UNIVERSITY OF CALIFORNIA
RIVERSIDE

Psychometric Evaluation of the Listening Sentence Span Task: A Working Memory Measure for English Language Learners

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Arts

in

Education

by

Joseph A. Rios

June 2011

Thesis Committee:
Dr. H. Lee Swanson, Chairperson
Dr. Marsha Ing
Dr. George Marcoulides

The Thesis of Joseph A. Rios is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE THESIS

Psychometric Evaluation of the Listening Sentence Span Task:
A Working Memory Measure for English Language Learners

by

Joseph A. Rios

Master of Arts, Graduate Program in Education
University of California, Riverside, June 2011
Dr. H. Lee Swanson, Chairperson

The Listening Sentence Span Task is a widely used measure of working memory capacity for children. However, this measure has not been analyzed from an IRT framework nor has it been adapted to non-English languages. Study 1 of this paper examined the Classical Test Theory summed-score statistics, construct equivalence via a structural equation modeling framework, item parameter estimation utilizing Item Response Theory, and concurrent validity of a newly adapted Spanish-version of the Listening Sentence Span Task (LSST-S) for 491 English language learners (ELLs) in grades 1-3. Results of the analysis demonstrated that the majority of items on the measure displayed low item-total correlations and low internal consistency reliability. In addition, a very low coefficient α was obtained for the overall measure. A confirmatory item factor analysis demonstrated that the LSST-S measured a distinct latent construct when compared to its English predecessor, implying that construct non-equivalence was present between the two measures. Lastly, the LSST-S exhibited poor concurrent validity with measures of reading comprehension, fluid intelligence, and arithmetic computation. Study 2 examined

differential item functioning of the Listening Sentence Span Task English-version in a mixed language-status sample, which was comprised of ELL (n=491) and non-ELL (n=315) children. This analysis demonstrated that uniform and non-uniform DIF was present. Recommendations for improving the LSST-S and LSST-E for use with ELLs are provided.

Table of Contents

List of Tables

List of Figures

Psychometric Evaluation of the Listening Sentence Span Task-Spanish Version: A Working Memory Measure for English Language Learners

The study of individual differences in working memory has gained prominence in the fields of cognitive psychology and cognitive neuroscience over the past 25 years. Since the seminal article published by Baddeley & Hitch (1974), numerous theoretical advances have occurred, resulting in multiple models of working memory (Shah & Miyake, 1999). Although multiple models exist, working memory can be generally conceptualized as a limited capacity processing system that is necessary for the simultaneous storage and manipulation of information (Baddeley, 1992; Baddeley & Logie, 1999). The practical applications of examining individual differences in working memory lie in the assumption that it supports learning throughout the lifespan (Gathercole, 2007). This assumption is supported by consistent findings across various age groups that demonstrate a relationship between working memory and performance on tasks, such as vocabulary acquisition (Atkins & Baddeley, 1998; Gathercole & Baddeley, 1990; Gathercole & Baddeley, 1993; Masoura, Gathercole, & Bablekou, 2004), counting (Gathercole, Durling, Evans, Jeffcock, & Stone, 2008; Noël, 2009), arithmetic computation (Barrouillet, Lépine, & Camos, 2008), word-problem solving (Swanson, Jerman, & Zheng, 2008), and reading comprehension (Daneman & Carpenter, 1980; Swanson, Kehler, & Jeman, 2010). In addition, numerous studies have found a strong relationship between working memory and fluid intelligence (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Laughlin, Tuholski, & Conway, 1999; Kane, Hambrick, Tuholski, Wilhelm, Payne, & Engle, 2004; Kane, Hambrick, & Conway,

2005; Friedman, Miyake, Corley, Young, DeFries, & Hewitt, 2006; Oberauer, Schulze, Oliver, & Süß, 2005; Shelton, Elliott, Matthews, Hill, & Gouvier, 2010), which has led some researchers to suggest that working memory is a more powerful predictor of academic success than IQ (Alloway & Alloway, 2010). Therefore, the accurate measurement of working memory may be an effective approach towards understanding cognitive abilities, as well as predicting academic achievement (Vock & Holling, 2008).

Researchers in the field of cognitive psychology primarily study individual differences in working memory by utilizing complex span tasks to assess working memory capacity (WMC; Engle, 2010). Complex span tasks require participants to simultaneously store and process other, potentially distracting, information. Although these tasks rely on some speech-based and visuo-spatial-based coding, they primarily assess executive processing, which is responsible for control and regulation of cognitive processes (Baddeley & Logie, 1999). In general, these WMC assessments have been shown to demonstrate construct validity and acceptable ranges of reliability (ranging from .7-.9) throughout multiple studies (Engle & Kane, 2004). Many of the working memory span tasks used today are based off of the seminal article published by Daneman and Carpenter (1980) who first introduced the *Reading/Listening Span Task.*

The Reading/Listening Span Task has been used in multiple studies and has been adapted to study various populations (e.g., brain-damaged patients; Tompkins, Bloise, Timko, & Baumgaertner, 1994), as it has long been considered the classic instrument for assessing working memory (Elosúa, Carriedo, & García-Madruga, 2009); however, very few adaptations/translations have been developed for use with non-English speaking

populations.[1] This is surprising as the number of general test adaptations has increased greatly over the past few decades (Casillas & Robbins, 2005). The few non-English adaptations in the literature were developed for adults in French (Desmette, Hupet, Schelstraete, & Van der Linden, 1995), Spanish (Gutiérrez, Jiménez, & Castillo, 1996), and Japanese (Kondo & Osaka, 2000). Given the critical role that this measure could play in assessing cognitive ability and predicting academic achievement, an adaptation for use with children who are language minorities in the educational context is necessary.

Currently, very few children's non-English working memory measures exist. However, the need for such a measure is evident in the fact that the number of non-native English speakers in the U.S. public education system has increased 53% from 1998-2008 to include a total of 5.3 million students (U.S. Department of Education, 2010). Estimates suggest that the number of English language learners (ELLs) in the U.S. will continue to increase and will eventually comprise 25% of the total student population by the year 2025 (Educational Testing Service, 2009). Therefore, adapting a WMC measure will allow for fairness in assessment as individuals can be tested in their first language (AERA, APA, & NCME, 1999), as well as facilitate comparative studies across populations (Van de Vijver & Hambleton, 1996). In order to fill the need for a children's non-English working memory assessment, a Spanish-version of the *Listening Sentence Span Task* (Swanson, 1992; Swanson, 1996a; Swanson, 1999) was developed and piloted.[2]

The Listening Sentence Span Task is one of the most extensively used instruments to assess WMC in children. This measure is an adaptation of Daneman and

Carpenter's Listening Span Task; however, it differs in two ways. First, this version was designed specifically for use with children by controlling for sentence complexity (mean sentence-reading level is approximately 3.8), word frequency, and imagery (Swanson, 1992; Swanson & Beebe-Frankenberger, 2004; Swanson, 2008). Secondly, it requires an accurate response to the process question, whereas the original measure does not require an accurate response. Besides these distinctions the format between the original measure and the adaptation are the same. Previous research on the Listening Sentence Span Task has reported an equivalent-form reliability of .95 (Swanson, 1992), as well as an acceptable range of internal reliability (.77-.95; Swanson, 1992; Swanson, 1996b; Swanson & Beebe-Frankenberger, 2004; Swanson, 2008). In addition, convergent validity for this measure has been demonstrated through strong correlations with 11 different working memory measures (Swanson, 1992). As previous research demonstrates, the English-version of the Listening Sentence Span Task is a psychometrically sound and valid measure, which makes it acceptable for adaptation.

Study 1

The present study describes the translation and piloting of the Listening Sentence Span Task for bilingual students in grades 1-3. Objectives of the study were to analyze the psychometric properties of both the English- and Spanish-versions with a bilingual sample, and to determine whether the two versions were structurally equivalent. More specifically, the analysis was concerned with the following: a) evaluation of Classical Test Theory summed-score statistics for both the LSST-E and LSST-S; b) assessment of the structural equivalence for both the English- and Spanish-versions; c) model parameter

estimation for both versions via an Item Response Theory (IRT) framework; and d) assessment of the concurrent validity for the Listening Sentence Span Task (both English and Spanish) with measures of reading comprehension, general fluid intelligence, and arithmetic computation.

Method

Data Source

Participants in the study were first identified by the school district to be English language learners (ELLs). Once identified, the project director sent consent forms to the parents of the ELL students. If the parents provided written informed consent as approved by the respective institutional review board, the students were then included in the study. The sample consisted of 500 bilingual students from two public school districts in southern California; however, 9 participants dropped out of the study before being administered the LSST-E and LSST-S, resulting in the inclusion of 491 students in the final analysis. At the time of data collection all students were in grades 1, 2, and 3. The socioeconomic status of the sample ranged from low to lower-middle class based on parental income level. Descriptive statistics of the sample's demographic characteristics are provided in Table 1.

English and Spanish language proficiency of the sample was determined by two measures that assessed both receptive and expressive language proficiency. The Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn & Dunn, 2007) was administered to assess receptive language proficiency, while expressive language proficiency was determined by the Expressive One-Word Picture Vocabulary Test,

Spanish-Bilingual Edition (EOWPVT-SBE; Brownell, 2001). Descriptive statistics of the sample's language proficiency are provided in Table 2.

Measures

Listening Sentence Span Task (LSST-E). The LSST-E consists of four levels of unrelated declarative sentences, 7-10 words in length, which are read to the participant with a 5-second pause to indicate the end of each sentence. Each level is comprised into random sets of two, three, four, or five sentences. Once all of the sentences in the level are read, the participant is required to correctly recall the last word of each sentence after correctly answering a process question about one of the sentences. The process question is placed into the measure in order to ensure that participants do not merely try to remember the target word or treat the task as one of short-term memory. If the process question is answered incorrectly, the assessment is discontinued. However, if the participant correctly answers the process question and recalls all items within the level, s/he goes onto the preceding level, where the sentences gradually increase. Raw scores were calculated by allotting one point for every process question answered correctly and one point for every item recalled in the same order presented. A raw score of zero was given to any participant who missed the process question for the first level, irrespective of whether any of the last words within the level were recalled correctly. Working memory capacity was defined as the highest set of correctly recalled words in which the process question was recalled correctly. The possible number of words to recall ranged from 0 to 8 (See Appendix A). The mean and standard deviation of the LSST-E raw score is presented in Table 3.

Translation of Spanish Version (LSST-S). The English-version of the Listening Sentence Span was translated into Spanish by a fully bilingual graduate researcher. This translation was then reviewed and edited by a language and content consultant whose first language was Spanish and who specialized in working memory research. These changes were presented to an additional language expert to ascertain whether the translation was linguistically similar to the English-version. From these reviews a final version was produced to be piloted with a bilingual sample. All individuals involved in the translation process were both fluent in the target and primary language of the Listening Sentence Span Task. In addition, all translators were knowledgeable of the intended population's culture as they themselves were members of the same cultural group (See Appendix B). The mean and standard deviation of the LSST-S raw score is presented in Table 3.

Fluid Intelligence. Raven's *Coloured Progressive Matrices* (CPM; Raven, 1956) was used to assess the concurrent validity of the LSST-E and LSST-S with nonverbal or fluid intelligence as it has long been considered the gold standard measure (Engle, 2010). The CPM is comprised of 36 items, which are represented as patterns. Each pattern is missing a section, requiring the participant to choose the correct replacement piece from six different options. Items progressively increase in difficulty as patterns become more intricate. The dependent variable for this measure is the number of correct responses, which ranges from 0 to 36. An internal consistency reliability of .80 to .90 was reported in the technical manual (Raven, Court, & Raven, 1990). An alpha coefficient of .87 was obtained for the sample included in this study.

Reading Comprehension. The *Passage Comprehension/Comprensión de textos* subtest of the *Woodcock-Muñoz Language Survey-Revised (WMLS-R)* was used as an indicator of reading comprehension (Woodcock, Muñoz-Sandoval, Ruef, & Alverado, 2005a; Woodcock, Muñoz-Sandoval, Ruef, & Alverado, 2005b). This subtest involves matching a pictographic representation of a word with an actual picture of the object. Early items require the participant to choose the picture represented by a phrase from multiple choices. The remaining items require the participant to identify a missing key word that makes sense in the context of a story that has been read. The items become more difficult as the passages increase in length and level of vocabulary. In an attempt to shorten testing time basals and ceilings were cut from 5 to 3. Besides this change, all other test administration procedures were the same as described in the manual. The Comprehensive Manual for the WMLS-R reports a median reliability of .82 for ages 5 to 19 years. The median standard error of measurement for the standard scores is 5.95 (Alvarado, Ruef, & Schrank, 2005). English Passage Comprehension obtained an internal consistency reliability of .90 for the sample included in this study, while the Spanish-version obtained an alpha coefficient of .89.

Arithmetic Computation. The written arithmetic subtest of the *Wide Range Achievement Test-3rd Edition* (WRAT-A; Wilkinson, 1993) was used to assess arithmetic computation. This subtest consists of 40 computational items administered to participants in a group format for 15 minutes. Each item is worth 1 point, total scores range from 0-40. An internal consistency coefficient ranging from .82-.95 and an alternate forms

reliability of .89 were reported for the WRAT-A. An alpha coefficient of .87 was obtained for the sample included in this study.

Procedure

Before data collection began, instructional meetings provided 7 bilingual graduate researchers with training in test administration. The English- and Spanish-versions of the Listening Sentence Span Task were part of a battery consisting of 45 assessments which were administered individually to each student as part of a larger project.[3] A single-group design was chosen, requiring each participant to be evaluated in both English and Spanish within the same test administration. As a result, testing time was divided up between two time points consisting of English and Spanish test administration with both lasting approximately 45 minutes to 1 hour. Test administration took place in a quiet room conducive to testing (empty classroom) on each school campus. For each student both test order (6 different combinations) and language order were counterbalanced. Inter-rater reliability was assessed by allowing a second test administrator to observe and individually score the test battery for 6 participants (1.27% of total sample). Data were collected from September 2009-June 2010.

Data Analysis

Item-level means, standard deviations, item-total correlations, and coefficient alphas were calculated for each item in IRTPRO 2, Beta version (Cai, du Toit, & Thissen, forthcoming). An overall internal consistency reliability statistic was also computed for both the LSST-E and LSST-S. Item-total correlations below .30 were considered to

reflect a possible problematic item, while an internal consistency coefficient above .70 was considered to reflect adequate reliability (Nunnally & Bernstein, 1994).

Following the analysis of the item-level descriptive statistics, structural equivalence between the LSST-E and LSST-S was examined via a categorical item factor analysis from a structural equation modeling (SEM) framework in Mplus, version 5 (Muthén & Muthén, 2007). This analysis was conducted as test developers/publishers must demonstrate that an adapted measure assesses the same construct as its predecessor in order to ensure the adequacy of using the adapted measure with an unintended population (International Test Commission, 2010). As a result, the following four models were tested: 1) a unidimensional model for the LSST-E items; 2) a unidimensional model for the LSST-S items; 3) a unidimensional model combining the items from both the LSST-E and LSST-S; and 4) a simple structure model where the items from the LSST-E were conceptualized to measure one latent construct, while the items from the LSST-S were modeled to measure a second distinctive factor. This analysis examined an exploratory analysis using confirmatory factor analytic models as the lack of variability in item responses did not allow the traditional two-step process of examining dimensionality. That is, if sample size and item response variability permit, a dataset is traditionally subset to conduct an exploratory analysis, which can then be validated via a confirmatory analysis; however, this was not possible in this study as few participants correctly endorsed items 4-7, not allowing for one to subset the dataset as the minimal variability in those items would be reduced in half for each dataset. Applying categorical data to common linear factor models has been demonstrated to result in biased parameter

estimates (DiStefano, 2002); however, use of modified weighted least squares estimators for categorical data have been suggested to be adequate methods for conducting item factor analysis from a structural equation modeling framework (Wirth & Edwards, 2007). As a result, the weighted least squares with mean and variance adjustment (WLSMV) estimator was used in this analysis. To determine the fit of the confirmatory models the chi-square *p*-value, comparative fit index (CFI), Tucker-Lewis Index (TLI), and root mean square error of approximation (RMSEA) were assessed. More specifically, cut-off levels to indicate good model fit were >.05 for the chi-square *p*-value, >.95 for both the CFI and TLI, and <.06 for the RMSEA estimate (Curran, Bollen, Chen, Paxton, & Kirby, 2003; Hu & Bentler, 1999). Although these fit indices were originally suggested for use with continuous variables, they have also been found to be accurate with categorical variables (Yu & Muthén, 2001).

Based on the results of the dimensionality analysis, Item Response Theory (IRT) models were conducted to estimate item parameters in IRTPRO 2, Beta Version, utilizing the Bock-Aitkin method and the cross-product approximation standard error algorithm. One challenge presented with the Listening Sentence Span Task was that the process question for each level shared content with one of the recall items. As a result, this shared content would conceptually lead to local dependence, which would violate the IRT assumption of local independence. To address this issue, the strategy first introduced by Thissen, Steinberg, and Mooney (1989) was adopted, whereby items that shared similar content were collapsed into a single polytomous item. As a result, mixed-item format unidimensional models were developed, whereby dichotomous items within each level

that were conceptually locally dependent (items 1 and 3; items 4 and 6) were collapsed into a single polytomous item with three possible thresholds (zero correct, one correct, and both items endorsed correctly). The remaining items were left dichotomous. Two competing models were tested to determine the level of constraints on the model that best fit the data. That is, the first model analyzed the dichotomous items with a 2-parameter logistic (2-PL) model, while the Nominal Response Model (NRM) was used to analyze the polytomous items. The second model utilized the Rasch and Graded Response Models to analyze the dichotomous and polytomous items, respectively. A 3-parameter logistic (3-PL) model was not estimated for the dichotomous items as previous research has demonstrated that the guessing parameter within the 3-PL model is not appropriate for working memory measures (Vock and Holling, 2008). Model fit for each IRT model was assessed by examining the sample size free RMSEA estimate associated with the M_2 statistic. The M_2 statistic is a limited-information test that provides improved model fit estimation relative to Pearson's χ^2 and the likelihood ratio statistic G^2 when sparseness is present in 2^n contingency tables, which are tables comprised of *n* items with two possible responses (See Maydeu-Olivares & Joe, 2005; Maydeu-Olivares & Joe, 2006). An RMSEA value at or below .06 was determined a-priori to indicate adequate fit to the sample data (Browne & Cuddeck, 1993; Hu and Bentler, 1999). Local dependence (LD) violations were examined via the standardized LD χ^2 statistic proposed by Chen and Thissen (1997).

Lastly, concurrent validity was examined by correlating the raw scores on the Listening Sentence Span Task (English- and Spanish-versions) with raw scores on

measures of reading comprehension, fluid intelligence, and arithmetic computation. This was accomplished as working memory capacity has been suggested to be strongly related to these constructs (Baddeley, 1992). Both attenuated and disattenuated correlations were calculated using SAS, version 9.2 (SAS Institute, Inc, 2008). Particular attention was given to the correlation between the LSST-E and LSST-S as a criterion for the defensibility of linking the LSST-E and LSST-S scales (Creswell, 2010). All correlations were examined at α=.05.

Results

Item-Level Summed Score Statistics

Item-level descriptive statistics and frequencies were calculated for all items on the LSST-E and LSST-S. Results of this initial analysis demonstrated that participants reached level 3 (items 8-13) of the LSST-E, while level 3 for the LSST-S was not administered, as the highest level reached was level 2 (items 4-7) (See Table 4 & Table 5). Examination of the frequencies demonstrated that items 8-13 on the LSST-E received less than 1% of correct responses from the sample. As a result of the low correct endorsement percentage and for ease of examining structural equivalence between the LSST-E and LSST-S scales, items (8-13) on the LSST-E were dropped from all analyses (See Table 6), resulting in only items 1-7 on both versions being included in further analyses.

LSST-E Item-Level Descriptives. Examination of the item-level descriptive statistics revealed that the percentage of correct responses was dramatically greater for level 1 (items 1-3) when compared to level 2 (item 4-7). This was expected for two

reasons: 1) the measure was discontinued for participants who did not correctly answer all questions on the first level; and 2) the number of sentences increased for each level requiring greater working memory capacity. Item-total correlations revealed that items 2-7 were all above the acceptable level of .30 (ranged from .398-.663), while item 1 had an item-total correlation of .261. This result suggests that item 1 may be a problematic item that requires further examination. Internal consistency reliability was acceptable for items 1 (α=.713), 4 (α=.713), and 7 (α=.713), while levels below the a-priori cut-off level of .70 were obtained for items 2 (α=.695), 3 (α=.669), 5 (α=.661), and 6 (α=.676). However, an acceptable overall reliability coefficient for the LSST-E was demonstrated, α=.732.

LSST-S Item-Level Descriptives. The LSST-S item-level descriptive statistics revealed that fewer participants correctly responded to each item on the measure when compared to their results on the LSST-E. In addition, the frequencies demonstrated that items 4 (.61%) and 7 (.41%) received less than 1% of correct responses, while the remaining items in level 2 also received low endorsement rates. Items 2, 3, 5 and 6 displayed acceptable item-total correlations, while the item correlations for items 1, 4, and 7 were all below .30, suggesting that they may be problematic items (See Table 5). Item-level internal consistency reliabilities for the LSST-S were all below .60, and the overall reliability coefficient for the measure was very low, α=.483. The results of the item-level descriptive statistics for the LSST-S revealed that the measure contains problematic items and low reliability.

Structural Equivalence

One-Factor Models. Unidimensional models for the LSST-E and LSST-S were conducted separately to determine whether each measure assessed one latent construct, which was conceptualized to be working memory. Results of the unidimensional analysis for the LSST-E demonstrated adequate fit to the sample data, χ^2=34.791, p<.001, CFI=.991, TLI=.988, RMSEA=.083 (See Table 7). Although the RMSEA statistic was slightly larger than the a-priori cutoff-level, very high CFI and TLI estimates provided justification for a one-factor model. The next analysis tested was the unidimensional model for the LSST-S. Results for this model provided evidence to support a one-factor pattern for the LSST-S, χ^2=4.559, p=.472, CFI=1.00, TLI=1.00, RMSEA<.001 (See Table 7). As evidence was obtained to confirm that both the LSST-E and LSST-S each measured one latent construct, two competing models were next examined. The first competing model tested looked to determine whether structural equivalence was apparent for the items from both the LSST-E and LSST-S. That is, no restrictions were placed on invariant factor loadings or invariant intercepts, instead this analysis looked to uncover whether the items from the LSST-E and LSST-S measured the same latent construct (See Figure 1). Results of this analysis demonstrated that the model did not provide adequate fit to the sample data, χ^2=135.32, p<.001, CFI=.920, TLI=.915, RMSEA=.123 (See Table 7). This result revealed that construct equivalence was not present between the LSST-E and LSST-S. As a result, the next model tested was a two-factor structure.

Two-Factor Model. The simple "Thurstonian" two-factor structure proposed that the LSST-E and LSST-S measured two distinctive latent constructs. As a result, the

LSST-E items were modeled as indicators of one latent construct, while the LSST-S items were purported to be indicators of the second latent construct (See Figure 2). Results of the confirmatory factor analysis demonstrated that the model provided adequate fit to the sample data, χ^2=45.154, p<.001, CFI=.982, TLI=.983, RMSEA=.055 (See Table 7). Examination of the factor loadings demonstrated strong relationships among all LSST-E items with factor one, while strong relationships were also demonstrated between the second factor and the LSST-S items, except item 1 (See Table 8). Item 1 on the LSST-S had a weak relationship with its respective factor (λ=.353), which may suggest that further content analyses need to be conducted to examine possible issues with wording. The covariance between the two latent constructs was weak, Φ=.365.

Item Parameter Estimation

As the confirmatory factor analysis demonstrated that the LSST-E and LSST-S each measured a separate latent construct, two competing mixed item format unidimensional models were estimated for both versions. More specifically, the first model (Model 1) tested utilized the Nominal Response Model to estimate the collapsed polytomous items (items 1 and 2; items 4 and 6) to examine whether the categories were ordered, while a 2-parameter logistic model was tested for the remaining dichotomous items (items 2, 5, and 7). The second model (Model 2) tested was more constrained in that the Graded Response Model was used to estimate item parameters for the polytomous items, while a Rasch model was estimated for the dichotomous items.

LSST-E. Results of the Model 1 analysis demonstrated adequate fit to the sample data, M_2=24.93, p=.015, RMSEA=.05 (See Table 9), suggesting that the Nominal Response Model and 2-PL adequately fit the polytomous and dichotomous items for the LSST-E. The next model tested was Model 2, which placed greater constraints on the item parameters by estimating the Graded Response and Rasch models. Results demonstrated that this model provided poor fit to the sample data, M_2=145.85, p<.001, RMSEA=.12 (See Table 9), revealing that by placing greater constraints on the item parameters model fit significantly deteriorated. As a result, Model 1 was retained and item parameters are described below for this model (See Table 10).

Examination of the local dependence (LD) χ^2 statistics for Model 1 revealed no major violations (See Table 11), which provided evidence to support that LD violations were avoided by collapsing items that shared similar content. Therefore, the IRT assumption of local independence was met. In examining level 1, the combined polytomous item (items 1 and 2) demonstrated, α_0=.00, α_1=1.25, α_2=4.27, and c_0=.00, c_1=.92, c_2=-1.63. These results indicated that as the categories increased, discrimination increased and correct endorsement of items became less popular (See Figure 3). This result was expected as correctly recalling more items required greater working memory capacity (WMC), thus discriminating between individuals with high and low WMC, which led to fewer participants correctly answering both items 1 and 2 when compared to only answering one of the items correctly. Item 3 demonstrated a high slope, a_3=2.05, and a location parameter, b_3=.51, that was slightly above "average" (θ=0) WMC (See Figure 4). Inspection of the item parameters for the polytomous item (items 4 and 6) revealed,

α_0=.00, α_1=9.07, α_2=10.34, and c_0=.00, c_1=-12.90, c_2=-15.61 (See Figure 5). These parameters support the idea that recalling two items was more discriminating and less popular than recalling only one item; however, the parameter estimates demonstrated that recalling just one item was very highly discriminating and very unpopular in itself. Item 5 was a far less discriminating item, a_5=2.14, when compared to item 7, a_7=26.76; however, the location parameters demonstrated that item 5 was a more difficult item, b_5=2.14, than item 7, b_7=1.50 (See Figures 6 and 7, respectively).

Overall, the test information curve revealed that the LSST-E provided a considerable amount of information for students with a trait estimate of θ=1.5. This result implies that the LSST-E was most sensitive for a very narrow range, suggesting that this measure was effective at differentiating respondents with trait estimates in the 1.3 to 1.75 range (See Figure 8). The test characteristic curve is provided in Figure 9.

LSST-S. Two competing unidimensional models were tested to determine which model best fit the sample data. Results revealed that Model 1 provided adequate fit, M_2=9.91, p=.625, RMSEA=<.01, while Model 2 was a poor fit to the sample data, M_2=151.66, p<.001, RMSEA=.13 (See Table 9). As a result, Model 1 was retained and item parameters for that model are described below (See Table 12).

Examination of the local dependence χ^2 statistics for Model 1 revealed no LD violations (See Table 13). The item parameters for the polytomous item (items 4 and 6) in level 1 demonstrated, α_0=.00, α_1=1.38, α_2=4.31, and c_0=.00, c_1=.07, c_2=-5.67 (See Figure 10). Item 2 exhibited a high slope, a_2=1.64, and a location parameter slightly above the mean, b_2=.90 (See Figure 11). The polytomous item (items 4 and 6) for level 2 displayed,

α_0=.00, α_1=71.10, α_2=49.45, and c_0=.00, c_1=-.141.09, c_2=-98.91 (See Figure 12). These extremely large parameter estimates reflect the low frequency of endorsement for level 2 (See Table 5). Interestingly, the parameter estimates for this item suggest that category 2 was more discriminating and less popular than category 3; however, this may be due to unstable parameter estimates. Item 5 displayed less discrimination, α_5=6.76, when compared to item 7, α_7=16.38; however, item 5, b_5=2.31, was more difficult than item 7, b_7=1.93 (See Figures 13 and 14, respectively). This finding was similar to the one obtained for the LSST-S suggesting that a closer examination of items 5 and 7 is warranted, especially as the items were designed to be progressively more difficult. The test information curve revealed that this measure provided the greatest amount of information for children with a trait estimate of θ=2.00. In addition, the LSST-S was found to be most sensitive to differentiating respondents with ability traits ranging from 1.9 to 2.1, suggesting that this measure was mainly appropriate for participants with very high working memory capacity (See Figure 15). The test characteristic curve is provided in Figure 16.

Concurrent Validity

Descriptive statistics for the raw scores of the measures included in the analysis of concurrent validity are included in Table 14. The attenuated zero-order correlations between both versions of the LSST and measures of fluid intelligence, reading comprehension, and arithmetic computation are presented in Table 15. The results demonstrated weak concurrent validity between the LSST-E and measures of fluid intelligence, (r=.211, p<.001), reading comprehension, (r=.313, p<.001), and arithmetic

computation, (r=.312, p<.001). Weak correlations were also obtained between the LSST-S and measures of fluid intelligence, (r=.136, p<.05), reading comprehension, (r=.144, p<.05), and arithmetic computation, (r=.206, p<.001). The correlation between the raw scores of the LSST-E and LSST-S was weak but statistically significant, (r=.157, p<.001). However, the low reliability must be taken into consideration when examining correlations of the LSST-S with other measures. That is, low reliability would suggest that the majority of variance in the LSST-S was due to error, which in turn attenuates its correlations with other measures. As a result, correlations correcting for attenuation were calculated by dividing the observed correlation with the product of the square roots of their reliabilities (Raykov & Marcoulides, 2011). As shown in Table 16, the disattenuated correlations slightly increased when compared to the attenuated coefficients; however, the strength of the relationships did not change between the Listening Sentence Span Task measures and tests of fluid intelligence, reading comprehension, and arithmetic computation. That is, weak-moderate correlations were revealed between the LSST-E and fluid intelligence, (r=.264, p<.001), reading comprehension, (r=.385, p<.001), and arithmetic computation, (r=.391, p<.001). In contrast, significantly weak correlations were obtained between fluid intelligence, (r=.210, p<.05), reading comprehension, (r=.220, p<.05), and arithmetic computation, (r=.318, p<.001), with the LSST-S. Most importantly, the correlation between the LSST-E and LSST-S did not change greatly as the coefficient was equal to .264, suggesting that the relationship between the adapted measure and its English-counterpart was weak.

Discussion

The psychometric properties of the newly adapted Listening Sentence Span Task-Spanish-version were evaluated in a large sample of ELL students. Overall, findings indicated that the LSST-S was a psychometrically unsound and invalid measure of working memory for ELL children. This was supported by low levels of item-total correlations, item-level internal consistency, and a low overall coefficient α. In addition, confirmatory factor analyses demonstrated non-structural equivalence between the LSST-E and LSST-S, revealing that the measures assess distinctive latent constructs. In addition, poor concurrent validity was exhibited for the LSST-S with measures of fluid intelligence, reading comprehension, and arithmetic computation. However, these results were most likely due to the low overall reliability obtained for the LSST-S (α=.483). One major influence of the psychometrically unsound nature related to the LSST-S was due to the issues present on the LSST-E. This analysis revealed item-level issues with the measure. More specifically, the analysis revealed that item 1 on the LSST-E possessed a low item-total correlation, and low (below .30) internal consistencies for items 2, 3, 5, and 6 were also obtained. As the item properties of a translated/adapted measure will be at best equivalent to the original measure, it is necessary to ensure that all items on the original measure are psychometrically sound (Hambleton, 2001). Therefore, many of the issues present on the LSST-S may have been due to the unsound item properties of the LSST-E.

Although there were item-level issues with the LSST-E, the measure demonstrated an acceptable overall level of internal consistency with the bilingual

sample. Somewhat surprisingly, the LSST-E poorly correlated with Raven's Coloured Progressive Matrices, the Passage Comprehension subtest of the Woodcock-Muñoz Language Survey-Revised, and the written arithmetic subtest of the Wide Range Achievement Test-3rd Edition. A possible explanation for the poor concurrent validity may be due to a floor effect, which may have resulted from a combination of the test's difficulty and young age of the participants. In the original administration of the listening span task, a sample of Carnegie-Mellon University students, recalled on average 2.95 words (*S.D.*=.72) with a range of 2 to 4.5 words (Daneman & Carpenter, 1980). In comparison, the mean number of recalled items on the LSST-E in this study was much lower (*M*=.55, *S.D.*=.98), demonstrating the difficulty of the test.

While, the LSST-E and LSST-S were found to measure distinctive latent constructs, it may be necessary for practical purposes to transform the scores from the different versions onto a common scale in order to correctly interpret scores across the two measures. To accomplish this, a scaling method referred to by Kolen (2004) as *battery scaling* can be adopted as the LSST-E and LSST-S were found to measure different constructs and both measures were administered to a common population of examinees (Holland, 2007). However, in utilizing the disattenuated correlation between the adapted-version and its predecessor as a criterion for linking the two scales, there appears to be no justification for scaling the two measures.

Study 2

As the adapted Spanish-version of the Listening Sentence Span Task was found to be non-equivalent to its English predecessor, it was necessary to analyze the

appropriateness of utilizing the English-version for use with English language learners as an appropriate Spanish-version is not currently available. More specifically, the International Test Commission's test adaptation guidelines suggest that statistical evidence must be provided for the equivalence of questions for all intended populations (Hambleton, 2001). In order to determine the appropriateness of this measure for use with an unintended language group, it was necessary to examine descriptive statistics, structural equivalence, and potential item bias across groups (Sireci, Harter, Yang, & Bhola, 2003). Therefore, the objective of this analysis was to examine item-level summed-score statistics across groups, conduct structural analyses via a confirmatory factor analysis from an IRT framework, and determine whether potential differential item functioning (DIF) was present for the English-version of the Listening Sentence Span Task for ELLs in grades 1-3.

Method

Data Source

The sample was comprised of 806 participants in grades 1-3 obtained from 6 public schools and 1 private school in southern California. Of the 806 students, 315 were non-English language learners obtained from Swanson and Beebe-Frakenberger's (2004) study, while the remaining 491 English language learning students were drawn from study 1 of this paper. The ethnic representation of the sample consisted of 147 White, 22 African-American, 618 Hispanic, 13 Asian-American, 2 Native-American, and 4 self-identified "other" students. The socioeconomic status of the sample ranged from low to upper-middle class based on parental education, occupation, and income level; however,

one must consider that SES varied greatly between ELLs and non-ELLs. That is, the mean SES of the non-ELL sample was primarily middle class, while the mean of the bilingual sample was low-lower middle class. Additional sample demographic information is provided in Table 17.

Measures

As the sample was derived from two separate studies, the non-ELL participants were administered a different battery of assessments than the bilingual participants. However, both groups were administered a battery of group- and individually-administered tests that were comprised of experimental and standardized measures. For a description of the tasks administered to the non-ELL participants in this study refer to Swanson and Beebe-Frakenberger (2004), while the measures administered to the bilingual participants were described in study 1. All participants included in this study were administered the English-version of the Listening Sentence Span Task (LSST-E), which was the primary focus of this analysis. For a description of the LSST-E refer to study 1 of this paper.

Procedure

For a full description of the procedures administered to the monolingual participants refer to Swanson and Beebe-Frankenberger (2004), while a description of the procedures for the bilingual participants can be obtained from study 1 of this paper.

Data Analysis

Item descriptive statistics were performed in IRTPRO 2, Beta version (Cai, du Toit, & Thissen, forthcoming) to examine the means, standard deviations, item-total

correlations, and coefficient alphas for each item across both language status groups (Non-ELL and ELL). In addition, the overall reliability coefficient was produced. An item-total correlation below .30 was considered to reflect a possible problematic item, while an internal consistency coefficient above .70 was considered to reflect adequate reliability.

Following the descriptive statistics analyses, three competing models were conducted to determine measurement invariance across groups via a multiple group IRT framework. More specifically, the structural analysis consisted of testing the following three competing models in IRTPRO 2, Beta version: 1) structural equivalence; 2) factor loading invariance; and 3) factor loading and intercept invariance. Model fit was assessed via the RMSEA estimate associated with the M_2 statistic with an a-priori cut-off level of .06.

Finally, a differential item functioning (DIF) analysis for language-status (ELL vs. non-ELL) was conducted via the logistic regression procedure first introduced by Swaminathan and Rogers (1990) in SAS, version 9.2. Although there are multiple techniques for assessing DIF (e.g., IRT-based methods, Angoff's delta plot, Mantel-Haenszel odds ratio), logistic regression has long been proposed to be a superior method (Rogers & Swaminathan, 1993; Zumbo, 1999). One of the major advantages of utilizing the LR procedure lies in its ability to assess both uniform and non-uniform DIF, while in comparison the Mantel-Haenszel method only has the capability to evaluate uniform DIF (Rogers & Swaminathan, 1993). Furthermore, there are fewer assumptions underlying the LR procedure when compared to IRT-based methods, such as smaller sample sizes are

required to fit the model, and effect sizes can be produced to determine the magnitude of DIF (Sireci, 2011). In utilizing the LR procedure to detect potential item bias, the analysis proceeded in stepwise fashion, involving three distinctive stages for each item. First, total test score (covariate) was entered into the first analysis, then group membership (predictor) was added to the covariate in the second equation, and lastly, the interaction effect between the covariate and predictor was added to total test score and group membership in the third analysis. The statistical significance (p<.01) of the Wald-test statistic for the group membership variable in the second analysis indicated uniform DIF, while non-uniform DIF was specified by the significance (p<.01) of the Wald-test statistic for the interaction effect in the third equation (Sireci, 2011). The magnitude of DIF was assessed by utilizing the guidelines proposed by Jodoin and Gierl (2001), which suggest that small or negligible DIF is indicated by R_{Δ}^{2}<.035, moderate DIF is indicated by .035< R_{Δ}^{2}<.070, and large DIF is indicated by R_{Δ}^{2}>.070. R_{Δ}^{2} was computed between the first and second analyses to examine the effect size for uniform DIF, while the difference in R^2 between the first and third analyses was calculated to determine the effect size for non-uniform DIF. One must note that for the logistic procedure SAS produces the R^2 and R_{adj}^{2}, which is the max-rescaled R^2. The R_{adj}^{2} has been proposed to be superior when comparing models with different numbers of predictors as it corrects for overestimation by controlling for the number of predictors (Liao & McGee, 2003), and thus, was used in determining the magnitude of DIF in this analysis.

Results

Item-Level Summed-Score Statistics

Non-ELLs. In examining the frequency of endorsement, it appeared that items 1-3 were highly endorsed correctly (37%-58%), while items 4-7 were correctly endorsed less than 15% (ranged from 5%-13%). This large decrease in correct endorsement was expected as increasing information was presented in level 2 (4 sentences) when compared to level 1 (3 sentences). Summed-score statistics revealed moderate-strong item-total correlations (.42-.72) for all items, except for item 3. Item 3 had an item-total correlation of .28, which was below the a-priori cut-off level of .30, suggesting that it may be a problematic item that requires further content analysis. Internal consistency coefficients for each item were all above the acceptable range, α ranged from .73-.80 (See Table 18). The overall LSST-E demonstrated an acceptable internal reliability for non-ELL participants, α=.77.

ELLs. As with the non-ELL sample, the first level of questions was more highly endorsed than the second level of questions. Summed-score statistics demonstrated moderate item-total correlations (ranged from .39-.66), except for item 1. The item-total correlation for item 1 was equal to .26, which was below the a-priori cut-off level, suggesting that a closer analysis of this item is required. Interestingly, the LSST-E was also revealed to have a problematic item with the non-ELL sample, however, it was found to be a different item (item 3). Acceptable alpha coefficients were obtained for items 1, 4, and 7, while the alpha coefficients for items 2, 3, 5, and 6 were all slightly below the acceptable cut-off level (See Table 19). Yet, the overall reliability coefficient

for the LSST-E with an ELL sample was acceptable, α=.73, which was comparable to the non-ELL sample. As the traditional Classical Test Theory statistics revealed no major issues with the data, the next step was to determine the underlying dimensionality of the LSST-E with a mixed language-status sample.

Measurement Invariance

Multiple group item factor analytic models were conducted within an IRT framework to determine the measurement invariance across the two language status groups (Non-ELL and ELL). The first model tested examined whether the same dimensionality (one-factor structure) was present between the ELL and non-ELL samples. Results of the analysis demonstrated that both groups measured one latent construct, M_2=32.58, p=.250, RMSEA=.01 (See Table 20). The next model placed greater constraints by testing factor loading invariance across groups. The results of the analysis demonstrated that the model fit deteriorated, M_2=167.77, p<.001, RMSEA=.07 (See Table 20). Lastly, factor loading and intercept invariance across groups was analyzed. The item factor analysis revealed that this model greatly deteriorated fit to the sample data, M_2=371.61, p<.001, RMSEA=.10 (See Table 20), and was thus removed from the analysis as a plausible model. Although model 2 had a slightly larger RMSEA statistic than desired, a likelihood ratio test was conducted to determine whether the structural equivalence and factor loading invariance models were statistically different. Results of the likelihood ratio test revealed that the χ^2 difference value of 95.72 exceeded the χ^2 critical value of 14.07 at α=.05 with 7 degrees of freedom, demonstrating that the

structural equivalence model best fit the data. This result suggests that both the LSST-E and LSST-S measure the same latent construct, allowing for valid group comparisons.

Differential Item Functioning

Differential Item Functioning (DIF) was examined to determine potential item bias for individuals who were dichotomously classified as English language learners (ELLs; n=491) when compared to non-English language learners (non-ELLs; n=315). The logistic regression method was chosen to assess DIF. Analysis of item 1 demonstrated a significant slope coefficient for the language status variable in equation 2 (Wald χ^2=81.38, p<.001), suggesting that uniform DIF was present. Examination of the R_Δ^2, which was equal to .0930, revealed that the uniform DIF for this item was significantly large. Item 2 displayed no DIF as indicated by the non-significant logistic regression coefficient for group membership variable in equation 2 (Wald χ^2=.15, p=.69) and the raw score by language status variable in equation 3 (Wald χ^2=4.22, p=.03). A significant slope coefficient for the interaction effect in equation 3, (Wald χ^2=11.56, p<.001), was obtained for item 3. Analysis of the effect size revealed large DIF, R_Δ^2=.1171. Non-uniform DIF was also obtained for item 4, as the slope coefficient for the equation 3 interaction effect was significant (Wald χ^2=7.61, p<.01), and the magnitude of the DIF was found to be moderate, R_Δ^2=.1171. A significant slope coefficient for language status in equation 2 was obtained (Wald χ^2=10.97, p<.01); however, the magnitude was established to be small or negligible for item 5, R_Δ^2=.0231. Negligible uniform DIF was also obtained for item 6 (Wald χ^2=8.97, p<.01, R_Δ^2=.0217). Non-significant slope coefficients for the language status variable in equation 2 (Wald χ^2=0.12,

p=.72) and the interaction effect for equation 3 were obtained (Wald χ^2=0.86, p=.35), revealing that no DIF was present for item 7 (See Table 21).

Discussion

Examination of the LSST-E with a mixed language status sample demonstrated acceptable item-level summed score statistics (e.g. item-total correlations and item-level internal reliability coefficients) and an adequate overall coefficient α. Furthermore, structural equivalence was supported for the ELL and non-ELL samples when conducting a multiple group categorical item factor analytic model from an IRT framework. As the most basic form of construct validity was achieved, valid group comparisons were permitted, which allowed for an analysis of potential item bias (Gierl, 2000). The analysis revealed that uniform and non-uniform differential item functioning was present for all items, except for items 2 and 7, on the LSST-E for ELLs. Unfortunately, as noted by Zumbo (2007), the current generation of DIF procedures does not shed light on the sources or causes of item bias. As a result, these outcomes suggest that a content review board comprised of working memory and language experts must be formed in order to more carefully examine the content of the DIF items. Hence, further test development is required to improve the LSST-E for use with English language learners.

General Discussion

The two studies demonstrated that there are psychometric concerns with using the Listening Sentence Span Task in assessing working memory capacity for English language learners. As there are few other English language WMC measures for children, it is clear that revisions must be employed to improve the LSST for use with diverse

linguistic populations. A major concern on the LSST-E is possible differential item functioning, suggesting that a content analysis be conducted to examine potential bias in wording for English language learners. In addition, a closer examination of the item properties is warranted as undesirable item-total correlations and item-level alpha coefficients were obtained. Furthermore, it is recommended that a more diverse sample, in terms of age, be administered the LSST to obtain greater variability in responses, particularly for levels 3 and 4. One key issue when examining WMC measures is that unless participants correctly respond to 100% of the items within a level, the measure is discontinued as it is assumed that a participant's WMC has been exceeded. Therefore, in order to examine item characteristics for the entire measure an older sample may be required. Ultimately, it is vital to improve the item properties of the LSST-E in order to further advance the development of the LSST-S, as an adapted measure will at best be equivalent to its predecessor. In anticipation of such an improvement process, suggestions for enhancing the adaptation procedure and test validation of the adapted measure are offered below.

In developing a Spanish-language version of the LSST, it is suggested that proper adaptation procedures be adopted. For example, the International Test Commission published adaptation guidelines, which suggest that acceptable translation procedures are back-translation of an instrument, development of two independent translations with review from a third party, and/or a combination of the two (Hambleton, 2001). Translation of the LSST-S in this study was conducted by a single individual, and then was individually reviewed by two bilingual speakers to evaluate whether the translation

was equivalent to the original version; however, no collaboration occurred amongst the reviewers. This approach has been suggested to have major shortcomings in evaluating a test adaptation (Hambleton, 2001), as translation by a committee may allow for a greater combination of linguistic and psychological expertise, which may result in a superior translation than if done by a single individual (Van de Vijver & Hambleton, 1996). In addition, the LSST-S was based on a direct translation of the original version. Translation of a measure from one language to another does not guarantee equivalence (Sireci & Berberoğlu, 2000). Instead, proper adaptation may require excluding items and replacing them with others that are more appropriate in terms of frequency of use, length of syllables, and appropriateness in the targeted culture (Gudmundsson, 2009). Item equivalence is particularly vital when assessing working memory, as both word duration and complexity influence decay in the phonological loop (Word Length Effect; Baddeley & Logie, 1999), which is a critical aspect in assessing working memory capacity when using verbal complex span tasks.

In validating the LSST-S, it is recommended that a monolingual Spanish-speaking sample be used. One limitation of this paper was the utilization of a single-group design, which makes the assumption that participants are equally proficient in both languages (Sireci, 2005). As was demonstrated, assessment of the sample's receptive and expressive language proficiency revealed that participants' competence in English was much stronger than in Spanish. When this occurs, psychometric evaluation of the measure may not be accurate, making the findings difficult to generalize to a monolingual population. Furthermore, proficiency in a test's language is particularly important when assessing

working memory as difficulty in comprehending the language will place greater constraints on one's central executive system leading to construct irrelevant variance. Furthermore, using a monolingual sample will assist in avoiding issues associated with a single-group study, such as fatigue, motivation, and practice effects (Sireci & Berberoğlu, 2000). Nevertheless, further test development of the LSST should be conducted as the need for such a children's working memory assessment is needed.

References

AERA/APA/NCME (1999). *The standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Alloway, T.P., & Alloway, R.G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*(1), 20-29.

Alvarado, C.G., Ruef, M.L., & Schrank, F.A. (2005). Comprehensive Manual. *Woodcock-Muñoz Language Survey-Revised.* Itasca, IL: Riverside Publishing.

Atkins, P.W.B., & Baddeley, A.D. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics, 19*(4), 537-552.

Baddeley, A.D. (1992). Working Memory. *Science, 255*(5044), 556-559.

Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.H. Gordon (Ed.), *The psychology of learning and motivation* (vol. 8) (pp. 47-89). New York, NY: Academic Press.

Baddeley, A.D., & Logie, R.H. (1999). Working memory: The multiple component model. In A. Mikaye & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp.28-61). New York, NY: Cambridge University Press.

Barrouillet, P., Lépine, R., & Camos, V., (2008). Is the influence of working memory capacity on high-level cognition mediated by complexity or resource-dependent elementary processes?. *Psychonomic Bulletin & Review, 15* (3), 528-534.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.

Bollen. & J. Long (Eds.) *Testing Structural Equation Models*. pp. 136–162. Beverly Hills, CA: Sage.

Brownell, R. (2001). *Expressive one-word picture vocabulary test (Spanish-Bilingual Edition)*. Novato, CA: Academic Therapy Publications.

Cai, L., du Toit, S. H. C., & Thissen, D. (forthcoming). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Chicago, IL: Scientific Software International.

Cai, L., Yang, J.S., & Hansen, M. (in press). Generalized full-information item bifactor analysis.

Casillas, A., & Robbins, S.B. (2005). Test adaptation and cross-cultural assessment from a business perspective: Issues and recommendations. *International Journal of Testing, 5*(1), 5-21.

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using Item Response Theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Conway, A.R.A., Cowan, N., Bunting, M.F., Therriault, D.J., & Minkoff, S.R.B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183.

Creswell, M. (2010). Defending the quality of links between scores from different tests and exams. *Measurement, 8*(4), 157-160.

Curan, P.J., Bollen, K.A., Chen, F., Paxton, P., & Kirby, J.B. (2003). Finite

sampling properties of point estimated and confidence intervals of the RMSEA. *Sociological Methods & Research, 32* (2), 208-252.

Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and Reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466.

Desmette, D., Hupet, M., Schelstraete, M.-A. & Van der Linden, M. (1995). Adaptation en langue française du "Reading Span Test" de Daneman & Carpenter (1980) [A French version of M. Daneman and P. A. Carpenter's (1980) Reading Span Test]. *L'Année Psychologique, 95*(3), 459–482.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*(3), 327-346.

Dunn, L., & Dunn, L. (2007). *Peabody picture vocabulary test* (4th ed.). Minneapolis, MN: Pearson Assessments.

Educational Testing Service (2009). *ETS guidelines for the assessment of English language learners*. Princeton, NJ: Pitoniak, M.J., Young, J.W., Martiniello, M., King, T.C., Buteux, A., & Ginsburgh, M.

Elosúa, M.R., Carriedo, N., & García-Madruga, J.A. (2009). Dos nuevas pruebas de Memoria Operativa de Anáforas [Two new Anaphora Working Memory Tests]. *Infancia y Aprendizaje: Psicodinamica e Psicopatologia, 32*(1), 97-118.

Engle, R.W. (2010). Role of working-memory capacity in cognitive control. *Current Anthroplogy, 51*(1), S17-S26.

Engle, R.W., & Kane, M.J. (2004). Executive attention, working memory capacity, and a

two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (vol. 44) (pp. 145-199). New York, NY: Elsevier.

Engle, R.W., Laughlin, J.E., Tuholski, S.W., & Conway, A.R.A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309-331.

Friedman, N.P., Miyake, A., Corley, R.P., Young, S.E., DeFries, J.C., & Hewitt, J.K. (2006). Not all executive functions are related to intelligence. *Psychological Science, 17*(2), 172-179.

Gathercole, S.E. (2007). Working memory: A system for learning. In Wagner, R.K., Muse, A.E., & Tannenbaum, K.R. (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 233-248). New York, NY: Guilford Press.

Gathercole, S.E., & Baddeley, A.D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology, 81*(4), 439-454.

Gathercole, S.E., & Baddeley, A.D. (1993). Phonological working memory: A critical building block for reading development and vocabulary acquisition? *European Journal of Psychology of Education, 8*(3), 259-272.

Gathercole, S.E., Durling, E., Evans, M., Jeffcock, S., & Stone, S. (2008). Working memory abilities and children's performance in laboratory analogues of classroom activities. *Applied Cognitive Psychology, 22*(8), 1019-1037.

Gierl, M.J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280-296.

Gudmindsson, E. (2009). Guidelines for translating and adapting psychological instruments. *Nordic Psychology, 61*(2), 29-45.

Gutiérrez, M., Jiménez, A., & Castillo, D.M. (1996). Medida de la memoria operativa: Versión informatizada y adaptación al castellano de la tarea de "*Reading Span*" [Assessment of working memory capacity: Computerized version and *adaptation* to Spanish of the *Reading Span* task]. *Psicológica, 17*(2), 215-228.

Hambleton, R.K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hollen, P.W. (2007). A framework and history for score linking. In N. Dorans, M. Pommerich & P. Hollen (Eds.), *Linking and aligning scores and scales* (pp.5-30). New York, NY: Springer.

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

International Test Commission (2010). International Test Commission guidelines for translating and adapting tests. Retrieved from: http://www.intestcom.org

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating power and Type I error rates using an effect size with the logistic regression procedure for DIF. *Applied Measurement in Education, 14*(4), 329-349.

Kane, M.J., Hambrick, D.Z., & Conway, A.R.A. (2005).Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 66-71.

Kane, M.J., Hambrick, D.Z., Tuholski, S.W., Wilhelm, O., Payne, T.W., & Engle, R.W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189-217.

Kolen, M.J. (2004). Linking assessments: Concepts and history. *Applied Psychological Measurement, 28*(4), 219-226.

Kondo, H., & Osaka, N. (2000). Effect of concreteness of target words on verbal working memory: An evaluation using Japanese version of reading span test. *Japanese Journal of Psychology, 71*(1), 51-56.

Liao, J.G., & McGee, D. (2003). Adjusted coefficients of determination for logistic regression. *The American Statistician, 57*(3), 161-165.

Masoura, E.V., Gathercole, S.E., & Bablekou, Z. (2004). Contributions of phonological short-term memory to vocabulary acquisition. *The Journal of the Hellenic Psychological Society, 11*(3), 341-355.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*(471), 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713-732.

Muthén, L. K., & Muthén, B.O. (2007). *Mplus user's guide* (5^{th} ed.). Los Angeles, CA: Muthén & Muthén.

Noël, M.P. (2009). Counting on working memory when learning to count and to add: A

preschool study. *Developmental Psychology*, *45*(6), 1630-1643.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric theory* (3rd Edition). New York: McGraw Hill Inc.

Oberauer, K., Schulze, R., Oliver, W., & Süß, H.M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 61-65.

Raven, J.C. (1956). *Coloured Progressive Matrices*. London, UK: H.K. Lewis & Co. Ltd.

Raven, J.C., Court, J.H., & Raven, J. (1990). *Manual for Raven's progressive matrices and vocabulary scales: Coloured progressive matrices (1990 edition, with U.S. norms)* . Oxford, UK: Oxford Psychologists Press.

Raykov, T., & Marcoulides, G.A. (2011). *Introduction to psychometric theory.* New York, NY: Routledge.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

SAS Institute Inc. (2008). SAS/STAT User's guide, Cary, NC: SAS Institute Inc.

Shah, P., & Miyake, A. (1999). Models of working memory: An introduction. In A. Miyake, & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 1-27). New York, NY: Cambridge University Press.

Shelton, J.T., Elliott, E.M., Matthews, R.A., Hill, B.D., & Gouvier, W.D. (2010). The relationship of working memory, secondary memory, and general fluid

intelligence: Working memory is special. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 813-820.

Sireci, S.G. (2005). Using bilinguals to evaluate the comparability of different versions of a test. In R.K. Hambleton, P.F. Merenda, and C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp.117-138). Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S.G. (2011). Evaluating test and survey item for bias across languages and cultures. In D. Matsumoto & F Van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 216-240). New York, NY: Cambridge University Press.

Sireci, S.G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13*(3), 229-248.

Sireci, S.G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing, 3*(2), 129-150.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Swanson, H.L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology, 84*(4), 473-488.

Swanson, H.L. (1996). *Swanson-Cognitive Processing Test: Examiner's manual*. Austin, TX: PRO-ED.

Swanson, H.L. (1996). Individual and age-related differences in children's working memory. *Memory & Cognition, 24*(1), 70-82.

Swanson, H.L. (1999). Sentence span measure (SSM). In Frederickson, N., & Cameron, R.J. (Eds.), *Psychology in Education Portfolio: Memory and Listening Comprehension* (pp.25-26). Winsor, England: NFER-Nelson.

Swanson, H.L. (2008). Working memory and intelligence in children: What develops? *Journal of Educational Psychology, 100*(3), 581-602.

Swanson, H.L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problems solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 96*(3), 471-491.

Swanson, H.L., Cochran, K.F., & Ewers C.A. (1989). Working memory in skilled and less skilled readers. *Journal of Abnormal Child Psychology, 17*(2), 145-156.

Swanson, H.L., Jerman, O., Zheng, Xinhua, Z. (2008). Growth in working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 100*(2), 343-379.

Swanson, H.L., Kehler, P., & Jerman, O. (2010). Working memory, strategy knowledge, and strategy instruction in children with reading disabilities. *Journal of Learning Disabilities, 43*(1), 24-47.

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247-260.

Tompkins, C.A., Bloise, C.G.R., Timko, M.L., & Baumgaertner, A. (1994). Working memory and inference revision in brain-damaged and normally aging adults. *Journal of Speech & Hearing Research, 37*(4), 896-912.

U.S. Department of Education, Office of English Language Acquisition (2010). *The growing numbers of limited English proficient students: 1997/98-2007/08*. Retrieved from http://www.ncela.gwu.edu/files/uploads/9/growingLEP_0708.pdf

Van de Vijver, F., & Hambleton, R.K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2), 89-99.

Vock, M., & Holling, H. (2008). The measurement of visuo-spatial and verbal-numerical working memory: Development of IRT-based scales. *Intelligence, 36*(2), 161-182.

Wilkinson, G.S. (1993). *Wide Range Achievement Test: Administration manual* (3rd ed.). Wilmington, DE: Wide Range Inc.

Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: Current approaches and future directions. *Psychological methods, 12*(1), 58-79.

Woodcock, R.W., Muñoz-Sandoval, A.F., Ruef, M.L., & Alvarado, C.G. (2005). *Woodcock-Muñoz Language Survey-Revised, English*. Itasca, IL: Riverside Publishing.

Woodcock, R.W., Muñoz-Sandoval, A.F., Ruef, M.L., & Alvarado, C.G. (2005). *Woodcock-Muñoz Language Survey-Revised, Spanish Form*. Itasca, IL: Riverside Publishing.

Yu, C. Y., & Muthén, B. O. (2001). *Evaluation of model fit indices for latent variable*

models with categorical and continuous outcomes (Technical report). Los Angeles: University of California, Los Angeles, Graduate School of Education and Information Studies.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Footnotes

[1] The term *test adaptation* is preferable to *test translation* as it implies flexibility in substituting language that is more appropriate than a literal word-for-word translation (Sireci & Berberoğlu, 2000). Throughout this paper the term *test adaptation* will be used.

[2] This measure has been previously referred to as the *Children's Adaption of Sentence Span Task* (Swanson, 1992), *Sentence Span Task* (Swanson, Cochran, & Ewers, 1989; Swanson, 1996b), and the *Sentence Span Measure* (Swanson, 1999).

[3] Administration of the Listening Sentence Span Task was part of the *Growth in Literacy, Language, and Cognition in Children with Reading Disabilities who are English Language Learners* project (R324A090092), which is supported by a 4 year grant awarded to Dr. H. Lee Swanson from the U.S. Department of Education's Institute of Education Sciences

Tables

Table 1

Sample Demographic Characteristics

Variable	Value (Percentage)
Gender[a]	
Male	230 (46.84%)
Female	261 (53.16%)
Race or Ethnicity[a]	
Hispanic	489 (99.59%)
Self-Identified as "Other"	2 (.41%)
Age (years)[b]	M= 7.72 (S.D.=.93)
Grade[a]	
1st	159 (32.38%)
2nd	151 (30.75%)
3rd	181 (36.86%)
Language Spoken at Home[c]	
Spanish	384 (78.85%)
English	35 (86.04%)
Both Spanish & English	67 (13.76%)
Other[1]	1 (.21%)
Eligibility for Free or Reduced Lunch[d]	310 (96.27%)

Note. Demographic information was not available for all participants.
[a]n=491; [b]n=488; [c]n=487; [d]n=322
[1]The one student whose language spoken at home was "other" was reported to speak Spanish, English, and Arabic.

Table 2

Means and Standard Deviations for Language Proficiency Measures

Measure	*M*	*SD*
Receptive Language Proficiency		
PPVT (English-raw)[a]	97.47	20.05
PPVT (English-standard)[a]	83.29	9.81
PPVT (Spanish-raw)[d]	44.96	16.44
PPVT (Spanish-standard)[d]	81.61	16.25
Expressive Language Proficiency		
EOWPVT-SBE (English-raw)[b]	47.80	13.51
EOWPVT-SBE (English-standard)[c]	91.06	17.88
EOWPVT-SBE (Spanish-raw)[e]	27.87	16.72
EOWPVT-SBE (Spanish-standard)[f]	67.39	16.01

Note. PPVT=Peabody Picture Vocabulary Test; EOWPVT-SBE=Expressive One-Word Picture Vocabulary Test, Spanish-Bilingual Edition;
[a]n=486; [b]n=485; [c]n=484; [d]n=483; [e]n=472; [f]n=471

Table 3

Descriptive Statistics of LSST-E and LSST-S

Measure	N	Mean	*SD*	Minimum	Maximum
LSST-E[a]	491	1.217	1.351	0	7.00
LSST-S[a]	491	.690	.879	0	6.00

Note. If participants received a zero on the process question of either the LSST-E or LSST-S, the participant was given a zero raw score for that particular item and the assessment was discontinued.

Table 4

Initial Individual Item Summary for LSST-E

Item	Mean	*SD*	Frequency Correct	Percent Correct	Item-Total Correlation	Coefficent α
1	.633	.482	311	63.34	.254	.748
2	.348	.477	171	34.83	.473	.698
3	.279	.449	137	27.90	.546	.678
4	.051	.220	25	5.09	.434	.701
5	.090	.286	44	8.96	.672	.663
6	.073	.261	36	7.33	.627	.673
7	.041	.198	20	4.07	.490	.697
8	.004	.064	2	.41	.134	.729
9	.006	.078	3	.61	.375	.720
10	.006	.078	3	.61	.375	.720
11	.006	.078	3	.61	.343	.721
12	.004	.064	2	.41	.311	.724

Note. Testlet 3, which included items 8-12, was the highest level reached by participants. As a result, testlet 4 was not administered and thus, not included in the analysis.

Table 5

Individual Item Summary for LSST-S

Item	Mean	*SD*	Frequency Correct	Percent Correct	Item-Total Correlation	Coefficent α
1	.475	.500	233	47.45	.162	.561
2	.259	.438	127	25.87	.336	.391
3	.071	.258	35	7.13	.307	.412
4	.006	.078	3	.61	.288	.467
5	.020	.141	10	2.04	.475	.404
6	.020	.141	10	2.04	.441	.412
7	.004	.064	2	.41	.278	.473

Note. Testlet 2 (items 4-7) was the highest level reached by participants. As a result, testlets 3 and 4 were not administered and thus, not included in the analysis.

Table 6

Revised Individual Item Summary for LSST-E

Item	Mean	*SD*	Frequency Correct	Percent Correct	Item-Total Correlation	Coefficent α
1	.633	.482	311	63.34	.261	.765
2	.348	.477	171	34.83	.489	.695
3	.279	.449	137	27.90	.563	.669
4	.051	.220	25	5.09	.398	.716
5	.090	.286	44	8.96	.663	.661
6	.073	.261	36	7.33	.613	.676
7	.041	.198	20	4.07	.443	.713

Note. The item-correlation and coefficient α statistics reflect estimates based on the first two testlets. Results from testlet 3 were dropped for purposes of linking the LSST-S with the LSST-E.

Table 7

Measurement Model Fit Indices

Model		*df*	*p*	CFI	TLI	RMSEA
Unidimensional LSST-E	34.791	8	<.001	.991	.988	.083
Unidimensional LSST-S	4.559	8	.472	1.00	1.00	<.001
Unidimensional LSST-E & LSST-S	135.352	16	<.001	.920	.915	.123
Simple Structure	45.154	18	<.001	.982	.983	.055

Table 8

Factor Loadings for Two-Factor Structure

Item	λ_1	s.e.	λ_2	s.e.
LSST-E Item 1	.64	.03	---	---
LSST-E Item 2	.80	.03	---	---
LSST-E Item 3	.85	.03	---	---
LSST-E Item 4	.86	.05	---	---
LSST-E Item 5	1.00	.01	---	---
LSST-E Item 6	.96	.01	---	---
LSST-E Item 7	.87	.03	---	---
LSST-S Item 1	---	---	.35	.07
LSST-S Item 2	---	---	.80	.07
LSST-S Item 3	---	---	.79	.06
LSST-S Item 4	---	---	.80	.05
LSST-S Item 5	---	---	1.01	.02
LSST-S Item 6	---	---	.95	.02
LSST-S Item 7	---	---	.84	.11

Table 9

IRT Model Parameter Fit Statistics

Measure	Model	M_2	*df*	*p*	RMSEA	-2 loglikelihood	AIC	BIC
LSST-E	1	24.93	12	.015	.05	2021.51	2049.51	2108.27
	2	145.85	17	<.001	.12	2130.47	2148.47	2186.23
LSST-S	1	9.91	12	.625	<.01	1436.17	1464.17	1522.92
	2	151.66	17	<.001	.13	1477.34	1495.34	1533.11

Note. Model 1 consisted of the 2-PL model for dichotomous items and the Nominal Response Model for polytomous items. Model 2 consisted of the Rasch model for dichotomous items and the Graded Response Model for polytomous items.

Table 10

Mixed Format Unidimensional Item Parameter Estimates for LSST-E

Item	Category	1	2	3
Combined Items 1 and 2	*a*	.00	1.25	4.27
	c	.00	.92	-1.63
Combined Items 4 and 6	*a*	.00	9.07	10.34
	c	.00	-12.90	-15.61

Note. Nominal Response Model for polytomous items

Item	α	s.e.	*b*	s.e.
2	2.05	.34	.51	.08
5	2.14	.41	2.14	.21
7	26.76	1.11	1.50	.07

Note. 2-PL model for dichotomous items

Table 11

Local Dependence χ^2 statistics for LSST-E

Item	1	2	3	4
1				
2	-0.9			
3	-0.7	-0.1		
4	2.1	0.7	0.4	
5	-0.6	-0.5	-0.7	1.0

Table 12

Mixed Format Unidimensional Item Parameter Estimates for LSST-S

Item	Category	1	2	3
Combined Items 1 and 2	*a*	.00	1.38	4.31
	c	.00	.07	-5.67
Combined Items 4 and 6	*a*	.00	71.10	49.45
	c	.00	-141.09	-98.91

Note. Nominal Response Model for polytomous items

Item	α	s.e.	*b*	s.e.
2	1.64	.46	.90	.17
5	6.76	.67	2.31	.21
7	16.38	1.83	1.93	.27

Note. 2-PL model for dichotomous items

Table 13

Local Dependence χ^2 statistics for LSST-S

Item	1	2	3	4
1				
2	0.3			
3	0.8	1.1		
4	0.7	---	---	
5	0.9	0.8	1.2	---

Table 14

Descriptive Statistics of Measures Used for Concurrent Validation

Measure	N	Mean	S.D.	Minimum	Maximum
CPM	458	22.589	6.437	0	35.00
PC	470	12.791	5.122	1.00	23.00
CDT	479	7.022	4.031	0	19.00
WRAT-A	437	22.410	4.182	12.00	34.00

Note. LSST-E=English Listening Sentence Span Task; LSST-S= Spanish Listening Sentence Span Task; CPM= Raven's *Coloured Progressive Matrices*; PC= Passage Comprehension subtest of the *Woodcock-Muñoz Language Survey-Revised*; CDT= *Comprensión de textos* subtest of the *Woodcock-Muñoz Language Survey-Revised*; WRAT-A= Arithmetic subtest of the *Wide Range Achievement Test-3rd Edition*

Table 15

Concurrent Validity Attenuated Correlation Matrix

Measure	LSST-E	LSST-S	CPM	PC	CDT	WRAT
LSST-E	1.00					
LSST-S	.157[a**]	1.00				
CPM	.211[e**]	.136[e*]	1.00			
PC	.313[c**]	.159[c**]	.531[g**]	1.00		
CDT	.240[b**]	.144[b*]	.325[f**]	.450[d**]	1.00	
WRAT-A	.312[h**]	.206[h**]	.565[i**]	.640[k**]	.461[j**]	1.00

a=491; b=483; c=470; d=467; e=458; f=452; g=442; h=437; i=436; j=433; k=425;
*p<.05 **p<.001

Table 16

Concurrent Validity Disattenuated Correlation Matrix

Measure	LSST-E	LSST-S	CPM	PC	CDT	WRAT
LSST-E	1.00					
LSST-S	.264[a**]	1.00				
CPM	.264[e**]	.210[e*]	1.00			
PC	.385[c**]	.241[c**]	.600[g**]	1.00		
CDT	.297[b**]	.220[b*]	.369[f**]	.503[d**]	1.00	
WRAT-A	.391[h**]	.318[h**]	.649[i**]	.723[k**]	.524[j**]	1.00

a=491; b=483; c=470; d=467; e=458; f=452; g=442; h=437; i=436; j=433; k=425;
*p<.05 **p<.001

Table 17

Sample's Demographic Characteristics for Study 2

Variable	Value (Percentage)
Gender	
Male	395 (49.01%)
Female	411 (50.99%)
Race or Ethnicity	
White	147 (18.24%)
African American	22 (2.73%)
Hispanic	618 (76.67%)
Asian American	13 (1.61%)
Native American	2 (.25%)
Self-Identified as "Other"	4 (.50%)
Age (years)	M= 7.74 (S.D.=.94)
Grade	
1st	264 (32.75%)
2nd	236 (29.28%)
3rd	306 (37.97%)
Language Status	
Non-ELL	315 (39.08%)
ELL	491 (60.92%)

Table 18

LSST-E Item Descriptive Statistics for Non-ELLs

Item	Mean	*SD*	Frequency Correct	Percent Correct	Item-Total Correlation	Coefficent α
1	.403	.491	127	40.31	.541	.741
2	.375	.485	118	37.46	.582	.730
3	.584	.494	184	58.41	.281	.805
4	.133	.340	42	13.33	.728	.705
5	.095	.294	30	9.52	.622	.732
6	.076	.266	24	7.61	.572	.743
7	.057	.232	18	5.71	.421	.765

Table 19

LSST-E Item Descriptive Statistics for ELLs

Item	Mean	*SD*	Frequency Correct	Percent Correct	Item-Total Correlation	Coefficent α
1	.633	.482	311	63.34	.261	.765
2	.348	.477	171	34.82	.489	.695
3	.279	.449	137	27.90	.563	.669
4	.051	.220	25	5.09	.398	.716
5	.090	.286	44	8.96	.662	.661
6	.073	.261	36	7.33	.612	.676
7	.041	.198	20	4.07	.442	.713

Table 20

Measurement Model Fit Indices

Model	M_2	*df*	*p*	RMSEA	-2loglikelihood
Structural Equivalence	32.58	28	.25	.01	3726.62
Factor Loading Invariance	167.77	35	<.001	.07	3822.34
Factor Loading & Intercept Invariance	371.61	42	<.001	.10	4003.91

Table 21

DIF Analysis Results

	Uniform DIF		Non-Uniform DIF				
Item	Wald	*p*	Wald	*p*	Type of DIF Present		Effect Size
1	81.38	<.001	2.34	.12	Uniform	.0930	Large DIF
2	0.15	.69	4.22	.03	None	---	---
3	92.53	<.001	11.56	<.001	Non-Uniform	.1171	Large DIF
4	12.91	<.001	7.61	<.001	Non-Uniform	.0571	Medium DIF
5	10.97	<.001	2.91	.08	Uniform	.0231	Small DIF
6	8.97	<.01	0.11	.73	Uniform	.0217	Negligible DIF
7	0.12	.72	0.86	.35	None	---	---

Note. Although there was significant uniform and non-uniform DIF for items 3 and 4, non-uniform DIF was given precedence as it detects an interaction of group by ability, while uniform DIF only identifies a group interaction (Zumbo, 2007).

Figures

Figure 1

One-Factor Measurement Model for both the LSST-E and LSST-S

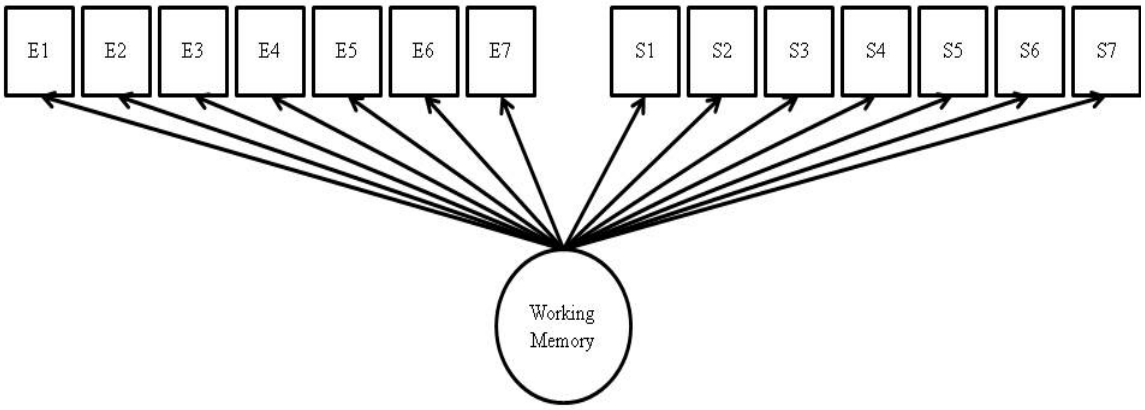


Figure 2

Simple Structure Model

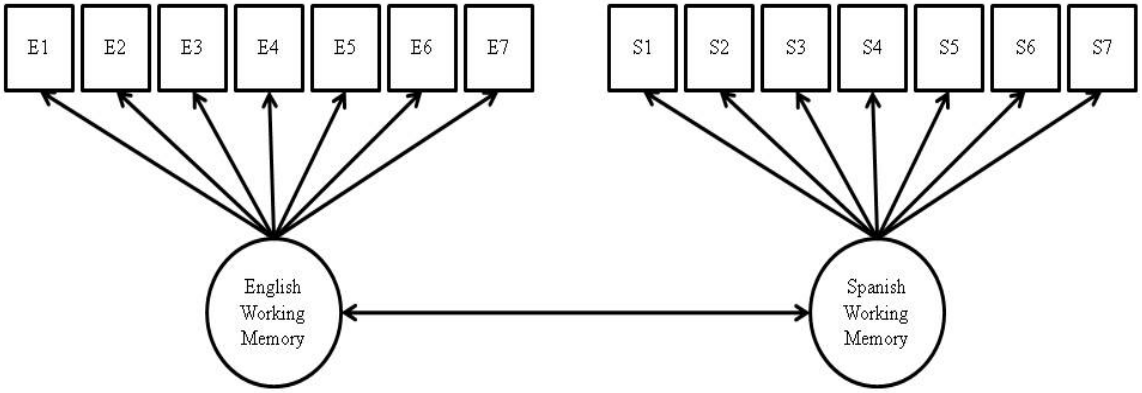


Figure 3

Trace Lines for Polytomous Item (Items 1 & 2) of LSST-E

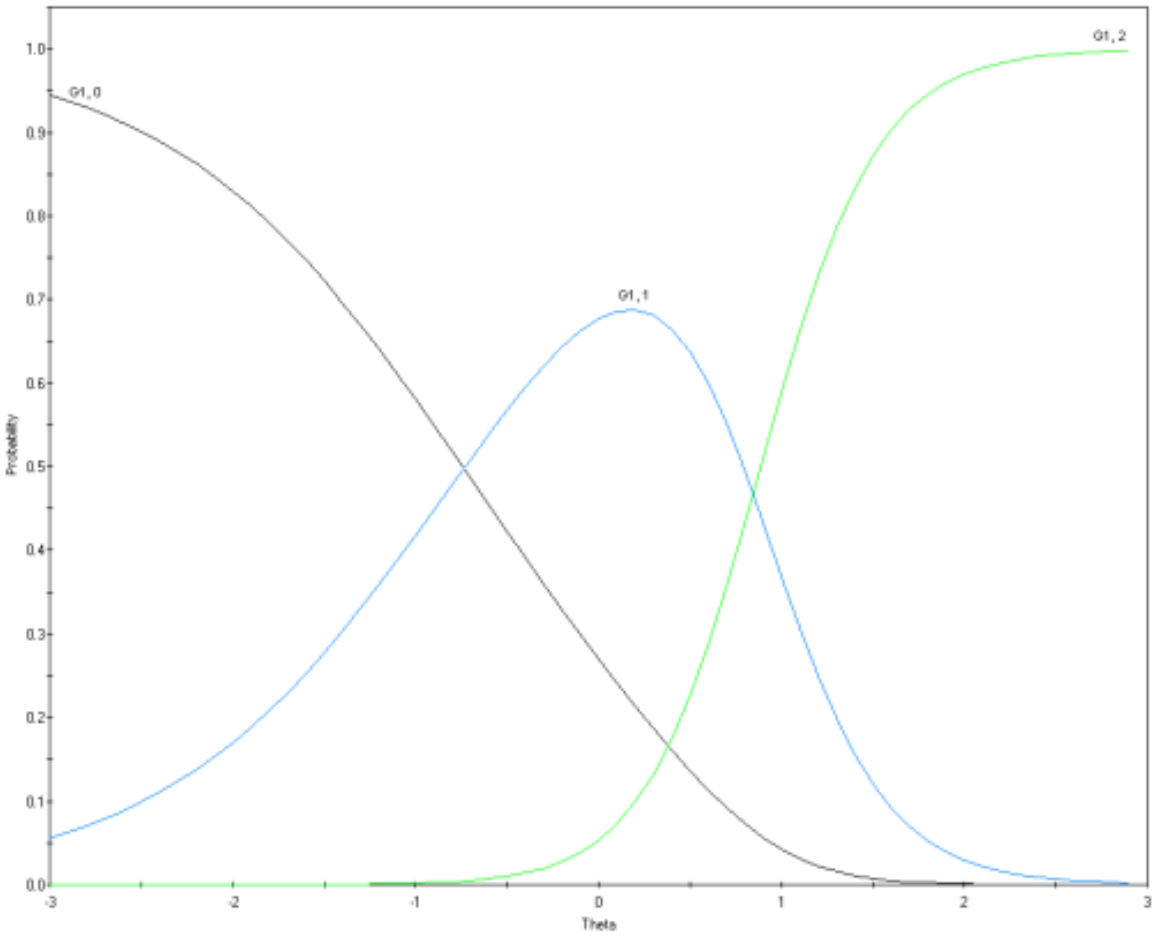


Figure 4

Trace Lines for Item 3 of LSST-E

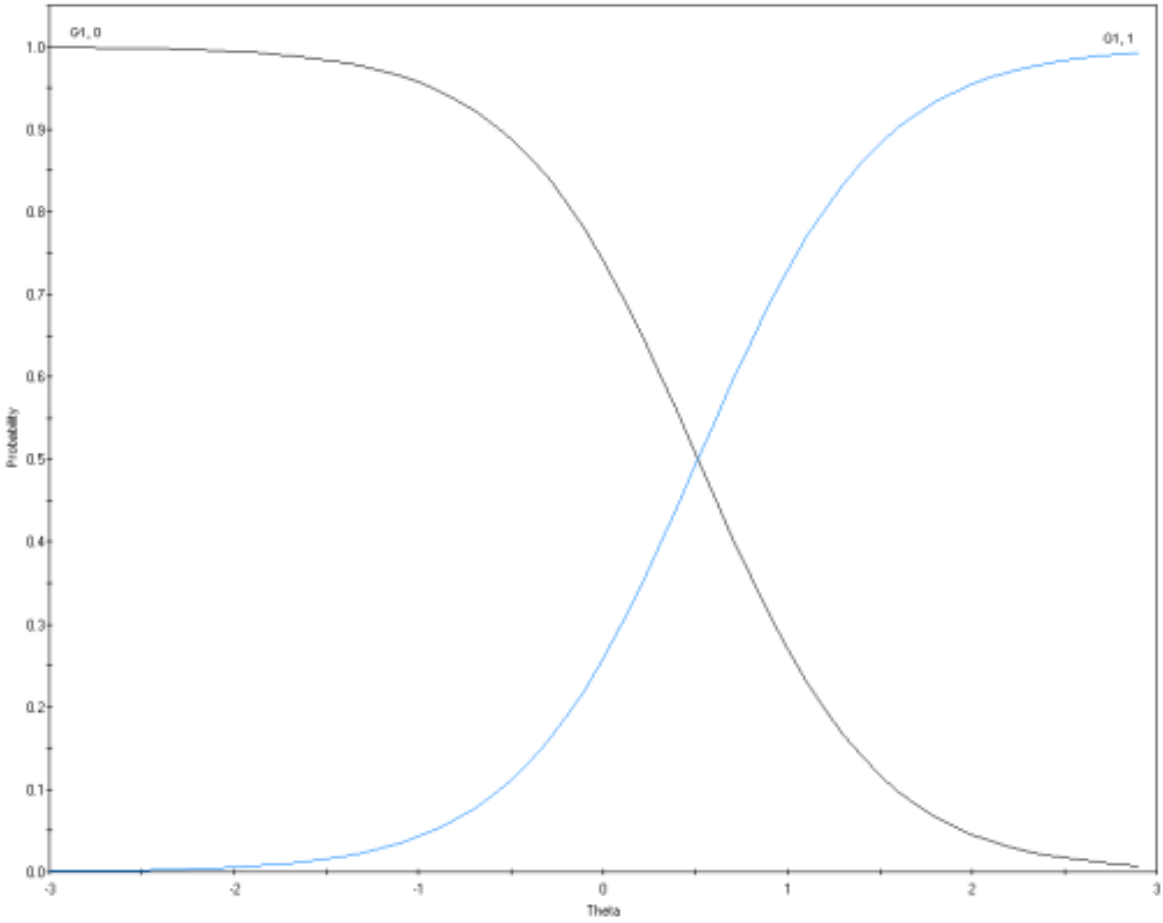


Figure 5

Trace Lines for Polytomous Item (Items 4 & 6) of LSST-E

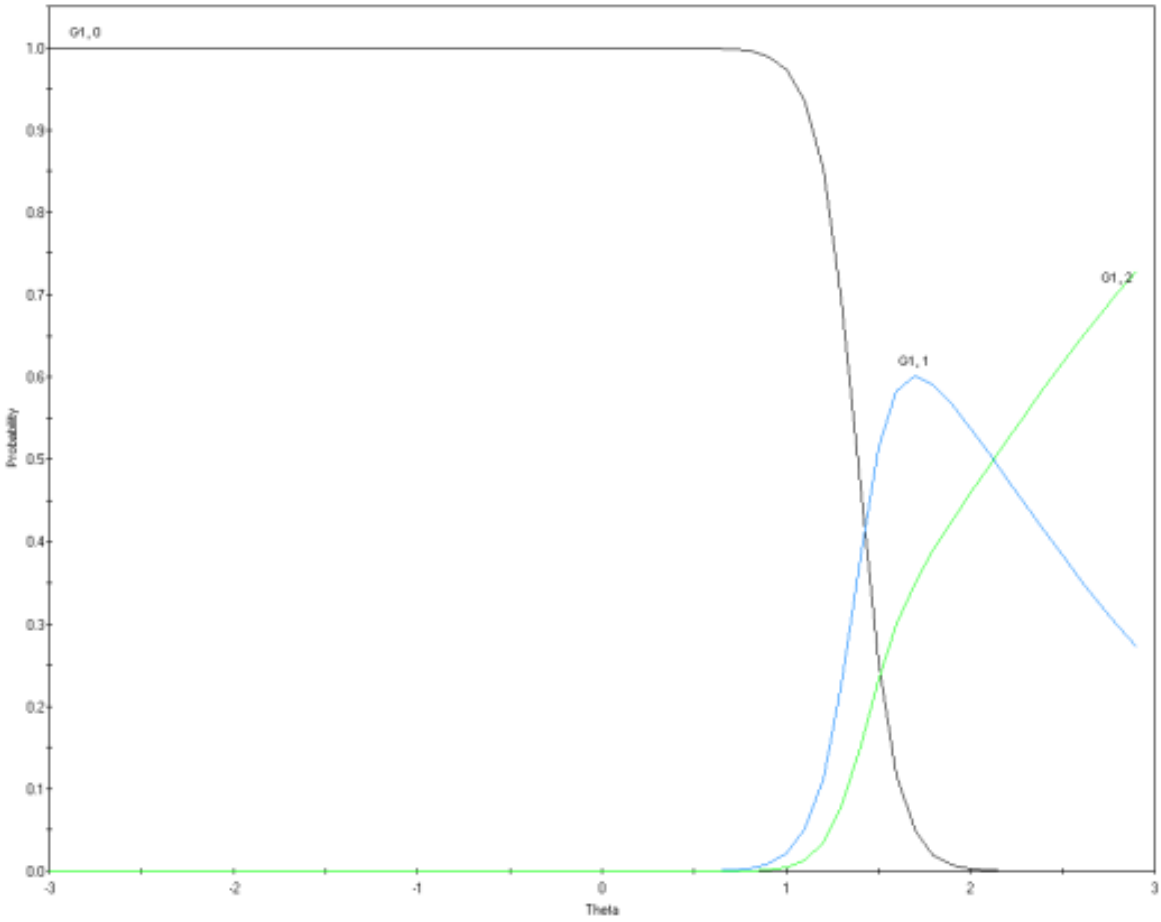


Figure 6

Trace Lines for Item 5 of LSST-E

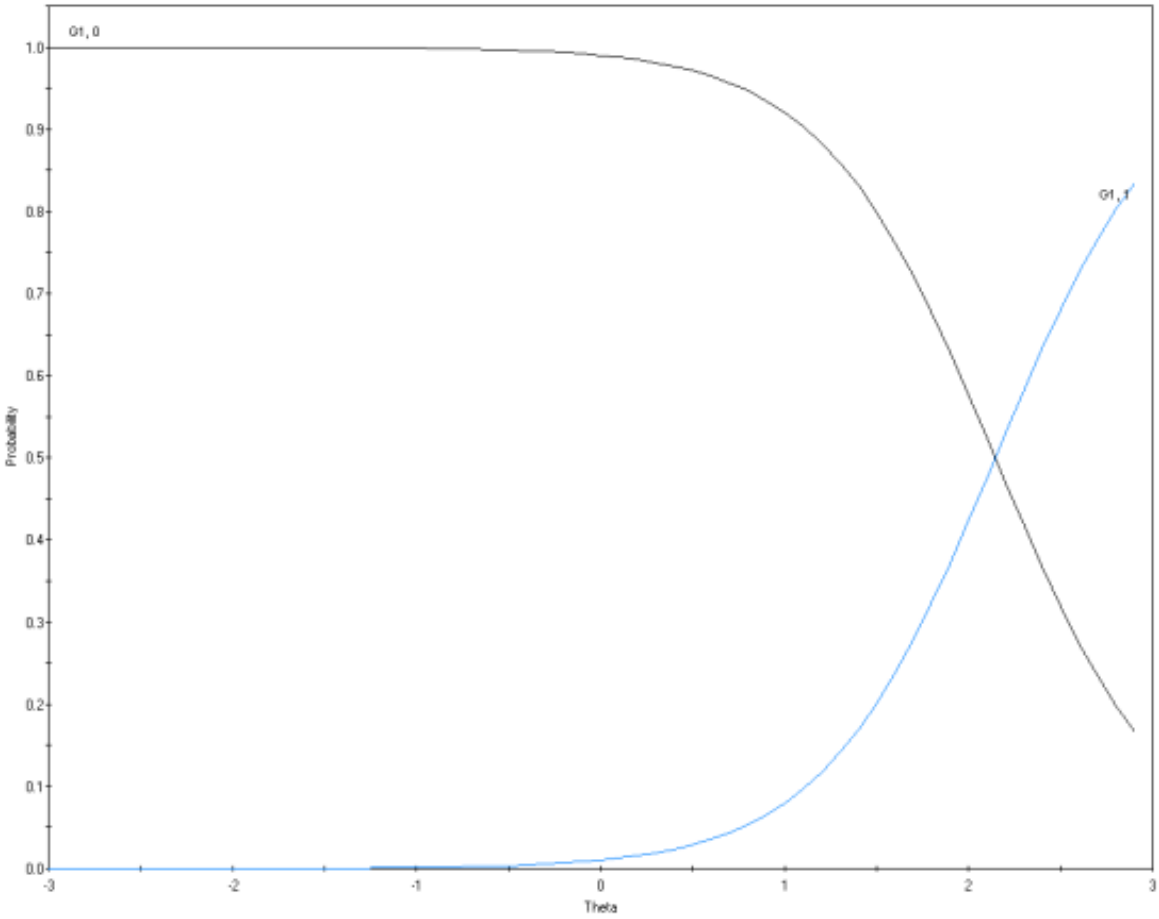


Figure 7

Trace Lines for Item 7 of LSST-E

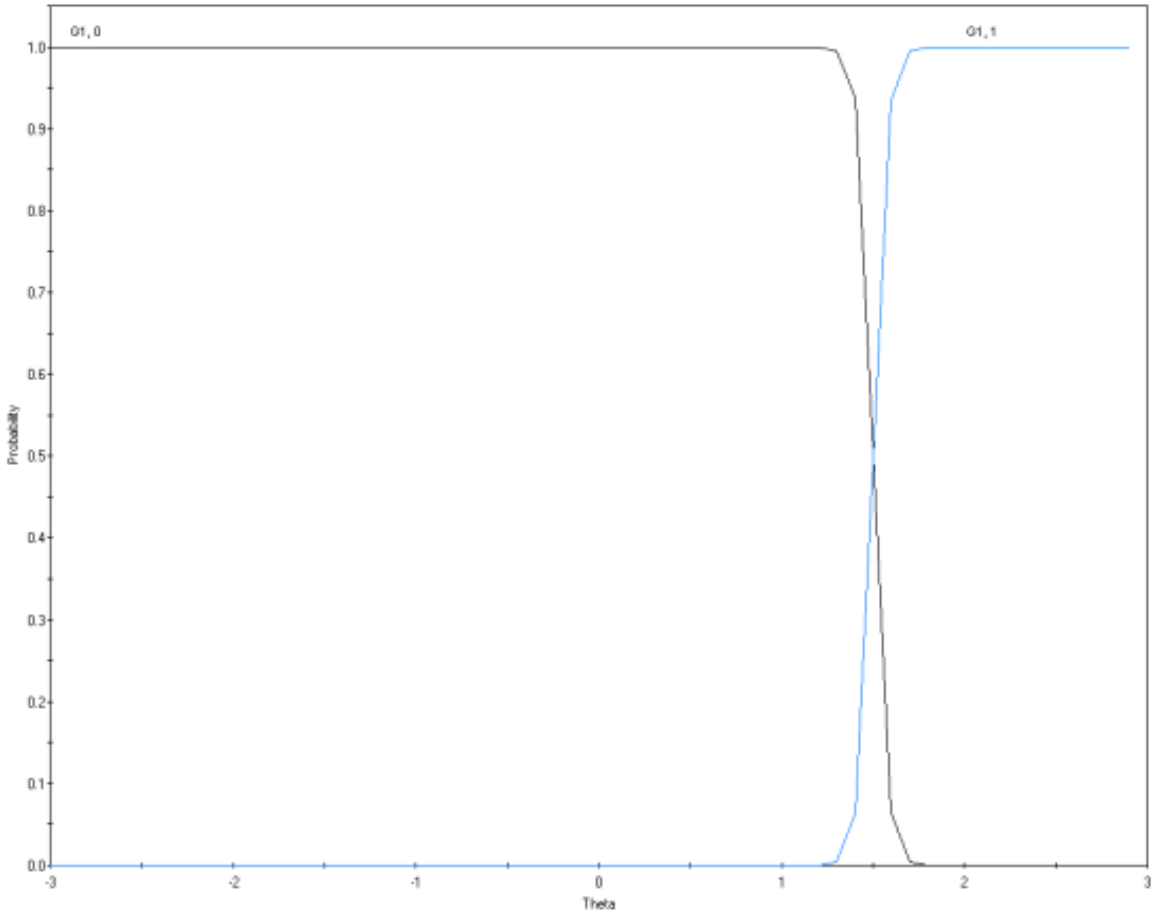


Figure 8

Test Information Curve for LSST-E

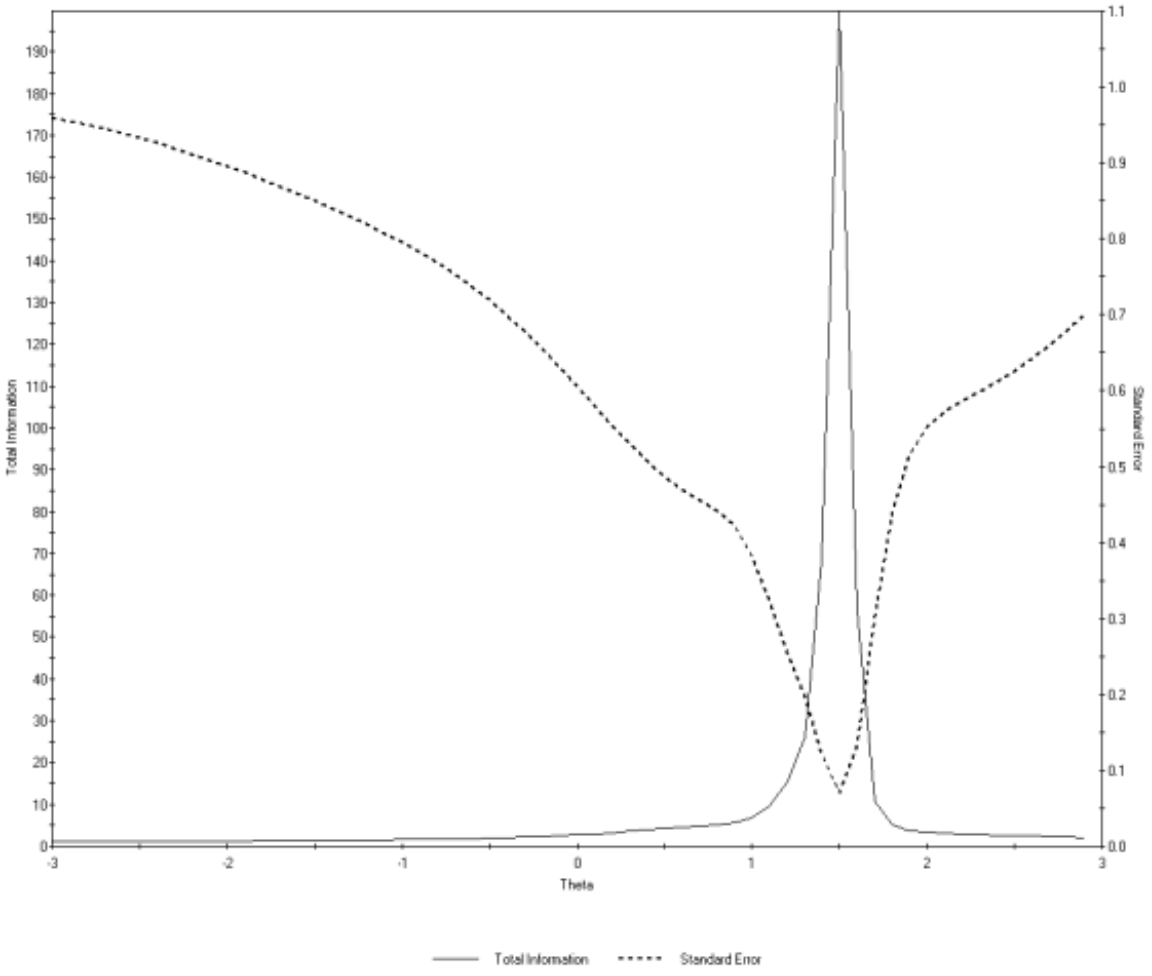


Figure 9

Test Characteristic Curve for LSST-S

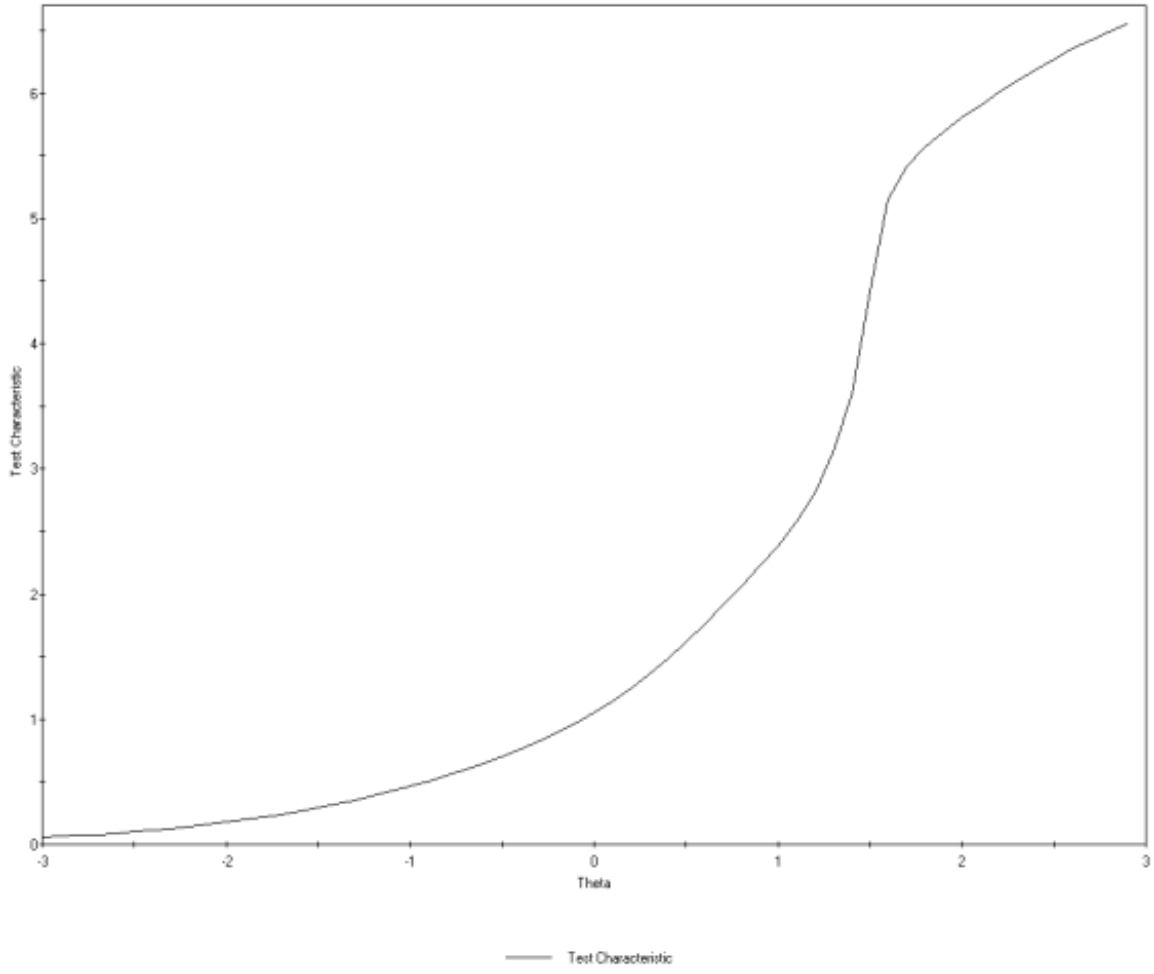


Figure 10

Trace Lines for Polytomous Item (Items 1 & 2) of LSST-S

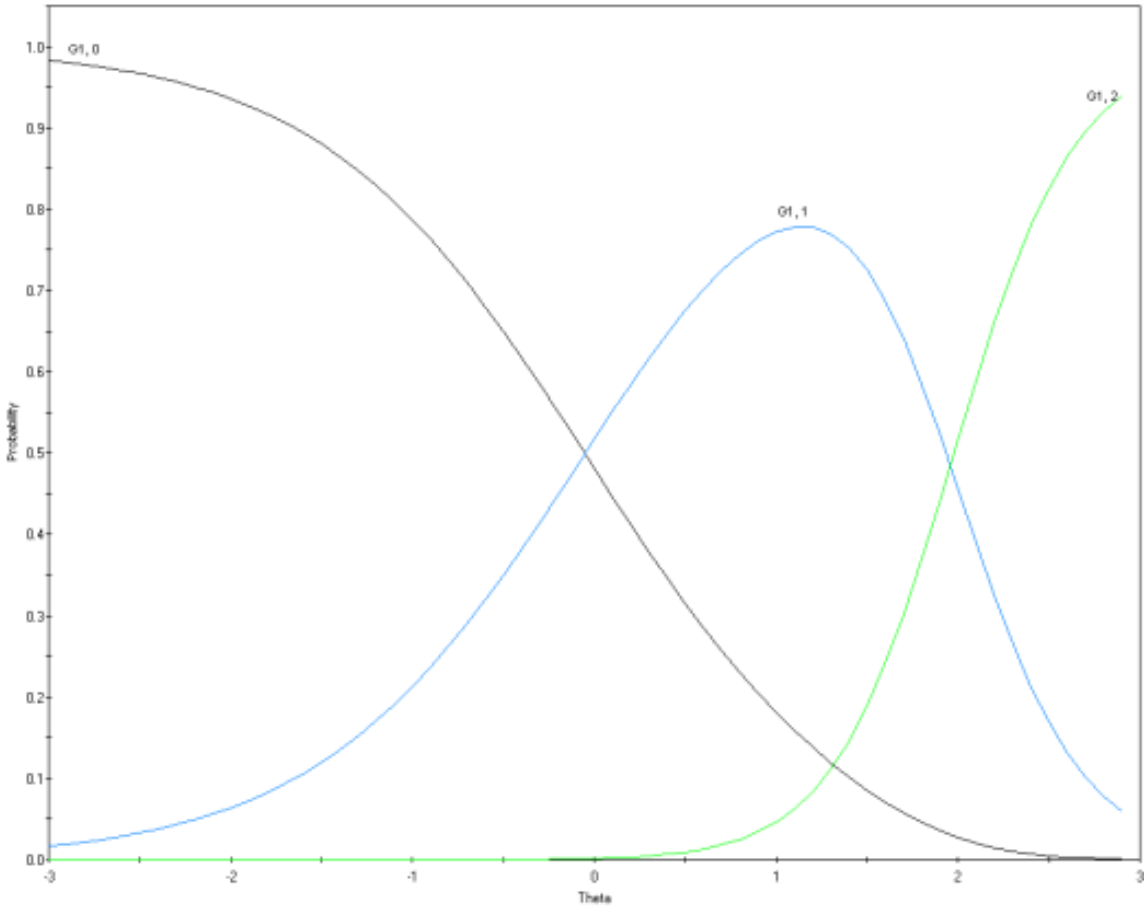


Figure 11

Trace Lines for Item 3 of LSST-S

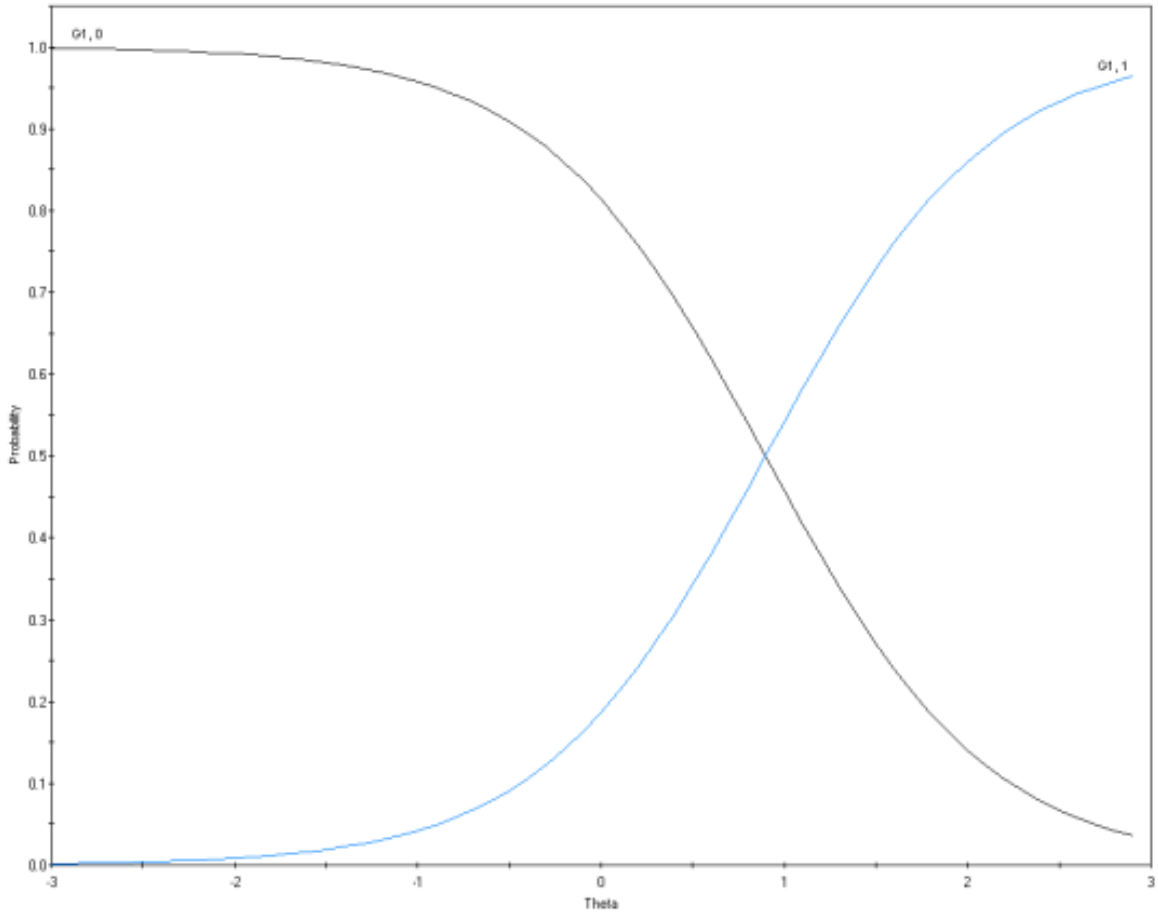


Figure 12

Trace Lines for Polytomous Item (Items 4 & 6) of LSST-S

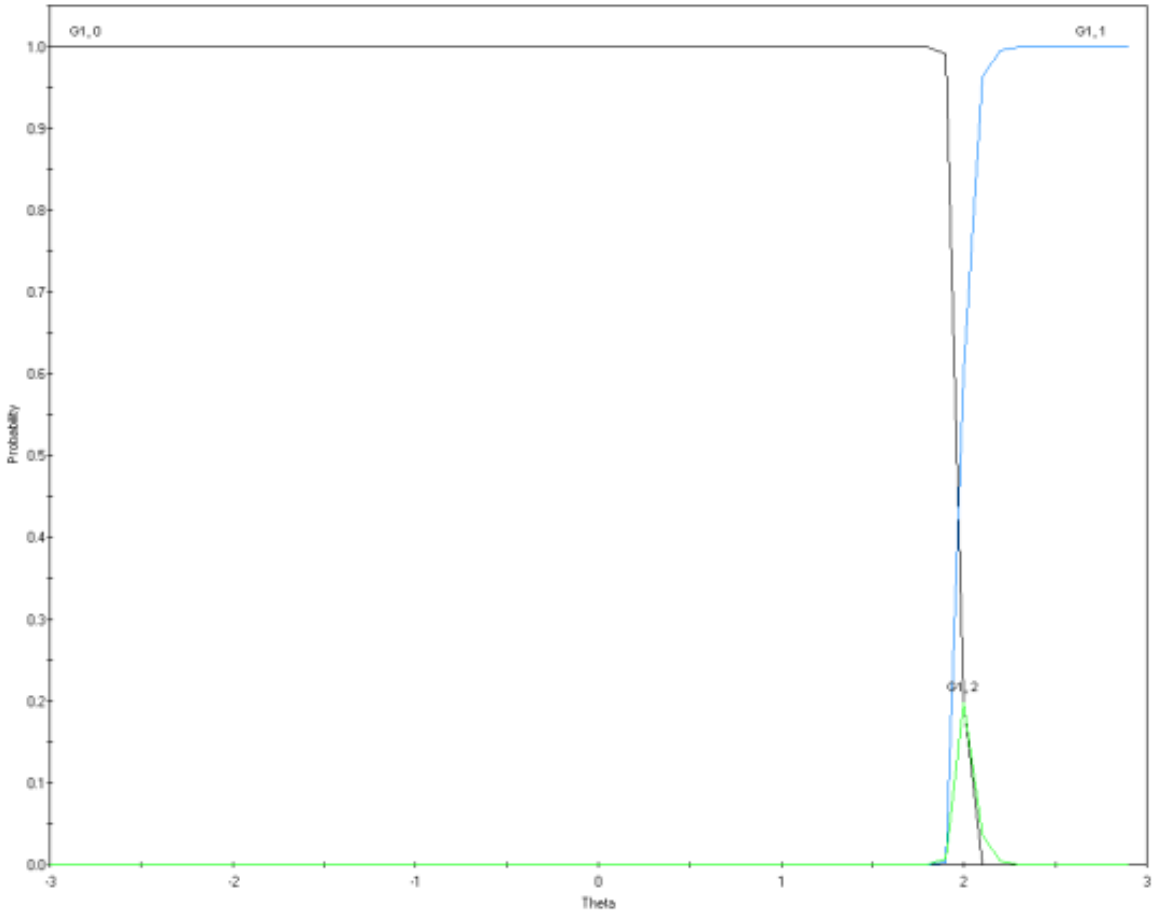


Figure 13

Trace Lines for Item 5 of LSST-S

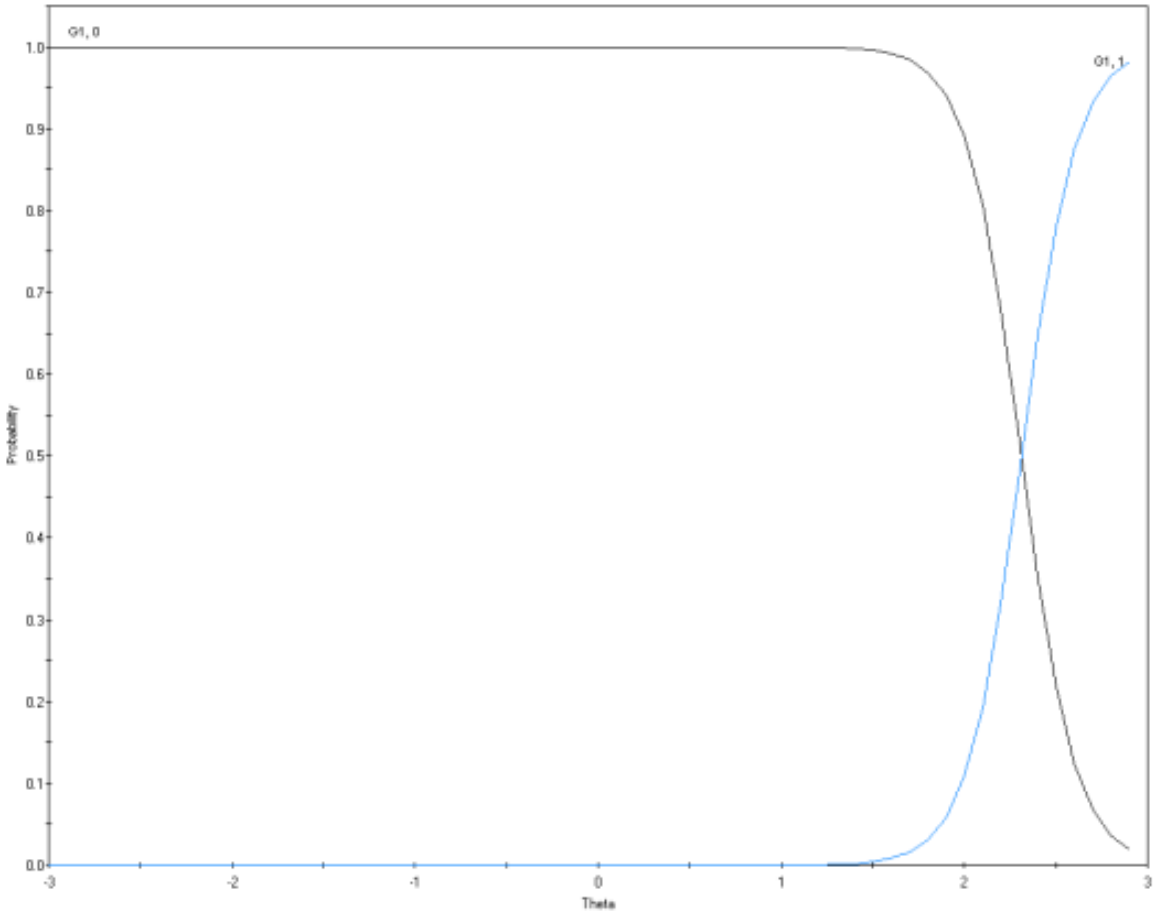


Figure 14

Trace Lines for Item 7 of LSST-S

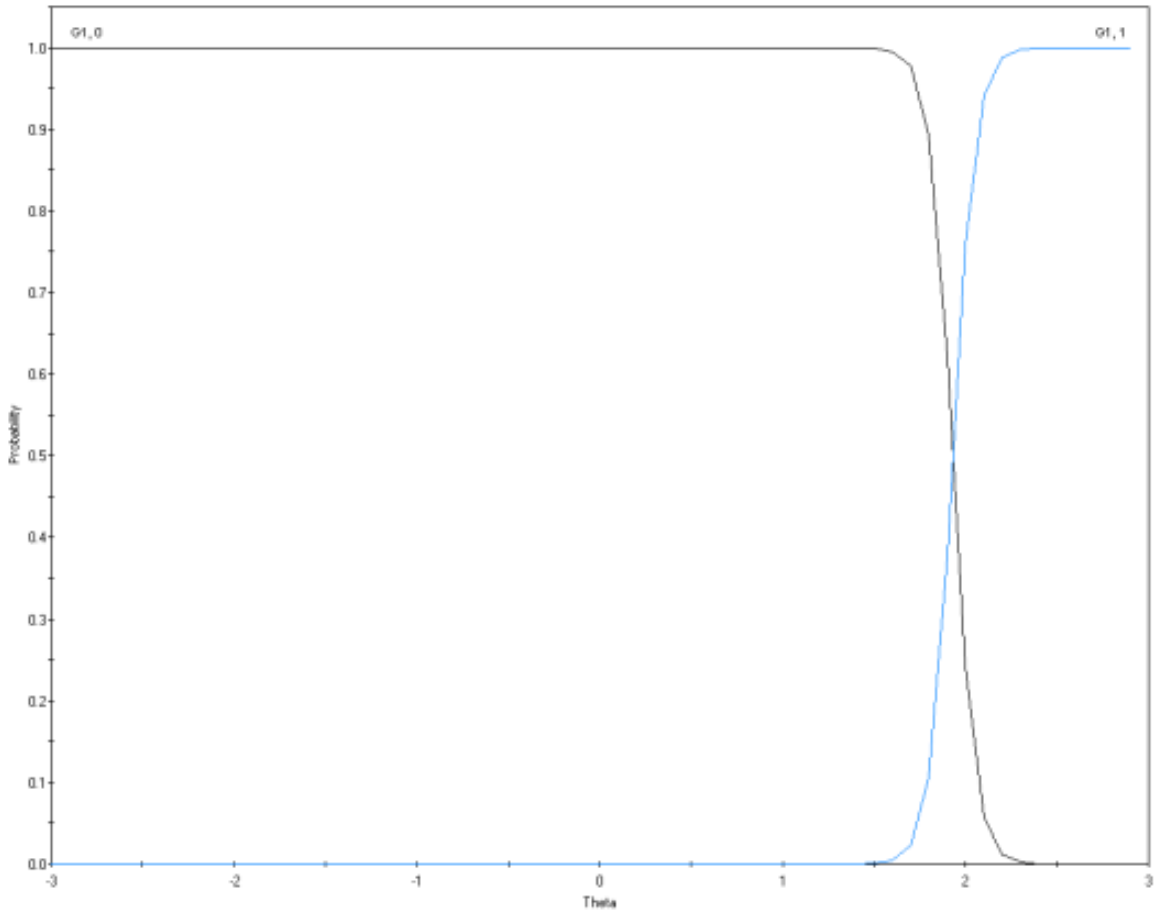


Figure 15

Test Information Curve for LSST-S

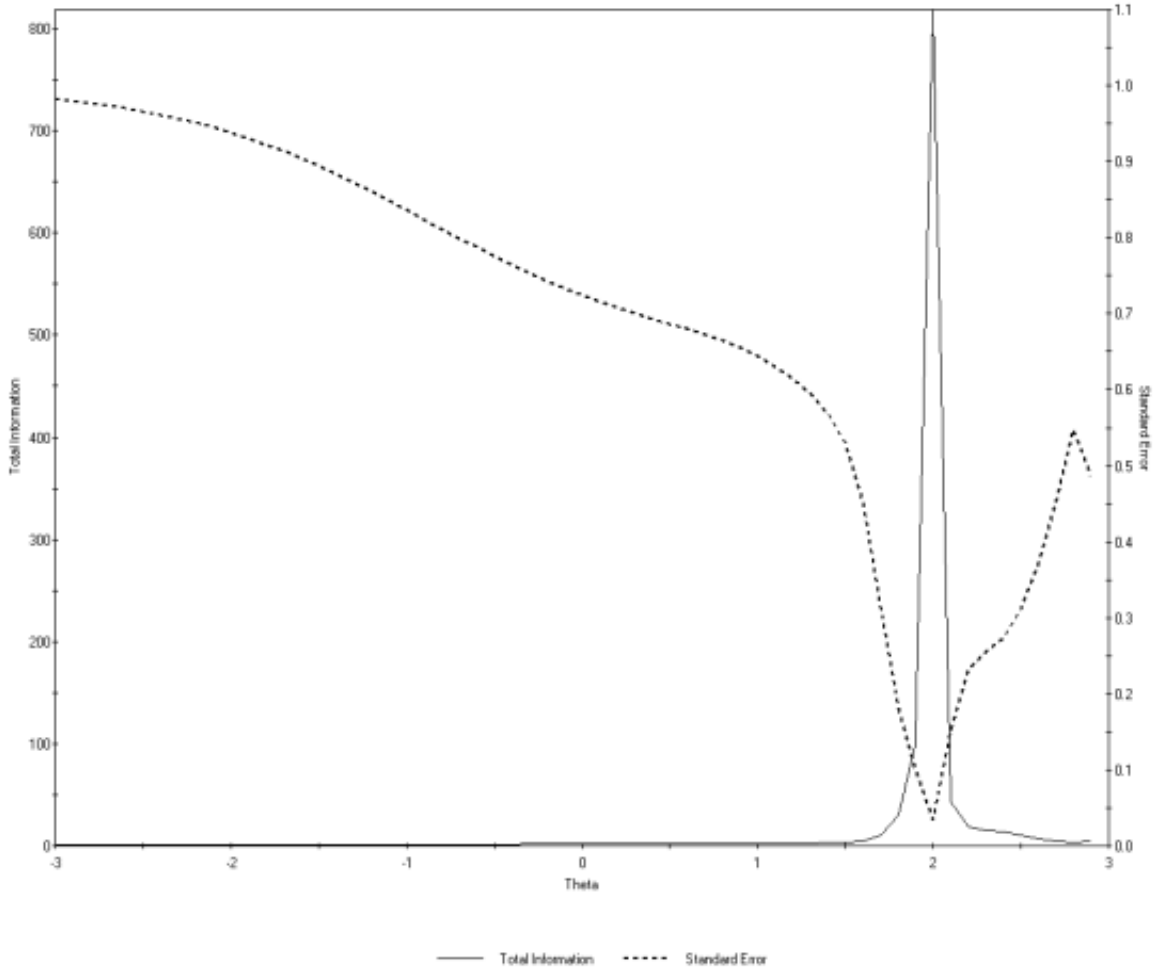
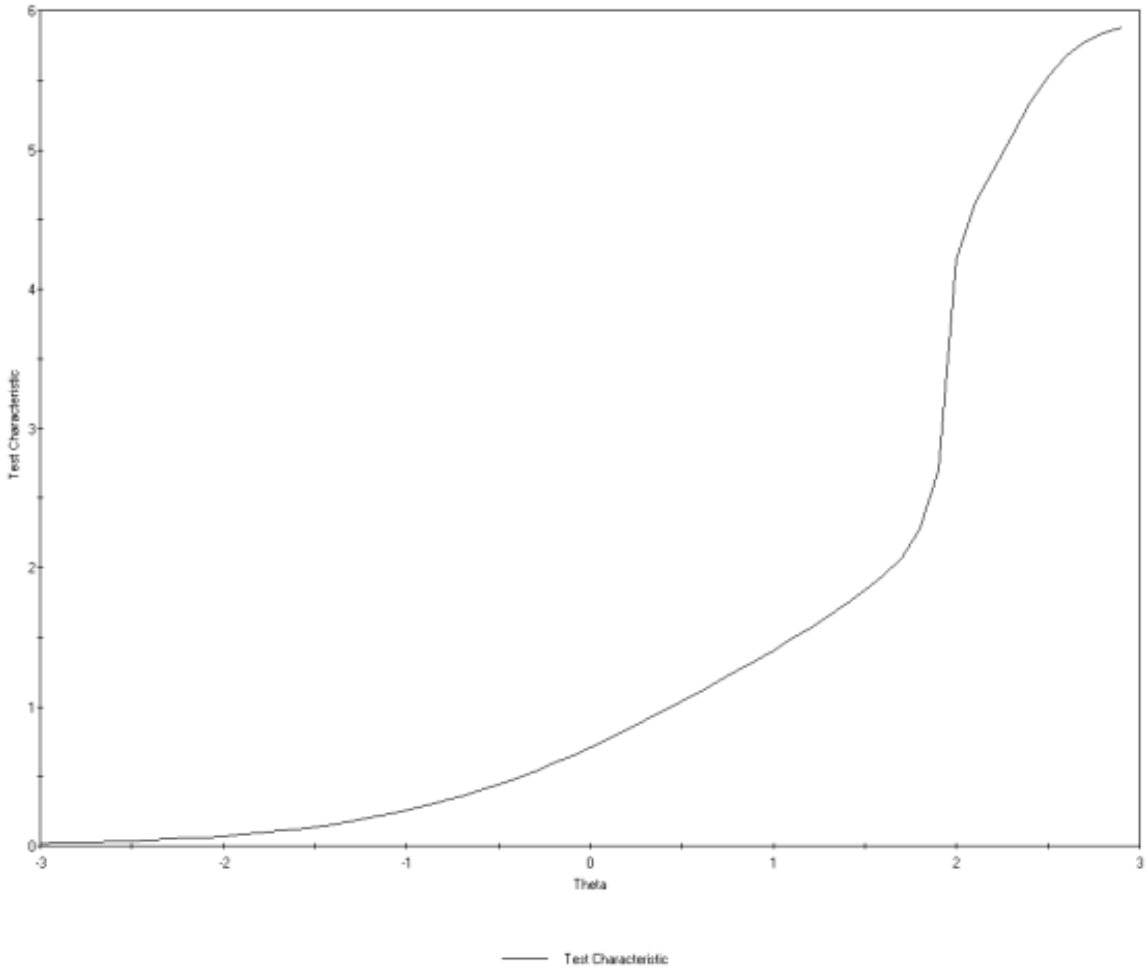


Figure 16

Test Characteristic Curve for LSST-S



Appendix A

LISTENING SENTENCE SPAN

Administer Practice Sets 1-3; Start All Children at Level 1.

DISCONTINUE when child misses one or more items OR the process question is incorrect. Record responses (even for the missed set or insertions) before moving to the next test.

Say, "**In this task I will read some sentences to you. Your job is to remember the LAST word of each sentence. First, I will read you a set of sentences. Then, I will ask you a question about one of the sentences. Then I will say 'Remember' and you are to tell me the last word of each sentence in correct order.**

So it's LISTEN, QUESTION, REMEMBER.

Let's do some practice ones first. LISTEN as I say the sentences. Then I'll ask you a QUESTION and then you REMEMBER the last word in each sentence in order. Ready for the first set?"

Examiner: Bold print is used to demarcate the correct answer and remember to say capitalized word prompts. Write student answers on the lines provided in the order spoken. Put a **CHECK MARK** on the line next to the comprehension questions if correct. If wrong comprehension answer is given, then **write in child's response**. Mark an X through the answers provided if incorrect and re-read the sentences from that set.

Practice Set 1 (provide feedback)	
LISTEN:	
1. Many animals live on the **farm**. [pause]	_____
2. People have used masks since early **times**. [pause]	_____
QUESTION: What have been used since early times?	**Masks** _____
REMEMBER?	
Practice Set 2 (Provide feedback)	

LISTEN:	
1. The baby's toy rolled under the **bed**. [pause]	_____
2. They walked around to the back of the **house**. [pause]	_____
QUESTION: What rolled under the **bed**?	**Toy** _____
REMEMBER?	
<u>Practice Set 3 (Provide feedback)</u>	
LISTEN:	
1. The squirrel hid the acorns in the hollow **tree**. [pause]	_____
2. It was so cold, the snow crunched under his **feet**. [pause]	_____
QUESTION: What crunched?	**Snow** _____
REMEMBER?	
<u>Examiner: Start all children at level 1.</u>	
DISCONTINUE when child misses one or more items OR the process question is incorrect.	
LEVEL 1: **LISTEN**:	
1. Sarah wants you to give her a **dollar**. [pause]	_____
2. Mary tried to tell her teacher the right **street**. [pause]	_____
QUESTION: Who did Mary try to **tell**?	**Teacher** _____
REMEMBER?	
LEVEL 2: **LISTEN**:	

1. The captain does not seem to have **friends**. [pause]	_____
2. Beth can't go because she didn't get **shoes**. [pause]	_____
3. Bob doesn't want to tell the **teacher**. [pause]	_____
QUESTION: Who can't go?	**Beth** _____
REMEMBER?	
LEVEL 3: **LISTEN**:	
1. My little brother went in the wrong **restaurant**. [pause]	_____
2. The teacher wanted to see me about my **book**. [pause]	_____
3. You will be sorry if you break the **window**. [pause]	_____
4. My friend wants to learn about **snakes**. [pause]	_____
QUESTION: Who will be sorry?	**You** _____
REMEMBER?	
LEVEL 4: **LISTEN**:	
1. I can study if you give me a **pencil**. [pause]	_____
2. Children like to read books about **animals**. [pause]	_____
3. I will give Cathy the sweets in a **bowl**. [pause]	_____
4. The good news gave Ann a feeling of **happiness**. [pause]	_____
5. Jeff likes to do homework in **ink**. [pause]	_____
QUESTION: What will I give to **Cathy**?	**Sweets** _____
REMEMBER?	

Appendix B

LISTENING SENTENCE SPAN

Administer Practice Sets 1-3; Start All Children at Level 1.
DISCONTINUE when child misses one or more items OR the process question is incorrect. Record responses (even for the missed set) before moving to the next test.

Say, "**En esta tarea, voy a leer algunas frases. Tú tienes que recordar la última palabra de cada frase. Primero, voy a leer las frases. Entonces, voy a hacer una pregunta acerca de una de las frases. Tú dirás la última palabra de cada frase en el orden correcto cuando yo diré RECUERDE.**

Vamos a practicar. ESCUCHE cuidadosamente a las frases. Entonces, haré una PREGUNTA acerca de las frases. Entonces, tú dirás la última palabra en cada frase en orden cuando yo diré "RECUERDE". ¿Entiendes?

ESCUCHA, PREGUNTA, RECUERDE. ¿Estás listo para empezar?"

Examiner: Bold print is used to demarcate the correct answer and remember to say capitalized word prompts. Write student answers on the lines provided in the order spoken. Put a **CHECK MARK** on the line next to the comprehension questions if correct. If wrong comprehension answer is given, then **write in child's response**. Mark an X through the answers provided if incorrect and re-read the sentences from that set.

Practice Set 1 (provide feedback)	
ESCUCHE:	
1. Muchos animales viven en la **granja** [pause].	_____
2. La gente ha utilizado máscaras desde épocas **tempranas**. [pause]	_____
PREGUNTA: ¿Qué se han utilizado desde épocas **tempranas**?	**Máscaras** _____
RECUERDE?	
Practice Set 2 (Provide feedback)	

ESCUCHE:	
1. El juguete del bebé rodó debajo de la **cama**. [pause].	_____
2. Ellos caminaron alrededor a la parte trasera de la **casa**. [pause.]	_____
PREGUNTA: Qué rodó debajo de la **cama**?	**Juguete** _____
RECUERDE?	
<u>Practice Set 3 (Provide feedback)</u>	
ESCUCHE:	
1. La ardilla escondió las bellotas en el árbol **hueco**. [pause]	_____
2. Estaba tan frío, la nieve crujia debajo de sus **pies**. [pause]	_____
PREGUNTA: ¿Qué crujió?	**Nieve** _____
RECUERDE?	
<u>Examiner: Start all children at level 1.</u>	
DISCONTINUE when child misses one or more items OR the process question is incorrect.	
LEVEL 1: **ESCUCHE:**	
1. Sara quiere que le de un **dólar**. [pause]	_____
2. Maria trato a decirle a su maestro la calle **derecha**. [pause]	_____
PREGUNTA: ¿A quién le trato Maria de decir?	**Maestro** _____
RECUERDE?	

LEVEL 2: **ESCUCHE:**	
1. Parece que el capitán no tiene **amigos**. [pause]	_____
2. Beth no puede ir porque no consiguió **zapatos**. [pause]	_____
3. Bob no quiere decirle al **maestro**. [pause]	_____
PREGUNTA: ¿Quién no puede ir?	**Beth** _____
RECUERDE?	
LEVEL 3: **ESCUCHE:**	
1. Mi hermano pequeño entró en el restaurante **equivocado**. [pause]	_____
2. El maestro quiso verme acerca de mi **libro**. [pause]	_____
3. Usted se va a arrepentir si quiebra la **ventana**. [pause]	_____
4. Mi amigo quiere aprender acerca de las **serpientes**. [pause]	_____
PREGUNTA: ¿Quién se va a arrepentir?	**Usted** _____
RECUERDE?	
LEVEL 4: **ESCUCHE:**	
1. Yo puedo estudiar si me da un **lápiz.** [pause]	_____
2. A los niños les gusta leer libros acerca de **animales**. [pause]	_____
3. Le voy a dar a Cathy los dulces en un **tazón**. [pause]	_____
4. Las buenas noticias le dieron a Ann una sensación de **felicidad**. [pause]	_____
5. A Jeff le gusta hacer la tarea en **tinta.** [pause]	_____

PREGUNTA: Que le voy a dar a **Cathy**?	**Dulces _____**
RECUERDE?	