and validity of the test to its test takers. This analysis-argument approach would address how fairness issues effect decisions and consequences made off score results. ME/I tests as originally proposed in this study or differential item functioning (DIF) through an item-response theory (IRT) framework (Putnick & Bornstein, 2016) can be used to further address and identify the fairness analysis issues.

When evaluating the fairness of an assessment, one must also consider the issue of bias. The three types of biases distinguished in van de Vijver & Tanzer's (2004) taxonomy need to be addressed when evaluating the Spanish-translated TAGG-S. First, a professional evaluation of construct bias is needed to ensure the underlying construct measured is the same across cultures.

This is particularly important when tests to confirm the final six-factor model are conducted. The constructs represented in the final six-factor model may be reflective of the constructs as they were originally conceptualized for English-testing populations but may also carry different meaning or irrelevancy for ELSWDs. Interviewing a content panel of potential test takers could address construct bias and irrelevancy concerns. Second, steps to ensure representative sampling of participants is considered when addressing issues of method bias. Purposeful sampling of students across disability categories, gender, age, geographic location, and racial/ethnic backgrounds is needed to ensure the sample is representative of the U.S. ELSWD population (or as close as possible). Third, an iterative process is needed to ensure that each assessment item is (a) properly translated and (b) reflects the same meaning across cultures. This process should include ELSWDs and their family members. For example, TAGG-S items regarding the student’s disability awareness and advocacy may not be a practice in the student’s culture and perhaps not relevant to their transition from school to community.

Identification of ELSWDs

With the overrepresentation of participant samples from the speech/language disability category present in this study, an additional issue should be considered: the identification of students who are English-language learners and have a disability. The identification of ELSWDs has been of concern for some time (Artiles & Ortiz, 2002; Counts et al., 2018; Hamayan et al., 2007; Zacarian, 2011). There is a significant lack of valid and reliable instruments that have been normed for EL populations (Artiles et al., 2005; Ford, 2012; Morgan et al., 2015; Rueda & Windmueller, 2006; Sullivan, 2011). The current normed tools for ELSWDs are not technically required and other (non-normed) tools are also used during the identification process (Ford, 2012; Hibel & Jasper, 2012; MacSwan & Rolstad, 2006; Morgan et al., 2015; Rueda &

Windmueller, 2006). While the population of ELSWDs is relatively small (14.3% of all EL students; NCES, 2020a), there exists a possibility of misidentification among the participant sample included in this study. This is impossible to control for in extant analysis but the inclusion of responses from misidentified students could explain anomalies in the data.

Additionally, students with a disability in speech/language may have a history of language processing issues, limited use and understanding of complex words and figurative language, reading issues and disorganized output (National Institute of Deafness and Other Communication Disorders, 2010). Per IDEA (2004), a disability diagnosis cannot be due to a limited proficiency in English. Presumably, if the speech/language disability exists for a student in their second language, it should also exist in their first language. In this case, the student's needs in language processing could affect their performance on the TAGG or other translated assessment especially if that assessment has not been properly translated and validated. The topic of ELSWD identification is far more complex than summarized here but requires further attention and consideration when confirming the final structure of the Spanish-translated TAGG-S.

Special Education Policy

Though the specific issue of translated assessments for students with disabilities has not been argued in special education court cases yet, there is a legal precedence wherein the court may order an agency to develop formal assessment tools to evaluate group needs if the needs of the group require significant attention (United States & Mellette v. Jones, 1996). While the majority of [language-related] special education court cases address procedural problems (e.g., providing translation services during IEP meetings and translating documents for parents of students with disabilities; Zimmerman, 2019), the likeliness of future cases addressing substantive problems is high especially when reflecting on the outcomes of the *Endrew* case (i.e., move from procedural to substantive focus in court). Bearing in mind the legal history of special education, it is imperative that we do not wait for a case to be argued in court to force our hand in validating translated transition assessments. Instead, we should be proactive and begin the validation process for in-use translated transition assessments and ensure they are appropriate tools whose results can confidently be used to guide transition planning. The *Standards* assert it is the responsibility of test developers to ensure all versions and adaptations of an assessment are measured and appropriate for the intended testing population. For the most part, that responsibility has gone unfulfilled in special education transition.

As a field driven by policy, transition professionals must set higher expectations for national special education transition resources. The national and publicly available transition assessment guides (see: NTACT, IRIS) and the professional home for special education transition (the Division for Career Development and Transition) must advocate for the use of validated transition assessment tools and provide resources to support educators in selecting validated assessments. In these efforts, there must also be a shift in language used to define validated assessments. The terms *formal* and *informal* used in education disciplines are confounding at best and use of standardized language aligned to the educational assessment research field is better suited to future of transition assessment. National special education transition resources should also reflect trends and practices seen in the education assessment research field to better bridge the gap from research-to-practice.

IDEA has not changed since 2004 and was supposed to be renewed in 2009 (Madlowitz, 2016) but policymakers have been cautious in making changes to the law. While the next reauthorization date is unknown, an update to the law is likely to occur sooner rather than later.

Future reauthorization of IDEA should formally mandate the use of at least one validated transition assessment, specificizing that translated assessment must also have validity evidence. As it stands, language regarding transition assessments in IDEA is vague and knowledge on *how* and *why* valid assessments are needed comes from best practice recommendations and outcomes of court cases (Prince et al., 2014). Any new iteration of IDEA should also emphasize the need to train special educators in evaluating and selecting validated assessments and provide guidance on *what* special educators can do when options for validated assessments, and particularly translated assessments, are limited.

Study Limitations

Three main limitations affected this study: (1) control for extraneous factors, (2) target group testing, (3) statistical program selection, and (4) user error. First, the data were analyzed extant which limited my ability to control for extraneous factors. The most significant issues faced were in the sampling of the focal group; there was a significant overrepresentation of students with a primary disability of speech/language impairment. Though speech/language impairment is the second-most identified disability category among all students with disabilities (NCES, 2020b), the difference between samples in this category outnumber samples of all other disability categories combined (*n* = 125), excluding specific learning disability (*n* = 55). Since the focal group data were matched by three variables including disability category, the reference group data also had an overrepresentation of students with a speech/language impairment. Even though the reference group did not show good fit in the initial CFA tests, three factors may explain poor model fit: (a) sampling, (b) large number of variables, and (c) fit indices. First, the sample demographics very likely skewed results as they were matched with the focal group demographics and are not representative of the entire English-language TAGG-S dataset.

Second, when a large number of variables are included in data analysis (34 variables were included per group for this study), deviation from model fit may be expected and warrants continued consideration. Third, the fit indices used to determine model fit are not standardized or generally agreed upon, so determining model fit is subjective when coupled with other factors. This meant I had to create my own thresholds for determining fit using recommended indices. I chose to utilize a less-stringent approach for exploratory tests to multiple-possible structures and a strict approach for confirmatory tests to ensure the best-fitting model was identified.

Second, extant data analysis also limited my knowledge and understanding of how the target group sample was tested. The language the TAGG-S was completed in (along with the student's primary disability category and their gender) was recorded in the data set and separated for inclusion in the target sample. However, how the assessment was introduced to the students, how the assessment was administered, the racial/ethnic representation of students, and the location of students taking the TAGG in Spanish remain unknown. I also cannot account for how students took the TAGG (e.g., did they read the questions to themselves, use the Spanish audio, or watch the English ASL videos?). The inability to gauge student perception of the assessment and translation after completion or to have bilingual students take the assessment in both languages was also limiting. Having student user feedback may have helped explain issues found in the original structure and informed if the translations were appropriate to Spanish-language test takers. Though limitations in understanding the population characteristics are not unique to extant data analysis, the lack of information on the target group sample only adds to the fact that little is actually known about ELSWDs' transition practices and needs (Trainor, 2016).

Third, the programs selected to analyze data were not entirely comprehensive for the specifications I desired. SPSS proved to be limiting since the AMOS extension could not use categorically treated data to identify fit and the program itself did not produce detailed EFA reports in a similar manner to that of R-Studio and FACTOR but results from the SPSS EFAs were helpful when comparing proposed factor structures. I would have also preferred to conduct EFAs on a program that was capable of a simultaneous promax rotation, ML estimation, and specification of a polychoric correlation, with the additional option to also examine a DWLS estimation. If I were analyzing true categorical data (e.g., ‘yes’/’no’ questions that are recorded as numerical values but do not have mathematical meaning) and not data that was *treated* as categorical (but was really interval/scaled), then FACTOR would have been the best program option. However, since none of the programs used could include all desired specifications, comparisons across each were conducted but judgement of which specifications fit best was rudimentary at best. Perhaps using the same programs as the Hennessey study (Mplus and SAS) would have produced different results for consideration.

Third, user error in program usage may have also occurred since I was most familiar with SPSS prior to this study and had to learn R-Studio and FACTOR as analysis occurred. While materials and guides for both programs exist and were consulted at length (as well as guidance from committee members), my skills and knowledge on usage of the programs was limited. R-Studio proved most difficult to use and perhaps the errors observed in the CFAs may have been mitigated or resolved from a more experienced user of the program. FACTOR proved quite easy to navigate but exploration of the program’s full capabilities was not possible given the timeframe.

Recommendations and Future Research

The TAGG

Given the study findings, the Spanish-translated TAGG-S cannot at this time be considered a formal, valid and reliable transition assessment. I believe the TAGG website should include a notice to the Spanish-translated TAGG-S users indicating of its current ‘informal transition assessment’ status and provide guidance on how the Spanish-translated TAGG-S results should be interpreted until is properly normed for the ELSWD population in the U.S. Along with this, I would not feel confident in suggesting any other translated transition assessment as appropriate for ELSWDs either since none contain sufficient validation evidence. Although the Chinese version of the TAGG-S is also available, it should not be advertised for use until tested appropriately. Validation tests for the American sign language (ASL) videos to TAGG-S items are also needed. Studies evaluating the TAGG ASL videos should include observations of the test takers and user feedback.

To that end, further analyses should be conducted on the final six-factor model identified in this study for the Spanish-translated TAGG-S. If the model continues to meet acceptable fit without output errors, items in each factor should be evaluated to see if they are still representative of the intended construct measured in the original TAGG-S. As shown in Table 21, items under six of the original TAGG constructs are represented in the newly identified model, excluding items with low factor loadings. Theoretically, the items under the new model’s structure should represent the same construct however, additional assessment of this items/constructs are needed. An evaluation of if the items are relevant to ELSWDs in the U.S. is also needed. A possible consideration for this type of evaluation is the inclusion of user feedback from bilingual (English and Spanish) students in the U.S. Bilingual students may be uniquely capable of testing in both languages, providing feedback, and sharing insights into the larger ELSWD population needs.

In the event the new model does not meet acceptable fit, the TAGG developers should begin the translation adaptation process again. This time they should follow guidelines from the international test commission (ITC, 2017) for translating and adapting tests. This involves a systematic translation and back-translation process that will aid in ensuring items in the TAGG-S are appropriately worded and carry the same intended meaning. Within the translation process, TAGG developers should also consider the reading level needs of the student and ensure they are within the same grade range as the English-language TAGG-S. The English-language TAGG-S items are written to be at a fifth-grade reading level (Martin et al., 2015) but the same cannot be said for the translated version. Perhaps the translation of the TAGG-S items are written at a higher level than ELSWDs can access or items are not translated in a manner which denotes the same meaning as the English version. Since an analysis of the translation is not available for review, these questions remain unanswered. Once a confirmed and documented translation is completed, EFA and CFA tests can proceed much in the same fashion as was done in this study provided a sufficiently large sample size. In future EFA/CFA tests, the sample demographics should be relatively equal or reflective of the national ELSWD population. Following a confirmed factor structure, tests for ME/I can proceed as originally planned for in this study.

As mentioned in sections above, the English-language TAGG-S can still be confidently used for secondary students with mild/moderate disabilities. Even though the English-language TAGG-S model was not a good fit to the data, previous studies confirm its structure and have added validity evidence to support its continued usage. The TAGG developers should set their sights towards revalidation of the TAGG in years to come. Revalidation studies are commonly used in other fields (see: Schwalbe, 2009; Staples et al., 2016) and should be conducted for all current English-language TAGG versions (TAGG-S, TAGG-P, and TAGG-F) to ensure the

TAGG remains an appropriate, valid and reliable special education transition assessment. Nevertheless, the TAGG is only about six years old so revalidation is not a significant issue to consider currently. When it does come time to provide additional validity evidence for the TAGG, researchers should note that revalidation studies do not need to follow the same lengthy procedures as initial validation studies and findings do not need to replicate exactly, in fact some degree of "shrinkage" is expected on revalidated assessments using independent samples (Silver et al., 2000).

Assessment Validation Research

The use of three statistical software programs allowed for comparison of different model specifications. Previous studies have conducted similar comparisons across programs for categorical exploratory and confirmatory tests (Beauducel & Herzberg, 2006; DiStefano & Morgan, 2014; Holgado et al., 2008; Li, 2015; Shi and Maudeu-Olivares, 2020). The consensus amongst comparison studies suggests that the use of WLSMV estimation and/or DWLS estimation (when available) is better suited to categorical data than ML estimations, particularly when evaluating factor loadings (Bandalos, 2014; Beauducel & Herzberg, 2006; Li, 2015; Muthén, du Toit, & Spisic, 1997; Newsom, 2018; Rhemtulla et al., 2012). When WLSMV and DWLS are directly compared across programs DiStefano and Morgan (2014) found both produce accurate parameter estimates.

Polychoric correlations also provide a more accurate reproduction of the measurement model than Pearson correlated data (Holgado et al., 2008), which explains why the SPSS models (using Pearson correlation) produced the least reliable models. Related research has also shown that CFAs with WLSMV or DWLS estimations should not be evaluated in the same manner as CFAs with ML estimations (Newsom, 2018; Shi and Maudeu-Olivares, 2020), which is why I used only four indices to evaluate fit as opposed to the full set of fit indices typically used in evaluation model fit of Pearson correlated, ML estimated models. For data treated as categorical, ML specifications may be the preferred estimation option; ML can also be used for comparison purposes when analyzing true categorical data, especially when there is no underlying hypothesis during exploratory analysis (Komorowski et al., 2016; NIST SEMATECH, 2012; Tukey, 1977). Additionally, a standardization method to evaluate model fit is needed for categorical data. I created my own approaches to determining model fit using recommended indices and though they were considerably rigorous (compared to model fit interpretation of continuous data), the formality of the approaches ensured that model fit thresholds were unobjective and clear. A similar standardized approach is needed from measurement evaluation researchers to the field.

Although SPSS results did not include options to further define EFA tests needed for categorically treated data, it does produce similar outputs as evidenced by the same items loading on the three factor models in SPSS and R-Studio. However, it became clear during model comparison that the specified models tested in R-Studio and FACTOR likely better represented the model fit. The differing factor loadings across the programs show that specifications must be thoroughly considered when conducting EFA and CFA tests, especially for data treated as categorical/polychoric. The final CFAs were conducted using a DWLS estimation which is in congruence with the recommended specifications for categorical data. Future iterations of this study should account for differences in model specification and include comparison of at least two programs if possible.

In special education transition assessment literature, there is a significant lack of attention paid to providing any sort of validity evidence for translated assessments. Researchers across academic fields often ignore invariance issues in evaluating translated assessments and compare scores across groups under the assumption that the factor structure holds even though the psychometric basis for doing so is not sound (Van De Schoot, 2015). Researchers and assessment developers must go beyond conducting a CFA and post hoc reliability tests often seen in literature as sole validity evidence of translated assessments (Fakhri, et al., 2012; Nakayama et al., 2015; Valentini et al., 2014). An emphasis on providing ME/I evidence across test taking groups is needed to address the consequential validity of adapting and using translated assessments (Zumbo, 2003).

Researchers and assessment developers should also make information and guidance documents for categorical CFA procedures as widely available as traditional ML CFA procedures. In my pursuit to learn R-Studio and FACTOR for this study and assess which specifications (ML, DWLS, polychoric correlation, etc…) were preferred and available for inclusion in each program, I found numerous resources for conducting EFA/CFAs under the assumption that data are on an interval/continuous scale but remarkably little information on conducting the same tests for categorical data. Peer-reviewed publications provided answers as to *why* certain tests should be run but user guidance on *how* to run such tests are needed to support novice researchers.

Special Education Transition Practice

"Providing access to a test construct becomes particularly challenging for individuals with more than one characteristic that could interfere with test performance; for example … English learners who have moderate cognitive disabilities" (AERA et al., 2014, p. 53). This quote from the *Standards* perfectly summarizes the issues discussed above and reflect why data analysis in this study was more complex than I originally conceptualized. Specific to education testing, the *Standards* suggests education professional may be justified in deviating from standard assessment procedures when an appropriate measurement tool is not available. While this is not an uncommon practice in special education, especially for students with more significant disabilities (Harshaw, 2013; Morningstar, 2010), deviation from the IDEA transition mandates should not be done lightly. I would recommend educators use scores from other standardized tests outside of transition, like state tests or the WIDA ACCESS for ELLs, to provide evidence of a formal transition assessment if the student's post-secondary goals are reliant on their ability to engage in English or in an academic sphere of some sort (e.g., college-bound, on-the-job training in an English speaking/reading environment, etc.). The use of a non-transition standardized test should be combined with informal assessments specifically related to the areas of transition to better inform the transition plan and services listed in the IEP. Although translated informal assessments also lack validity evidence, educators can supplement assessment results with information about the student including their understanding of the concepts tested. The lack of validated translated assessments places special educators in a challenging position. If they want to use formal translated transition assessments, the most appropriate alternative would be the SDI:SR since it has been normed on Spanish-testing populations. However, I would caution users since the norming sample includes students without disabilities and was conducted outside the U.S (Mumbard´o-Adam, 2018).

Perhaps a larger issue to consider is the preparation of educators to recognize which assessment adaptations provide scores that are comparable to the scores from the original, un-adapted assessment (AERA et al., 2014). Educators are not typically trained to evaluate assessments so much as they are to administer them (Rudner & Schafer, 2002). With a lack of training and understanding of concepts in validity, educators may select measurement tools that do not meet the IDEA and best practice recommendations for formal and informal assessments.

These errors in transition assessment selection could lead to costly and complicated legal consequences like in the *District of Columbia Public Schools, 111 LRP 26012 (SEA DC 2011)* case where failure to use assessment instruments specifically designed for the student's unique needs/abilities lead the court to order additional assessments, compensatory education, private tutoring, and counseling. To avoid these complications, districts/schools should provide assessment evaluation training to educators, especially special educators who have the ability to select any assessment they deem appropriate for their students (as opposed to special educators who are told which assessments to administer). Special education teacher preparation programs should also include components of evaluating an assessment's psychometric properties in teacher preparation courses to better prepare them for assessment evaluation when teaching.

Conclusion

In conclusion, the constructs measured in the TAGG did not function equivalently across English-language and Spanish-language TAGG-S groups. An alternative six-factor structure was identified but requires further testing. Once the final model is formally validated, the Spanish-translated TAGG-S will be the only formal, valid and reliable transition assessment for ELSWDs in the U.S. Though the results of this study are not as I originally hypothesized, I believe the work conducted and reported here is important to the future dissemination and use of the TAGG. I also believe this work will have a significant impact in ensuring ELSWDs have access to appropriately translated and validated assessment tools to better support their post-secondary outcomes once the Spanish-translated TAGG-S studies continue.

References

Adam Bujang, M., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for cronbach's alpha test: A simple guide for researchers. *Malays Journal of Medical Science, 25*(6), 85-99. http://doi.org/10.21315/mjms2018.25.6.9

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

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2), 201-238.

Anglim, J. (2014). *What to do when CFA fit for multi-item scale is bad?* https://stats.stackexchange.com/questions/111821/what-to-do-when-cfa-fit-for-multi-item-scale-is-bad

Ansell, D., Morse, J., Nollan, K. & Hoskins, R. (2004). Life Skills Guidebook. Casey Family Programs.

Artiles, A., & Ortiz. A. (2002). *English language learners with special education needs: Assessment, identification, and instruction.* Center for Applied Linguistics. https://eric.ed.gov/?id=ED482995

Artiles, A. J., Rueda, R., Salazar, J. J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children, 71*(3), 283–300. https://doi.org/10.1177/001440290507100305

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., & Tahan, K. (2011). *The condition of education 2011*. U.S. Department of Education, National Center for Education Statistics.

Bandalos, D. L. (2014). Relative performance of categorical diagnoally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 102-116. https://doi.org/10.1080/10705511.2014.859510

Banks, J. (2014). Barriers and supports to postsecondary transition: Case studies of african american students with disabilities. *Remedial and Special Education, 35*(1), 28-39. https://doi.org/10.1177/0741932513512209

Beauducel, A., & Yorck Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2

Blackorby, J., & Wagner, M. (1996). Longitudinal postschool outcomes of youth with disabilities: Findings from the National Longitudinal Transition Study. *Exceptional Children, 62*(5), 399-413. https://doi.org/10.1177/001440299606200502

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley.

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*(5), 713-734. https://doi.org/10.1177/0022022117749042

Bressani, R. V., & Down, A. C. (2002). Youth independent living assessment: Testing the equivalence of web and paper/pencil versions of the ansell-casey life skills assessment. *Computers in Human Behavior, 18*(1), 453-464. https://doi.org/10.1016/S0747-5632(01)00053-X

Brolin, D. E. (1983). Career education: Where do we go from here? *Career Development for Exceptional Individuals*, *6*, 3-14. https://doi/.org/10.1177/088572888300600101

Brooks-Bartlett, J. (2018). *Probability concepts explained: Maximum likelihood estimation.* Towards Data Science. https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

Brown, M. B. (2001). Review of the Self-Directed Search, 4th edition. In B.S. Blake & J. C. Impara (Eds.), *The fourteenth mental measurements yearbook.* http://web.b.ebscohost.com.ezproxy.lib.ou.edu/ehost/detail/detail?vid=4&sid=76d6ec70-c808-4573-bb75-942e38e54a51%40pdc-v-sessmgr05&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=mmt&AN=test.1978

Buchanan, E. M. (2020). *Graduate statistics.* Statstools + Statistics of Doom. http://doi.org/10.17605/OSF.IO/X5GNJ

Burnes, J. J., Martin, J. E., Terry, R., McConnell, A. E., & Hennessey, M. N. (2018). Predicting postsecondary education and employment outcomes results from the transition assessment and goal generator. *Career Development and Transition for Exceptional Individuals, 41*(2), 111-121. https://doi.org/10.1177/2165143417705353

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*(3), 289-311. https://doi.org/10.1207/s15327906mbr2903_5

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Lawrence Erlbaum Associates.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). Routledge.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.

Byrne, B. M., & van de Vijver, F. J. R. (2017). *Psicothema, 29*(4), 539-551. https://www.redalyc.org/pdf/727/72753218017.pdf

Cangur, S., & Ercan, I. (2015). Comparison of model fit indices used in structural equation modeling under multivariate normality. *Journal of Modern Applied Statistical Methods, 14*(1), 152-167. https://doi.org/10.22237/jmasm/1430453580

Capp, M. J. (2017). The effectiveness of universal design for learning: A meta-analysis of literature between 2013 and 2016. *International Journal of Inclusive Education, 21*(8), 791-807. https://doi.org/10.1080/13603116.2017.1325074

Carnine, D. (1997). Building the research-to-practice gap. *Exceptional Children, 63*(4). https://doi.org/10.1136/ip.2006.014159

Carter, E. W., Austin, D., & Trainor, A. A. (2012). Predictors of postschool employment outcomes for young adults with severe disabilities. *Journal of Disability Policy Studies, 23*(1), 50–63. https://doi.org/10.1177/1044207311414680

Chan, D. (2011). Advances in analytical strategies. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology* (Vol. 1, pp. 85–113). American Psychological Association.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Choiseul-Praslin, B. & Sinclair, T. E. (2021). *Analysis of the Spanish-translated Transition Assessment and Goal Generator* (*TAGG*). Submitted for publication.

Churchman, C. W. (1971). *Design of inquiring systems.* Basic Books Inc.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1). https://doi.org/10.1037/a0026975

Clark, G. M., & Patton, J. R. (2006). *Transition planning inventory– Updated version: Administration and resource guide.* ProEd.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98. https://doi.org/10.1037/0021-9010.78.1.98

Counts, J., Katsiyannis, A., & Whitfrd, D. K. (2018). Culturally and linguistically diverse learners in special education: English learners. *NASSP Bulletin, 102*(1), 5-21. https://doi.org/10.1177/0192636518755945

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4). https://doi.org/10.1037/h0040957

Daradkeh, S., & Khader, Y. S. (2008). Translation and validation of the Arabic version of the geriatric oral health assessment index (GOHAI). *Journal of Oral Science, 50*(4), 453-459. https://www.jstage.jst.go.jp/article/josnusd/50/4/50_4_453/_pdf

Deardorff, M. E. (2020). *The effects of professional development on transition plan components* [Unpublished doctoral dissertation]. University of Oklahoma.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 425-438. https://doi.org/10.1080/10705511.2014.915373

District of Columbia Pub. Sch., *111 LRP 26012.* (SEA DC, 2011)

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *The Journal of Technology, Learning, and Assessment, 3*(7), 1-33, https://ejournals.bc.edu/index.php/jtla/article/view/1660

Dugger, R. (1965). The vocational education act of 1963. *The Bulletin of the National Association of Secondary School Principles, 49,* 15-23. https://doi.org/10.1177/019263656504930104

El-Kazimi, N. (2012). *Influence of disability and gender on transition assessment and goal generator (TAGG) scores* [Unpublished doctoral dissertation]. University of Oklahoma.

Endrew F. v. Douglas County School District RE-I, 137 S. Ct. 988 (2017).

Fakhri, A., Pakpour, A. H., Burri, A., Morshedi, H., & Zeidi, I. M. (2012). The female sexual function index: Translation and validation of an Iranian version. *Journal of Sexual Medicine*, *9*(2), 514–523. http://doi.org/10.1111/j.1743-6109.2011.02553.x

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research ad practice in education: A conceptual framework. *Educational Researcher*, *47*(4), 235-245. https://doi.org/10.3102/0013189X18761042

Field, A. (2000). *Discovering Statistics using SPSS for Windows.* Sage publications.

Field, A. (2009). *Discovering statistics using SPSSL (and sex and drugs and rock 'n' roll)* (3rd ed.). Sage.

Flexer, R. W., Davis, A. W., Baer, R. M., McMahan Queen, R., & Meindl, R. S. (2011). An epidemiological model of transition and postschool outcomes. *Career Development for Exceptional Individuals*, *34*(2), 83-94. https://doi.org/10.1177/0885728810387922

Flora, D. B., & Flake, J. K. (2017). The purpose and practice in exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science, 49*(2), 78-88. https://doi.org/10.1037/cbs0000069

Ford, D. Y. (2012). Culturally different students in special education: Looking backward to move forward. *Exceptional Children, 78*(4), 391-405. https://doi.org/10.1177/001440291207800401

Frey, B. B. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation.* SAGE. http://methods.sagepub.com/Reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304-312. https://doi.org/10.1037/1040-3590.6.4.304

Gershwin Mueller, T. (2015). Litigation and special education: The past, present, and future direction for resolving conflicts between parents and school districts. *Journal of Disability Policy Studies, 26*(3), 135-143. https://doi.org/10.1177/1044207314533382

Glen, S. (2020a). Consequential validity: Definition, examples. https://www.statisticshowto.com/consequential-validity/

Glen, S. (2020b). *Polychoric Correlation*. https://www.statisticshowto.com/polychoric-correlation/

Greene, G. (2011). *Transition planning for culturally and linguistically diverse youth*. Paul H. Brookes Publishing Co.

Green, S. B., Lissitz, R.W., and Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*(4), 827–838. https://doi.org/10.1177/001316447703700403

Greenwood, C. G., & Abbott, M. (2011). The research to practice gap in special education. *Teacher Education and Special Education, 24*(4), 276-289. https://doi.org/10.1177/088840640102400403

Grigal, M., Hart, D., & Migliore, A. (2011). Comparing the transition planning, postsecondary education and employment outcomes of students with intellectual and other disabilities. *Career Development for Exceptional Individuals, 34*(1), 4-17. Halpern, A. S., (1985). https://doi.org/10.1177/0885728811399091

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Erlbaum.

Halpern, A. S., (1985). Transition: A look at the foundations. *Exceptional Children, 51*(6), https://doi.org/10.1177/001440298505100604

Halpern, A. S. (1992). Transition: Old wine in new bottles. *Exceptional Children, 58*(3), 202-211. https://eric.ed.gov/?id=EJ439575

Hamayan, E., Marler, B., Sanchez Lopez, C., & Damico, J. (2007). *Special education considerations for English language learners: Delivering a continuum of services*. Caslon.

Hambelton, R. K., Merenda, P. F., & Spielberg, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Lawrence Elrbaum Associates, Inc.

Harshaw, K. (2013). *Transition assessment in planning for students with most significant cognitive disabilities.* NSTTAC Institute. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Special-Education-Services/Documents/Transition/Callaway%20Transition%20Institute/Transition%20Assessment%20SWSCD.pdf

Hennessey, M. N., Terry, R., Martin, J. E., McConnell, A. E., & Willis, D. M. (2018). Factor structure and basic psychometric properties of the transition assessment and goal generator. *Career Development and Transition for Exceptional Individuals, 36*(2), 174-187. https://doi.org/10.1177/2165143417691021

Hibel, J., & Jasper, A. D. (2012). Delayed special education placement for learning disabilities among children of immigrants. *Social Forces, 91*(2), 503-530. https://doi.org/10.1093/sf/sos092

Holgado, F. P., Chacon-Mascoso, S., Barbero-Garcia, I., & Vila-Abad, E. (2010). Polychoric versus pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44*, 153-166. https://doi.org/10.1007/s11135-008-9190-y

Holland, J. L. & Messer, M. A. (2017). *Self-Directed Search, 5th edition.* https://www.parinc.com/products/pkey/396

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel- Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Erlbaum

Holland, J. L., Powell, A. B., & Fritzsche, B. A. (1997). *Self-directed search, 4th edition*. http://web.b.ebscohost.com.ezproxy.lib.ou.edu/ehost/detail/detail?vid=11&sid=a265680d-ffd4-4abd-ab76-5e074f9756a9%40pdc-v-sessmgr01&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#AN=test.1978&db=mmt

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. https://doi.org/10.1080/03610739208253916

Individuals with Disabilities Education Act, 20 U.S.C. § 1300 (2004). https://sites.ed.gov/idea/

International test commission. (2019). *ITC guidelines for translating and adapting tests - second edition.* https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

IBM. (2020). *Exploratory factor analysis with categorical variables.* https://www.ibm.com/support/pages/exploratory-factor-analysis-categorical-variables

IRIS. (2020). *Transition assessments.* https://iris.peabody.vanderbilt.edu/module/cou2/cresource/q2/p05/

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35,* 401-415. https://doi.org/10.1007/BF02291817

Kane, M. T. (2006). Validation . In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64) . American Council on Education/Praeger Publishers.

Kane, M. T. (2010). Validity and fairness. *Language Testing, 27*(2), 177-182. https://doi.org/10.1177/026553220934946

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment, 4*(2), 1-23. https://ejournals.bc.edu/index.php/jtla/article/view/1649

Kenny, D. A. (2020). *Measuring model fit.* Catalogue of Fit Indices. http://www.davidakenny.net/cm/fit.htm#:~:text=The%20SRMR%20is%20an%20absolute,correlation%20and%20the%20predicted%20correlation.&text=Because%20the%20SRMR%20is%20an,A%20value%20less%20than%20.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.

Kohler, P. D. (1998). Transition planning inventory. *Assessment for Effective Intervention, 24*(1), 249-256. https://doi.org/10.1177/153450849902400423

Komorowski M., Marshall D.C., Salciccioli J.D., & Crutain Y. (2016) Exploratory data analysis. In *Secondary Analysis of Electronic Health Records*. Springer. https://doi.org/10.1007/978-3-319-43742-2_15

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology, 9,* 2207-2230. https://doi.org/10.4236/psych.2018.98126

Laarson, C., Grangerg Axell, A., & Ersson, A. (2007). Confusion assessment method for the intensive care unit (C\AM-ICU): Translation, retranslation, and validation into Swedish intensive care settings. *ACTA Anaesthesiologica Scandinavia*, *51*, 888-892. https://doi.org/10.1111/j.1399-6576.2007.01340.x

Leake, D., & Black, R. (2005). *Cultural and linguistic diversity: Implications for transition personnel.* National Center on Secondary Education and Transition. https://files.eric.ed.gov/fulltext/ED495863.pdf

Lee, S. T. H. (2018). *Testing for measurement invariance: Does your measure mean the same thing for different participants.* https://www.psychologicalscience.org/observer/testing-for-measurement-invariance#:~:text=Testing%20for%20measurement%20invariance%20plays,are%20both%20meaningful%20and%20valid.

Li, C-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavioral Research Methods*, *48*, 936-949. https://doi.org/10.3758/s13428-015-0619-7

Lissitz, R. W. (2009). *The concepts of validity: Revisions, new directions, and applications.* Information Age Publishing Inc.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437-448. https://doi.org/10.3102/0013189X07311286

Lorenzo-Seva. U., & Ferrando, P. J. (2006). FACTOR. https://psico.fcep.urv.cat/utilitats/factor/index.html

MacSwan, J., & Rolstad, K. (2006). How language proficiency tests mislead us about ability: Implications for English language learner placement in special education. Teachers *College Record*, *108*, 2304-2328. https://doi.org/10.1111/j.1467-9620.2006.00783.x

Madaus, J. W., & Shaw, S. F., (2006). The impact of the IDEA 2004 on transition to college for students with learning disabilities. *Learning Disabilities Practice*, *21*(4), 273-281. https://doi.org/10.1111/j.1540-5826.2006.00223.x

Mandlawitz, M. (2016). Special education after 40 years: What lies ahead? *Policy Priorities, 22*(1), 1-7. http://www.ascd.org/publications/newsletters/policy-priorities/vol22/num01/Special-Education-After-40-Years@-What-Lies-Ahead%C2%A2.aspx

Mangal, S. K. (2010). *Statistics in Psychology and Education* (2nd ed.). PHI Learning.

Martella, R. C., Nelson, J. R., Morgan, R. L., & Marchand-Martella, N. E. (2013). *Understanding and interpreting educational research*. The Guilford Press.

Martin, J. D. (2013). *Examining the measurement invariance of the transition assessment and goal generator across percent of time spent in general education* [Unpublished doctoral dissertation]. University of Oklahoma.

Martin, J., Hennessey, M., McConnell, A., Terry, R., & Willis, D. (2015). *TAGG technical manual.* https://tagg.ou.edu/tagg/.

Martin, G. L., Thorsteinsson, J. R., Yu, C. T., Martin, T. L., & Vause, T. (2008). The assessment of basic learning abilities test for predicting learning of persons with intellectual disabilities: A review. *Behavior Modification, 32*(2), 228-247. https://doi.org/10.1177/0145445507309022

Massa-Carrol, I. (2018). Test review: Transition planning inventory – second edition (TPI-2). *Journal of Psychoeducational Assessment, 36*(3), 297-301. https://doi.org/10.1177/0734282916677434

McCarney, S. B., & Arthaud, T. J. (2012). *Transition behavior scale third edition (TBS-3).* https://www.hawthorne-ed.com/images/transition/samples/h04750.pdf

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain and corporate applications* (3rd Ed). Springer.

McConnell, A. E. (2012). *The relationships among academic, GPA, and the transition assessment and goal generator (TAGG) in students with mild to moderate disabilities* [Unpublished doctoral dissertation]. University of Oklahoma.

McConnell, A. E., Martin, J. E., & Hennessey, M. N. (2015). Indicators of postsecondary employment and education for youth with disabilities in relation to GPA and general education. *Remedial and Special Education, 36*(6), 327-336. https://doi.org/10.1177/0741932515583497

McConnell, A. E., Williams-Diehm, K. L., Sinclair, T., Suk, A., & Willis, D. (2020). Transition assessment and goal generator (TAGG): Useful tool to assess non-academic skills. In Yuen, M., Beamish, W., & Solberg, W. S. H. (Eds.). *Careers for students with special education needs*. Springer.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp.13-103). American Council on Education.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-121. https://doi.org/10.21500/20112084.857

Mithaug, D. E., Horiuchi, C. N., & Fanning, P. N. (1985). A report on the Colorado statewide follow-up survey of special education students. *Exceptional Children*, *51*(5), 397-404. https://doi.org/10.1177/001440298505100505

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook,M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal across five disability conditions. Educational Researcher, 44, 278-292. https://doi.org/10.3102/0013189X15591157

Morningstar, M. E. (2009). *Assessing students with significant disabilities for supported adulthood: Exploring appropriate transition assessments.* NCSTTA. https://www.crporegon.org/cms/lib/OR01928264/Centricity/Domain/45/Documents/ToolsforTransitionAssessmentforStudentswithSevereCognitiveDisabilitiesMorningstar.pdf

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Chapter 4: Validity in educational assessment. *Review of Research in Education, 30*(1), 109-162. https://doi.org/10.3102/0091732X030001109

Mumbard´o-Adam, C., Guardia-Olmos, J., Gine, C., Shogren, K. A., & Sanchez, E. V. (2018). Psychometric properties of the Spanish version of the self-determination inventory student self-report: A structural equation modeling approach. *American Journal on Intellectual and Developmental Disabilities, 123*(6), 545-557. https://doi.org/10.1352/1944-7558-123.6.545

Muthen, B. O. (2022). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81-117. http://cda.psych.uiuc.edu/CovarianceStructureAnalysis/Readings/Muthen-beyond_SEM.pdf

Muthén , B.O, du Toit, S., & Spisic,, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript. https://www.statmodel.com/download/Article_075.pdf

Nakayama, K., Osaka, W., Togari, T., Ishikawa, H., Yonekura, Y., Sekido, A., & Matsumoto, M. (2015). Comprehensive health literacy in Japan is lower than in Europe: a validated Japanese-language assessment of health literacy. *BMC Public Health, 15,* 505 (2015). https://doi.org/10.1186/s12889-015-1835-x

National Center for Education Outcomes. (2011). Understanding subgroups in common state assessments: Special education students and ELLs. https://nceo.umn.edu/docs/OnlinePubs/briefs/brief04/NCEOBrief4.pdf

National Center for Education Statistics. (2018). *Enrollment and percentage distribution of enrollment in public elementary and secondary schools, by race/ethnicity and region: Selected years, fall 1995 through fall 2025.* https://nces.ed.gov/programs/digest/d15/tables/dt15_203.50.asp

National Center for Education Statistics. (2019). Indicator 9: Students with disabilities. https://nces.ed.gov/programs/raceindicators/indicator_RBD.asp

National Center for Education Statistics. (2020a). *English language learners in public schools.* https://nces.ed.gov/programs/coe/indicator_cgf.asp

National Center for Learning Disabilities. (2020b). *Students with disabilities.* https://nces.ed.gov/programs/coe/indicator_cgg.asp

National Institute of Deafness and Other Communication Disorders. (2010). *Speech language impairment.* National Institute of Health. https://www.nidcd.nih.gov/health/specific-language-impairment

National Secondary Transition Technical Assistance Center & Office of Special Education Programs. (2007). *Indicator 13 checklist.*

https://transitionta.org/sites/default/files/transitionplanning/NSTTAC_ChecklistFormA.pdf

National Technical Assistance Center on Transition (NTACT). (2016). *Age appropriate transition assessment toolkit fourth edition.* *https://transitionta.org/system/files/toolkitassessment/AgeAppropriateTransitionAssessmentToolkit2016_COMPLETE_11_21_16.pdf*

Neubert, D. A., & Leconte, P. J. (2013). Age-appropriate transition assessment: The position of the Division on Career Development and Transition. *Career Development and Transition for Exceptional Individuals, 36*(2), 72-83. https://doi.org/10.1177/2165143413487768

Newman, L., Wagner, M., Cameto, R., & Knokey, A. (2009). *The post-high school outcomes of youth with disabilities up to 4 years after high school. A report of findings from the national longitudinal transition study-2.* https://files.eric.ed.gov/fulltext/ED505448.pdf

Newman, L., Wagner, M., Knokey, A.-M., Marder, C., Nagle, K., Shaver, D., Wei, X., with Cameto, R., Contreras, E., Ferguson, K., Greene, S., and Schwarting, M. (2011). *The post-high school outcomes of young adults with disabilities up to 8 years after high school. A report from the national longitudinal transition study-2 (NLTS2).* https://files.eric.ed.gov/fulltext/ED524044.pdf

Newsom, J. T. (2005). *A quick primer on exploratory factor analysis.* https://web.cortland.edu/andersmd/psy341/efa.pdf

Newsom, J. T. (2018). *Practical approaches to dealing with nonnominal and categorical variables.* http://web.pdx.edu/~newsomj/semclass/ho_estimate2.pdf

NIST SEMATECH. (2012). *E-handbook of statistical methods.* Engineering Statistics. https://www.itl.nist.gov/div898/handbook/index.htm

Nunnally, J. C. (1967). *Psychometric theory*. McGraw-Hill.

Office of Special Education and Rehabilitative Services (OSERS). (2017). *A Transition Guide to Postsecondary Education and Employment for Students and Youth with Disabilities*. U.S. Department of Education. https://www2.ed.gov/about/offices/list/osers/transition/products/postsecondary-transition-guide-may-2017.pdf

Oswald, D. P., Coutinho, M. J., & Best, A. M. (2002). Racial disparities in the identification, funding, and provision of special education. In D. J. Losen & G. Orfield (Eds.), *Racial inequity in special education* (pp. 1–14). Harvard Education Press.

PAR. (2021). *Self-directed search 5th edition.* https://www.parinc.com/

Parker, A., & Cross, C. T. (2020). *Response to "let's rebuild special education when schools reopen".* https://fordhaminstitute.org/national/commentary/response-lets-rebuild-special-education-when-schools-reopen

Patil, V. H., Singh, S. N., Mishra, S., & Donavan, D. T. (2017). *Parallel analysis engine to aid in determining number of factors to retain using R*. Gonzaga University. https://analytics.gonzaga.edu/parallelengine/

Patton, J. R., & Clark, G. M. (2021). *TPI-3: Transition planning inventory – third edition complete kit.* Proed.

Petcu, S. D., Yell, M. L., Cholewicki, J. M., & Plotner, A. J. (2014). Issue of policy and law in transition services: Implications from special education leaders. *Journal of Special Education Leadership, 27*(2), 66-75. https://eric.ed.gov/?id=EJ1088823

Phelan, C., & Wren, J. (2006). *Exploring reliability in academic assessment.* https://chfasoa.uni.edu/reliabilityandvalidity.htm

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburg, M. (2009). *The guidelines for the assessment of english-language learners.* https://www.ets.org/s/about/pdf/ell_guidelines.pdf

Pituch, K. A., & Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences* (6th ed.). Routledge.

Popham, W. (2008). A misunderstood grail. *Educational Leadership, 66*(1), 82-83. http://www.ascd.org/publications/educational-leadership/sept08/vol66/num01/A-Misunderstood-Grail.aspx

Povenmire-Kirk, T. C., Lindstrom, L., & Bullis, M. (2010). De escuela a la vida adulta/from school to adult life: Transition needs for Latino youth with disabilities and their families. *Career Development for Exceptional Individuals, 33*(1), 41-51. https://doi.org/10.1177/0885728809359004

Price, P. C., Jhangiani, R., Chiang, I C. A., Leighton, D. C., & Cuttler, C. (2017). *Research methods in psychology.* PressBooks.

Prince, A. M. T., Katsiyannis, A., & Farmer, J. (2013). Postsecondary transition under IDEA 2004: A legal update. *Intervention in School and Clinic, 48*(5), 286-293. https://doi.org/10.1177/1053451212472233

Prince, A. M. T., Plotner, A. J., & Yell, M. L. (2014). Postsecondary transition and the courts: An update. *Journal of Disability Policy Studies, 25*(1), 41-47. https://doi.org/10.1177/1044207314530469

Prince, A. M., Yell, M. L., & Katsiyannis, A. (2018). Endrew F. v. Douglas County School District (2017): The U.S. supreme court and special education. *Intervention in School and Clinic, 53*(5), 321-324. https://doi.org/10.1177/1053451217736867

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Rehfeldt, J. D., Clark, G. M., & Lee, S. W. (2012). The effects of using the transition planning inventory and a structured IEP process as a transition planning intervention on IEP meeting outcomes. *Remedial and Special Education, 33*(1), 48-58. https://doi.org/10.1177/0741932510366038

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, *17*(3), 354-373. https://doi.org/10.1037/a0029315

Rueda, R., & Windmueller, M. P. (2006). English language learners, LD, and overrepresentation: A multiple-level analysis. *Journal of Learning Disabilities, 39*(2), 99-107. https://doi.org/10.1177/00222194060390020801

Rudner, L. M., & Schafer, W. D. (2002). *What teachers need to know about assessment.* National Education Association. http://schoolofeducators.com/wp-content/uploads/2015/10/teachers.pdf

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116. https://doi.org/10.7334/psicothema2013.260

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://www.jstatsoft.org/article/view/v048i02

Rowe, D. A., Alverson, C. Y., Unruh, D. K., Fowler, C. H., Kellems, R., & Test, D. W. (2015). A delphi study to operationalize evidence-based predictors in secondary transition. *Career Development on Transition for Exceptional Individuals, 38*(2), 113-126. https://doi.org/10.1177/2165143414526429

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis. *Personality and Social Psychological Bulletin, 28*(2), 1629-1646. https://doi.org/10.1177/014616702237645

Russel, M., Hoffman, T., & Higgins, J. (2009). Meeting the needs of all students: A universal design approach to computer-based testing. *Innovate, 5*(4), 1-6. http://innovateonline.info/?view=article&id=676

Rutkowski, L., Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57. https://doi.org/10.1177/0013164413498257

Savalei, V. (2020). Improving fit indices in structural equation modeling with categorical data. *Multivariate Behavioral Research, 56,* 1-18. https://doi/org/10.1080/00273171.2020.1717922

Singer, E. A. (1959). *Experience and reflection.* (C. W. Churman, Ed.). University of Pennsylvania Press.

Singla, A. (2018). *An introductory guide to maximum likelihood estimation (with a case study in R)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/07/introductory-guide-maximum-likelihood-estimation-case-study-r/#:~:text=MLE%20is%20the%20technique%20which,best%20describe%20the%20given%20data.&text=These%20values%20are%20a%20good,get%20more%20robust%20parameter%20estimates.

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). Routledge.

Schwalbe, C. S. (2009). Risk assessment stability A revalidation stuy of the aizona risk/needs assessment instrument. *Research on Social Work Practice, 19*(2), 205-213. https://doi.org/10.1177/1049731508317297

Shi, D., & Maydeu-Olivares, A. (2020). The effects of estimation methods on SEM fit indices. *Educational and Psychological Measurement, 80*(3), 421-445. https://doi.org/10.1177/0013164419885164

Shogren, K. A., Wehmeyer, M. L., Little, T. D., Forber-Pratt, A. J., Palmer, S. B., & Seo, H. (2017). Preliminary validity and reliability of scores on the self-determination inventory: Student report version. *Career Development and Transition for Exceptional Individuals, 40*(2), 92-103. https://doi.org/10.1177/2165143415594335

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. https://doi.org/10.1007/s11336-008-9101-0

Silver, E., Smith, W. R., & Banks, S. (2000). Constructing actuarial devices for predicting recidivism: A comparison of methods. *Criminal Justice and Behavior, 27*(6), 733-764. https://doi.org/10.1177/0093854800027006004

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.)., *The concept of validity: Revisions, new directions, and applications* (p. 9-37). IAP Information Age Publishing.

Sitlington, P. L., & Clark, G. M. (2007). The transition assessment process and IDEIA 2004. *Assessment for Effective Intervention, 32*(3). https://doi.org/10.1177/15345084070320030201

Sitlington, P. L., Neubert, D. A., & Leconte, P. J. (1997). Transition assessment: The position of the division on career development and transition. *Career Development for Exceptional Individuals*, *20*(1). https://doi.org/10.1177/088572889702000106

Skiba, R. J., Poloni-Staudinger, L., Simmons, A. B., Feggins-Azziz, L. R., & Chung, C.-G. (2005). Unproven links: Can poverty explain ethnic disproportionality in special education? *The Journal of Special Education*, *39*(3), 130–144.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356–443). American Council on Education.

Staples, M., Zhu, L., & Grundy, J. (2016). Continuous validation for data analytics systems. ICSE '16: Proceedings of the 38th International Conference on Software Engineering Companion. https://doi.org/10.1145/2889160

Statista. (2019). *Languages spoken (at home) other than English in the United States by number of speakers in 2019. https://www.statista.com/statistics/183483/ranking-of-languages-spoken-at-home-in-the-us-in-2008/*

Steenkamp, J. B. E. M., & Baumgartner, H. (1995). Development and cross-national validation of a short form of CSI as a measure of optimum stimulation Level. International Journal of Research in Marketing, *12*(2), 97-104. https://doi.org/10.1016/0167-8116(93)E0035-8

Stevens, S. L. (1983). *An investigation of the content validity, stability, and internal consistency of the Spanish version of the transition planning inventory home form* [Doctoral dissertation, The University of Kansas]. ProQuest Dissertations and Thesis.

Suk, A. L., Martin, J. E., McConnell, A. E., & Biles, T. L. (2020). States decreased their required secondary transition planning age: Federal policy must change. *Journal of Disability Policy Studies*, *31*(2), 112-118. https://doi.org/10.1177/1044207320915157

Sullivan, A. L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children*, *77*(3), 317-334. http://debdavis.pbworks.com/w/file/fetch/81120626/journal%202.pdf

Tabachnick, B. G., & Fidell, L. S. (2013). *Multivariate statistics* (6th ed.). Pearson.

Tanzer, N. K. (1995). Cross-cultural bias in Likert-type inventories: Perfect matching factor structures and still biased? European Journal of Psychological Assessment, 11, 194-201. https://doi.org/10.1027/1015-5759.11.3.194

Tavakol, M., & Dennick, R. (2011). Making sense of cronbach's alpha. *International Journal of Medicine Education, 2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Test, D. W., Aspel, N. P., & Everson, J. M. (2006). *Transition methods for youth with disabilities*. Pearson.

Test, D. W., Mazzotti, V. L., Mustain, A. L., Fowler, C. H., Kortering, L., & Kohler, P. (2009). Evidence-based secondary transition predictors for improving postschool outcomes for students with disabilities. *Career Development for Exceptional Individuals, 32*(3), 160-181. https://doi.org/10.1177/0885728809346960

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments.* University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/onlinepubs/synth44.pdf

Tiffin-Richards, S. P., & Anand Pant, H. (2017). Arguing validity in educational assessment. In Leutner, D., Fleischer, J., Grünkorn, J., & Klieme, E. (Ed.), *Competence assessment in education: Research, models, and instruments* (pp. 469-486). Springer.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220. https://doi.org/10.1037/a0023353

Trainor, A., Murray, A., & Kim, H-J. (2016). English learners with disabilities in high school: Population characteristics, transition programs, and postschool outcomes. *Remedial and Special Education*, *37*(3), 146-158. https://doi.org/10.1177/0741932515626797

Todd, J., Barron, D., Aspell, J. E., Lin Toh, E. K., Zahari, H. S., Mohd Khatib, N. A., & Swami, V. (2020). Translation and validation of a Bahasa Malaysia (Malay) version of the multidimensional assessment of interoceptive awareness (MAIA). *PLoS ONE*, *15*(4), 1-19. https://doi.org/10.1371/journal.pone.0231048

Tukey, J. (1977). *Exploratory data anlysis.* Addison-Wesley. http://www.ru.ac.bd/wp-content/uploads/sites/25/2019/03/102_05_01_Tukey-Exploratory-Data-Analysis-1977.pdf

United States & Mellette v. Jones, 615 F. 3d 544, (1996). https://casetext.com/analysis/case-summary-united-states-mellette-v-jones

U.S. Department of Education. (2002). *The President's commision on excellence in special education.* https://www2.ed.gov/about/offices/list/ocr/frontpage/faq/rr/policyguidance/Supple%20Fact%20Sheet%203.21.20%20FINAL.pdf

Valentini, N. C., & Ramlho, M. H., & Olivera, M. A. (2014). Movement assessment battery for children - 2: Translation, reliability, and validity for Brazilian chidlren. *Research in Developmental Disabilities, 35*(3), 733-740. https://doi.org/10.1016/j.ridd.2013.10.028

Van De Schoot, R., Schmidt, P., De Beuckeler, A., Lek, K., & Zondervan-Zwijenburg, M. (2015). Editorial: Measurment invariance. *Frontiers in Psychology, 6,* 1-4. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01064/full

Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis for cross-cultural research. Sage.

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 54*(2), 119-135. https://doi.org/10.1016/j.erap.2003.12.004

Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invraiance methods and procedures. *Organizational Research Methods, 5*(2), 139-158. https://doi.org/10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

Vivek, P., Singh, S. N., Mishra, S., & Donavan, D. T. (2017). *Parallel analysis engine to aid in determining number of factors to retain using R.* https://analytics.gonzaga.edu/parallelengine/

Wagner, M., Newman, L., Cameto, R., Levine, P., & Marder, C. (2007). *Perceptions and expectations of youth with disabilities: A special topic report of findings from the*

National Longitudinal Transition Study-2 (NLTS2). https://ies.ed.gov/ncser/pdf/20073006.pdf

Weaver, B. (2020). *Formal vs. informal assessments.* https://www.scholastic.com/teachers/articles/teaching-content/formal-vs-informal-assessments/

Widaman, K. (2014). Handbook of research methods in social and personality psychology (Vol. 2). Cambridge University Press.

Widaman, K.F. & Reise, S.P. (1997). Exploring the measurement equivalence of psychological instruments: Applications in the substance use domain. In K.J. Bryant, M. Windle, & S.G. West (Eds.), *The science of prevention* (pp. 281–324). American Psychological Association.

Will. M. (1984). *OSERS programming for the transition of youth with disabilities: bridges from school to working life*. U.S. Dept. of Education.

Workforce Innovation and Opportunities Act, 29 U.S. Code § 701 (2014). https://www.dol.gov/agencies/eta/wioa

Xi, X. (2009). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170. https://doi.org/10.1177/0265532209349465

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavioral Research Methods*, *51*, 409-428. https://doi.org/10.3758/s13428-018-1055-2

Yang, W., Lance, C. E., & Hui, H. C. (2006). Psychometric properties of the Chinese self-directed search (1994 edition). *Journal of Vocational Behavior*, *68*(1), 560-576. https://doi.org/10.1016/j.jvb.2005.12.003

Yell, M. L., Rogers, D., & Lodge Rogers, E. (2019). In Yell, M. L. (Ed.), *The law and special education* (5th ed., pp. 37-53). Pearson.

Yoo, H. J., Ahn, S. H., Eremenco, S., Kim, H., Kim, W. K., Kim, S. B., & Han, O. S. (2005). Korean translation and validation of the functional assessment of cancer therapy-breast (FACT-B) scale version 4. *Quality of Life Research*, *14*, 1627-1632. https://doi.org/10.1007/s11136-004-7712-1

Zacarian, D. (2011). *The over- and under-identification of ELLs in special education.* Colorin colorado. https://www.colorincolorado.org/article/over-and-under-identification-ells-special-education

Zeigler, K., & Camarota, S. A. (2018). Almost half speak a foreign language in america's largest cities. https://cis.org/Report/Almost-Half-Speak-Foreign-Language-Americas-Largest-Cities

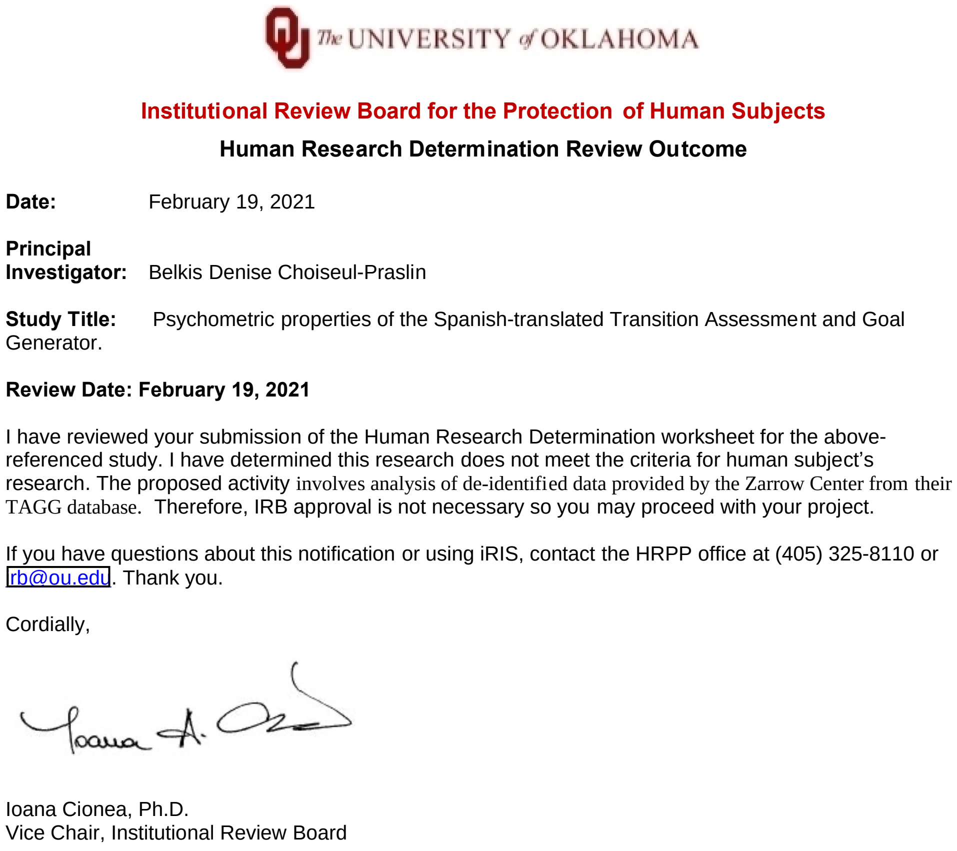
Zhang, D., & Benz, M. R. (2006). Enhancing self-determination of culturally diverse students with disabilities: Current status and future directions. *Focus on Exceptional Children, 38*(9), 3-12. https://doi.org/10.17161/foec.v38i9.6823

Zimmerman, A. (2019). *Why don't you learn English?: Lawsuit blasts NYC translation services for special education families*. Chalkbeat New York. https://ny.chalkbeat.org/2019/6/6/21108315/why-don-t-you-learn-english-lawsuit-blasts-nyc-translation-services-for-special-education-families

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*(2), 136-147. https://doi.org/10.1191/0265532203lt248oa

Appendix A

IRB Approval to Use Extant Data



The UNIVERSITY of OKLAHOMA

Institutional Review Board for the Protection of Human Subjects

Human Research Determination Review Outcome

Date: February 19, 2021

Principal Investigator: Belkis Denise Choiseul-Praslin

Study Title: Psychometric properties of the Spanish-translated Transition Assessment and Goal Generator.

Review Date: February 19, 2021

I have reviewed your submission of the Human Research Determination worksheet for the above-referenced study. I have determined this research does not meet the criteria for human subject's research. The proposed activity involves analysis of de-identified data provided by the Zarrow Center from their TAGG database. Therefore, IRB approval is not necessary so you may proceed with your project.

If you have questions about this notification or using iRIS, contact the HRPP office at (405) 325-8110 or rb@ou.edu. Thank you.

Cordially,

Ioana Cionea, Ph.D.
Vice Chair, Institutional Review Board

Appendix B

Syntax Used in R-Studio

1. English TAGG-S Initial CFA Code Used in R-Studio

```
o   Library(lavaan)

o   Library(REdaS)

o   Library(readr)

o   English <- read.csv("EnglishData.csv", header = TRUE, sep = ",")

o   str(English)

## CFA with ML estimation.

o   CFAEng<-'

    Strength=~1*item1+item2+item3+item4

    DisAware=~1*item5+item6+item7+item8

    Persistence=~1*item9+item10+item11+item12+item13

    Interacting=~1*item14+item15+item16

    Goals=~1*item17+item18+item19+item20+item21+item22

    Emply=~1*item23+item24+item25+item26

    StudentInv=~1*item27+item28+item29+item30

    SuppCom=~1*item31+item32+item33+item34'

o   Fit <- cfa(CFAEng, data = English)

o   Summary(fit, fit.measures = TRUE, stand = TRUE)


## CFA with DWLS estimation
```

```
o  Fit2 <- cfa(CFAEng, data = English, ordered =
   c("item1","item2","item3","item4","item5","item6",
   "item7","item8","item9","item10","item11","item12","item13","item14","item15"
   ,"item17","item18","item19","item20","item21","item22","item23","item24","ite
   m25","item26","item26","item27","item28","item29","item30","item31","item32"
   ,"item33","item34"))
o  Summary(fit2, fit.measures = TRUE, stand = TRUE)
o  MI <- modindices(fit2)
```

2. Spanish TAGG-S Initial CFA Code Used in R-Studio

```
o  Library(lavaan)
o  Library(REdaS)
o  Library(readr)
o  Spanish <- read.csv("SpanishData.csv", header = TRUE, sep = ",")
o  str(Spanish)
## CFA with ML estimation.
o  CFASpa <-'
   Strength=~1*item1+item2+item3+item4
   DisAware=~1*item5+item6+item7+item8
   Persistence=~1*item9+item10+item11+item12+item13
   Interacting=~1*item14+item15+item16
   Goals=~1*item17+item18+item19+item20+item21+item22
   Emply=~1*item23+item24+item25+item26
   StudentInv=~1*item27+item28+item29+item30
```

```
     SuppCom=~1*item31+item32+item33+item34'
o    Fit <- cfa(CFASpa, data = Spanish)
o    Summary(fit, fit.measures = TRUE, stand = TRUE)
## CFA with DWLS estimation
o    Fit2 <- cfa(CFASpa, data = English, ordered =
     c("item1","item2","item3","item4","item5","item6",
     "item7","item8","item9","item10","item11","item12","item13","item14","item15"
     ,"item17","item18","item19","item20","item21","item22","item23","item24","ite
     m25","item26","item26","item27","item28","item29","item30","item31","item32"
     ,"item33","item34"))
o    Summary(fit2, fit.measures = TRUE, stand = TRUE)
o    MI <- modindices(fit2)
```

3. Exploratory (unrotated) EFA Code Used in R-Studio

```
o    Library(psych)
o    Library(REdaS)
o    Spanish <- read.csv("SpanishData.csv", header = TRUE, sep = ",")
o    str(Spanish)
o    attach(Spanish)
o    bart_sphere(Spanish)
o    KMO(Spanish)
o    fa(Spanish, nfactors = 34,rotate = "oblimin")
o    ExploreEFA <- fa(Spanish, nfactors = 8, rotate = "oblimin")
o    fa.diagram(ExploreEFA, main = "Spanish")
```

4. Six-, four-, and three-factor EFA Code Used in R-Studio

- Library(psych)
- Library(REdaS)
- Library(readr)
- Library(knitr)
- Library(GPArotation)
- Spanish <- read.csv("SpanishData.csv", header = TRUE, sep = ",")
- str(Spanish)
- nrow(Spanish)
- attach(Spanish)
- bart_spher(Spanish)
- KMO(Spanish)
- fa.parallel(Spanish, fm = "ml", fa = "fa")

Six-factor EFA code

- fa(Spanish, nfactors = 6, rotate = "oblimin", fm = "ml")

  ##Used .40 as the factor loading cut off, eliminated items listed below

- fa(Spanish[, -c(1,2,3,4,11,12,13,14,15,16,17,22,25,26,34)], nfactors = 6, rotate = "oblimin", fm = "ml")

  ##Equation for CFI

- sixfactor = fa(Spanish[, -c(1,2,3,4,11,12,13,14,15,16,17,22,25,26,34)], nfactors = 6, rotate = "oblimin", fm = "ml") 1 - ((sixfactor$STATISTIC - sixfactor$dof)/(sixfactor$null.chisq - sixfactor$null.dof))

##Six factor reliability; raw alpha is the reliability coefficient

-
  ```
  factor1 = c(18,19,20,21)
  factor2 = c(27,28,29,30)
  factor3 = c(23,24)
  factor4 = c(31,32,33)
  factor5 = c(5,6,7)
  factor6 = c(9,10)
  ```

-
  ```
  psych::alpha(Spanish[, factor1])
  psych::alpha(Spanish[, factor2])
  psych::alpha(Spanish[, factor3])
  psych::alpha(Spanish[, factor4])
  psych::alpha(Spanish[, factor5])
  psych::alpha(Spanish[, factor6])
  ```

Four-factor EFA code

-
  ```
  fa(Spanish, nfactors = 4, rotate = "oblimin", fm = "ml")
  ```

##Used .40 as the factor loading cut off, eliminated items listed below

-
  ```
  fa(Spanish[, -c(7,8,14,15,16,22,25,26,32,33,34)], nfactors = 4, rotate = "oblimin", fm = "ml")
  ```

  ##Equation for CFI

-
  ```
  fourfactor = fa(Spanish[, -c(7,8,14,15,16,22,25,26,32,33,34)], nfactors = 4, rotate = "oblimin", fm = "ml") 1 - ((fourfactor$STATISTIC - fourfactor$dof)/(fourfactor$null.chisq - fourfactor$null.dof))
  ```

```
     ##Four factor reliability; raw alpha is the reliability coefficient
  o  factor1 = c(11,12,13,17,18,19,20,21)
     factor2 = c(1,2,3,4,5,6,9,10)
     factor3 = c(27,28,29,30)
     factor4 = c(23,24)
  o  psych::alpha(Spanish[, factor1])
     psych::alpha(Spanish[, factor2])
     psych::alpha(Spanish[, factor3])
     psych::alpha(Spanish[, factor4])
 ## Three-factor EFA code
  o  fa(Spanish, nfactors = 3, rotate = "oblimin", fm = "ml")
 ## Used .40 as the factor loading cut off, eliminated items listed below
  o  fa(Spanish[, -c(4,5,6,7,8,9,10,15,22,25,26,34)], nfactors = 3, rotate = "oblimin",
     fm = "ml")
## Equation for CFI
  o  threefactor = fa(Spanish[, -c(4,5,6,7,8,9,10,15,22,25,26,34)], nfactors = 3, rotate =
     "oblimin", fm = "ml") 1 - ((fourfactor$STATISTIC -
     fourfactor$dof)/(fourfactor$null.chisq - fourfactor$null.dof))
## Three factor reliability; raw alpha is the reliability coefficient
  o  factor1 = c(1,2,3,14,16,23,24,31,32,33)
     factor2 = c(11,12,13,17,18,19,20,21)
     factor3 = c(27,28,29,30)
  o  psych::alpha(Spanish[, factor1])
```

psych::alpha(Spanish[, factor2])

psych::alpha(Spanish[, factor3])

psych::alpha(Spanish[, factor4])

5. Post Hoc CFA Code Used in R-Studio

- Library(lavaan)
- Library(REdaS)
- Library(readr)
- Spanish <- read.csv("SpanishData.csv", header = TRUE, sep = ",")
- str(Spanish)

Six-factor model identified by R-Studio

- SixR <-'

  Factorone=~1*item18+item19+item20+item21

  FactorTwo=~1*item27+item28+item29+item30

  FactorThree=~1*item23+item24

  FactorFour=~1*item31+item32+item33

  FactorFive=~1*item5+item6+item7

  FactorSix=~1*item9+item10'
- Fit <- cfa(SixR, data = Spanish, ordered = c("item5","item6","item7","item9","item10","item18","item19", "item20","item21","item23","item24","item27","item28","item29","item30","item31","item32","item33"))
- Summary(fit, fit.measures = TRUE, stand = TRUE)

```
## Four-factor model identified by R-Studio
o   FourR <-'
    FactorOne=~1*item11+item12+item13+item17+item18+item19+item20+item21
    FactorTwo=~1*item1+item2+item3+item4+item5+item6+item9+item10
    FactorThree=~1*item27+item28+item29+item30
    FactorFour=~1*item23+item24'
o   Fit <- cfa(FourR, data = Spanish, ordered =
    c("item11","item12","item13","item17","item18","item19","item20",
    "item21","item1","item2","item3","item4","item5","item6","item9","item10","ite
    m27","item28","item29","item30","item23","item24"))
o   Summary(fit, fit.measures = TRUE, stand = TRUE)
## Six-factor model identified by FACTOR
o   SixF <-'
    FactorOne=~1*
    item1+item2+item14+item16+item23+item24+item31+item32+item33
    FactorTwo=~1* item25+item26
    FactorThree=~1* item27+item28+item29+item30
    FactorFour=~1* item11+item12+item17+item18+item19+item20+item21
    FactorFive=~1* item3+item5+item6+item7+item8
    FactorSix=~1* item9+item10+item13'
o   Fit <- cfa(SixF, data = Spanish, ordered =
    c("item1","item2","item14","item16","item23","item24","item31",
    "item32","item33","item25","item26","item27","item28","item29","item30","item
```

```
   11","item12","item17"+"item18"+"item19"+"item20"+"item21"+"item3"+"item5
   "+"item6"+"item7"+"item8"+"item9"+"item10"+"item13"))
o  Summary(fit, fit.measures = TRUE, stand = TRUE)
## Four-factor model identified by FACTOR
o  FourF <-'
   FactorOne=~1* item3+item4+item5+item6+item7+item8+item9+item10
   FactorTwo=~1* item14+item16+item23+item24+item31+item32+item33
   FactorThree=~1* item27+item28+item29+item30
   FactorFour=~1* item11+item13+item17+item18+item19+item20+item21'
o  Fit <- cfa(FourR, data = Spanish, ordered =
   c("item3","item4","item5","item6","item7","item8","item9","item10","item11","it
   em13","item14","item16","item17","item18","item19","item20","item21","item23
   ","item24","item27","item28","item29","item30","item31","item32","item33"))
o  Summary(fit, fit.measures = TRUE, stand = TRUE)
## Three-factor model identified by FACTOR
o  threeF <-'
   FactorOne=~1*item1+item2+item3+item14+item16+item23+item24+item31+ite
   m32+item33
   FactorTwo=~1*item11+item12+item13+item17+item18+item19+item20+item21
   FactorThree=~1*item27+item28+item29+item30
o  fit <- cfa(threeF, data = threeCFA, ordered =
   c("item1","item2","item3","item14","item16","item23","item24","item31","item3
```

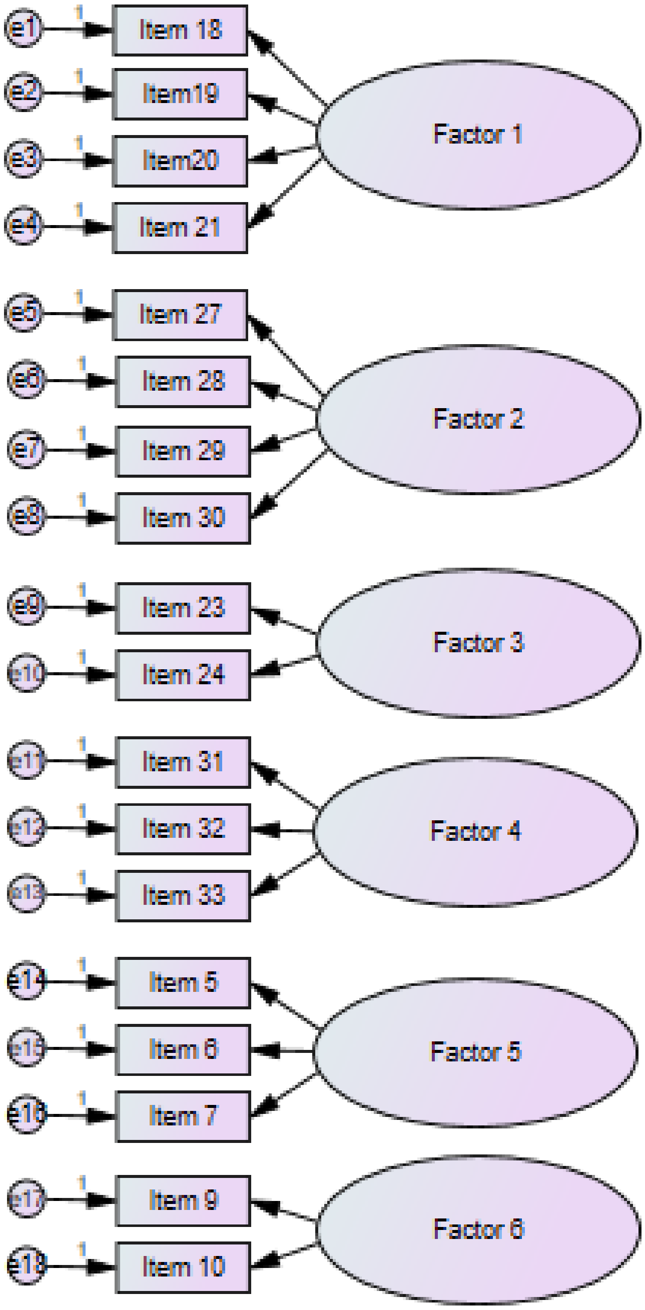
2","item33","item11","item12","item13","item17","item18","item19","item20","item21","item27","item28","item29","item30"))

- summary(fit, fit.measures = T, stand = T)

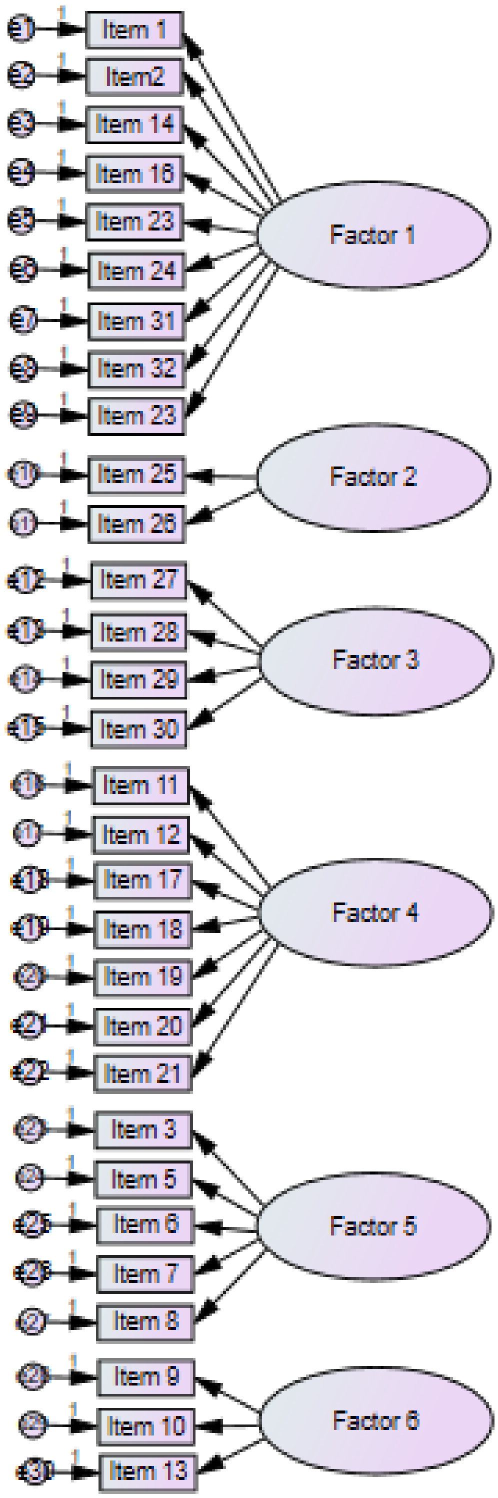
Appendix C

Model images (drawn on SPSS-AMOS)

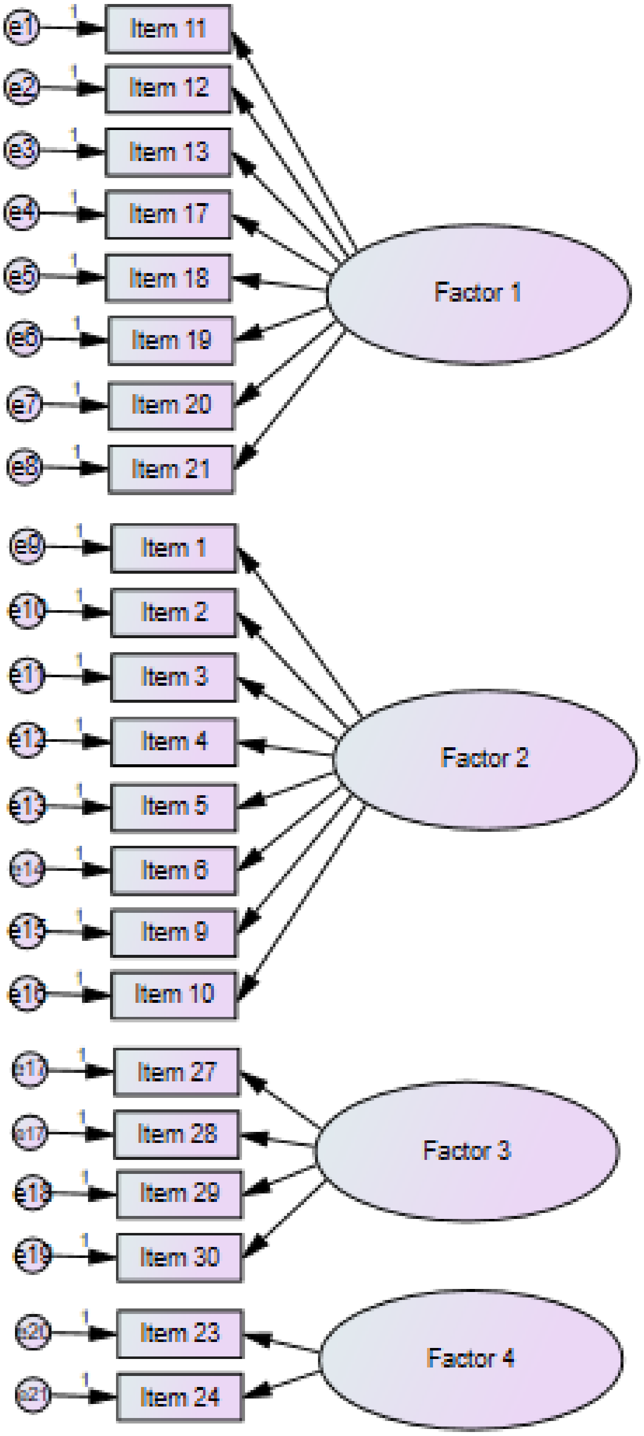
R: Six-factor model from post hoc CFA



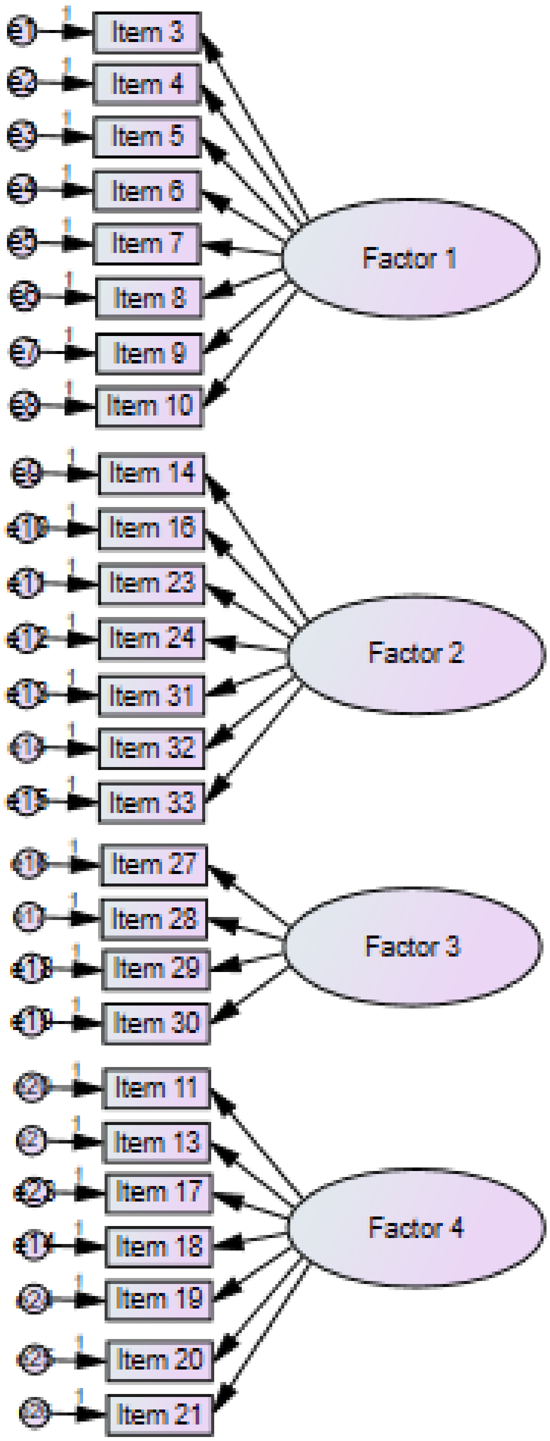
FACTOR: Six-factor model from post hoc CFA



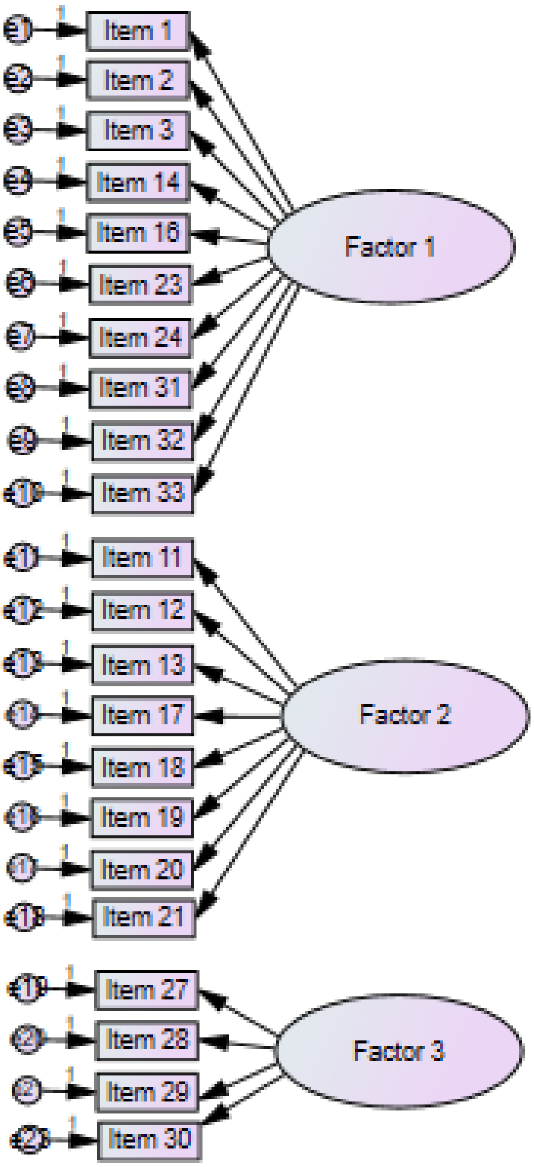
R: Four-factor model from post hoc CFA



FACTOR: Four-factor model from post hoc CFA



R-Studio: Three-factor model from post hoc CFA


e1
e2
e3
e4
e5
e6
e7
e8
e9
Item 1
Item 2
Item 3
Item 14
Item 16
Item 23
Item 24
Item 31
Item 32
Item 33
Factor 1
Item 11
Item 12
Item 13
Item 17
Item 18
Item 19
Item 20
Item 21
Factor 2
Item 27
Item 28
Item 29
Item 30
Factor 3