

Nonword Item Generation: Predicting Item Difficulty in Nonword Repetition

by

Gareth Morgan

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved November 2011 by the
Graduate Supervisory Committee:

Joanna Gorin, Chair
Roy Levy
Shelley Gray

ARIZONA STATE UNIVERSITY

December 2011

ABSTRACT

The current study employs item difficulty modeling procedures to evaluate the feasibility of potential generative item features for nonword repetition. Specifically, the extent to which the manipulated item features affect the theoretical mechanisms that underlie nonword repetition accuracy was estimated. Generative item features were based on the phonological loop component of Baddeley's model of working memory which addresses phonological short-term memory (Baddeley, 2000, 2003; Baddeley & Hitch, 1974). Using researcher developed software, nonwords were generated to adhere to the phonological constraints of Spanish. Thirty-six nonwords were chosen based on the set item features identified by the proposed cognitive processing model. Using a planned missing data design, two-hundred fifteen Spanish-English bilingual children were administered 24 of the 36 generated nonwords. Multiple regression and explanatory item response modeling techniques (e.g., linear logistic test model, LLTM; Fischer, 1973) were used to estimate the impact of item features on item difficulty. The final LLTM included three item radicals and two item incidentals. Results indicated that the LLTM predicted item difficulties were highly correlated with the Rasch item difficulties ($r = .89$) and accounted for a substantial amount of the variance in item difficulty ($R^2 = .79$). The findings are discussed in terms of validity evidence in support of using the phonological loop component of Baddeley's model (2000) as a cognitive processing model for nonword repetition items and the feasibility of using the proposed radical structure as an item blueprint for the future generation of nonword repetition items.

ACKNOWLEDGMENTS

I would like to thank my wife, Meagan, and my parents, Patrick and Christina, for their unwavering support throughout this process. I love you all very much. A special thanks to my brother, Stuart, for his critical help in programming the nonword generator. You are an artist with code and I feel very lucky to have you to work with. I look forward to the next time that we can work together.

I would also like to thank my thesis committee, Drs. Joanna Gorin (Chair), Roy Levy, and Shelley Gray. Your support over the years has been nothing less than fantastic. You all do a very good job of knowing when to support me and when to push me. I greatly appreciate your mentorship and I consider myself very lucky to have been able to work with you all.

Lastly, I would like to thank Dr. Laida Restrepo for her support, mentorship, and friendship. You believe in me and my work and I always know that you are in my corner.

TABLE OF CONTENTS

| | Page |
|--|------|
| LIST OF TABLES | vi |
| LIST OF FIGURES | vii |
| CHAPTER | |
| 1 INTRODUCTION | 1 |
| 2 LITERATURE REVIEW..... | 2 |
| Automatic Item Generation | 2 |
| Examples of AIG | 6 |
| Nonword Repetition AIG..... | 8 |
| Phonological Short-term Memory | 10 |
| Nonword Repetition Processing Model..... | 11 |
| NWR Radicals | 15 |
| Item Difficulty Modeling | 17 |
| Purpose | 22 |
| 3 METHODS..... | 23 |
| Participants | 23 |
| Nonword Repetition Task | 24 |
| Procedures..... | 26 |
| Analyses..... | 27 |
| 4 RESULTS | 33 |
| Descriptive Statistics..... | 33 |
| Dimensionality | 33 |
| Rasch Modeling..... | 34 |

| CHAPTER | Page |
|---|------|
| Regression Analyses | 35 |
| LLTM Analyses..... | 38 |
| 5 DISCUSSION AND LIMITATIONS | 41 |
| A Cognitive Model for Nonword Repetition..... | 41 |
| Nonword Item Generation | 43 |
| Limitations | 47 |
| Conclusion | 50 |
| Human Subjects | 51 |
| References | 52 |
| Appendix | |
| A Planned missing data design | 96 |
| B Example nonword repetition task | 98 |
| C Sample syntax | 101 |
| D Q Matrix of Item Features | 105 |
| E Human subjects documentation | 107 |

LIST OF TABLES

| Table | | Page |
|-------|--|-------|
| 1. | Nonword Item Descriptive Statistics | 60 |
| 2. | Inter-Item Correlations | 61-63 |
| 3. | Rasch Model Item Difficulty Estimates and Fit Statistics | 64 |
| 4. | Descriptive Statistics of Item Features | 65 |
| 5. | Inter-Feature Correlations and Their Correlations with Proportion Correct | 66 |
| 6. | Regression Model Statistics | 67 |
| 7. | Regression Model Parameter Estimates | 68 |
| 8. | Rasch and LLTM Model Summary Statistics | 69 |
| 9. | Final LLTM and LLTM with Incidentals Structural Weights | 70 |
| 10. | LLTM and Rasch Item Parameter Estimates | 71 |

LIST OF FIGURES

| Figure | | Page |
|--------|--|-------|
| 1. | Baddeley's (2000) Multicomponent Model of Working Memory | 72 |
| 2. | The Phonological Loop | 73 |
| 3. | Empirical Versus Model Implied Item Characteristic Curves | 74-91 |
| 4. | Rasch Person by Item Map | 92 |
| 5. | Plot of Total Information Function and Standard Error | 93 |
| 6. | Plot of b Parameters for Rasch and Final LLTM Model..... | 94 |
| 7. | Plot of b Parameters for Rasch and LLTM Model with Incidentals.... | 95 |

Chapter 1

INTRODUCTION

Recent research efforts in test development have incorporated the application of cognitive models for automatic item generation (Arendasy & Sommer, 2007; Bejar, 2002; Embretson, 1998; Embretson & Gorin, 2001; Gorin, 2005, 2006; Gorin & Embretson, 2006; Holling, Bertling, & Zeuch, 2009). Automatic item generation (AIG) is the process of algorithmically creating items based on a specific set of features that underlie the processing needed to successfully answer an item. The algorithms needed for AIG can be created by identifying the item features that accurately predict items' psychometric properties. Accurate prediction of such properties is dependent on a comprehensive understanding of the items' response processes and identifying the *controllable* item features that correspond to the cognitive processes that represent the measured construct (Bejar, 1993). The current study employs item difficulty modeling procedures to evaluate the feasibility of potential generative item features for nonword repetition tasks. Specifically, the extent to which the manipulated item features correspond to the theoretical mechanisms that underlie nonword repetition accuracy and their impact on item difficulty is estimated. AIG model radicals are hypothesized based on the phonological loop from Baddeley's model of working memory that addresses phonological short-term memory (PSTM; Baddeley, 2000, 2003; Baddeley & Hitch, 1974). Results from this study provide evidence to support the validity argument for nonword repetition tasks as a measure of language ability and advances efforts to fully automate item generation for future research and assessments.

Chapter 2

LITERATURE REVIEW

Automatic Item Generation

Automatic item generation is an approach to test-item development where items are automatically generated based on a predetermined set of item features that are tied to the cognitive mechanisms that drive the interpretation of the responses to the items. AIG has a number of benefits. For example, some types of tests (e.g., computer adaptive tests) require large pools of items with varying levels of item parameters (e.g., difficulty) to gain precision in measurement of a person's score. Creating large pools of items using traditional methods of item writing (e.g., human item writers) is very expensive and time consuming. Automatic item generators have the potential to create an infinite number of items, in real time, once item features and their corresponding parameters have been established for generation. AIG also helps to address issues of test security. Since the cost and time investment in traditional item writing is very high, testing companies are forced to reuse items, which increases an item's exposure. Increased item exposure creates greater potential for items being compromised. AIG can generate new items that have never been seen before and can be retired after a single use with no loss to cost or efficiency; further, AIGs can be used to create parallel forms of tests. Lastly, Wainer (2002) brings up a somewhat unintended benefit of AIG, which is that if we are able to clearly identify the item features to the extent that we are able to automatically generate items, then we have also gained a better understanding of the construct that we are measuring.

Components of AIG. AIG has several requisite components, the most fundamental of which is a list of item features – characteristics of items that vary from one test question to another. For example, a set of math word problems that test end of elementary school math ability may vary in terms of the length of the word problem (e.g., how many words), the types of operations required to solve the problem (e.g., multiplication, division, addition, and subtraction), and the context in which the problem is presented. Such item features can be divided into two types: incidentals and radicals (Irvine, 2002). *Radicals* are item features that systematically impact the item’s psychometric parameters, such as item difficulty, while *incidentals* are surface features of items that do not impact item parameters. Incidentals can be used to create *isomorphs*, which are items that are psychometrically the same, but look different on the surface. In contrast, *variants* are items that differ in terms of their radicals and are psychometrically different. An example of isomorphs and variants in first grade math could be in addition problems, where the radical is the number of digits in the numbers being added together and the incidentals could be the actual numbers. Therefore, an item with two digit addition (e.g., $22 + 75$) and an item with single digit addition (e.g., $2 + 3$) would be theorized to be cognitively and systematically different and therefore would be considered variants. Two items that require single digit addition, but are adding different numbers (i.e., $4 + 5$ vs. $3 + 5$) would be considered psychometrically equivalent and therefore would be considered isomorphs. When thinking about radicals and incidentals it is always important to remember that they are in reference to a population. In other words, a radical for one population could easily be an incidental for another.

Evaluation of AIG. The AIG system can be evaluated by measuring the impact of radicals on the psychometric properties of the items. This can be done using item difficulty modeling via multiple regression analysis (Embretson, 1998, 2002) and explanatory item response models (De Boeck & Wilson, 2004), among other ways. When possible, item difficulty modeling is highly desirable in assessment development because it enables the evaluation of the construct representation since radicals are systematically related to the cognitive processes of the latent trait (Embretson, 1983). Therefore, the psychometric properties (e.g., item difficulty) can then be explained in terms of the knowledge structures and cognitive processes of the latent trait, thus extending the argument for construct validity to the item level (Embretson, 1998, 2002; Embretson & Gorin, 2001).

Once the radical structure has been validated as adequately representing the response processes of the construct, a set of rules can be created to automatically generate items. The extent to which the item generation process can be automated is determined by whether the rules that govern the generation of items allow for automatic or at least semiautomatic item generation. In some cases this may be a technical limitation, such as it is too difficult or impractical to program item generation rules into software or there may be a theoretical limitation where the item features that represent the cognitive process of the latent trait to be measured are not adequately defined.

Approaches to AIG. According to Bejar (1993, 2002) the different approaches to AIG are limited by (1) the strength of the theoretical foundation supporting the item material and (2), the extent to which the item generation

process can be automated. For example, in some cases there is no existing or accepted theoretical or cognitive model of the latent trait. In these cases, the latent trait to be measured and its salient item features used for item generation are left to be defined by psychometricians, test developers, and content area experts. This type of approach to AIG is referred to as *functional* item generation (Bejar, 2002) or a bottom-up approach (Arendasy & Sommer, 2007). Such approaches are usually considered exploratory since a number of item features may be evaluated in terms of their impact on the measured construct. One risk of using a bottom-up approach is that it can lead to inaccurate predictions of what contributes to variation in item response processes because psychometricians, test developers, and content area experts may not properly identify the item features that are responsible for the variation in the latent trait (Nathan & Petrosino, 2003). In contrast to a bottom-up approach, a top-down approach assumes and proceeds from a theoretical model of the latent trait to be measured. The role of the theoretical model is to define the latent trait in terms of the cognitive processes and knowledge structures that are utilized in the item response process (Embretson, 1983, 1994). In addition, the theoretical model connects the cognitive processes and knowledge structures to item features. The top-down is considered confirmatory and could be considered one way to test the theoretical model of the latent trait and its processes.

Ideally the development of an automatic item generator uses a top-down approach. The construct will have strong theoretical and empirical evidence to support the knowledge structures and cognitive processes that are used to create the radical structure of the items. Item difficulty modeling will be used to evaluate the radical structure. Finally, assuming the radicals are defined in a way

that enables the creation of a set of rules that can be used to automatically generate items, isomorphs and variant items will be created for use.

Examples of AIG

The application of AIG methods is evident in a few domains. The earliest efforts in applying AIG was with visual-spatial items that intended to measure fluid intelligence, such as abstract reasoning (Embretson, 1999), assembly of objects (Embretson & Gorin, 2001), hidden figures (Bejar & Yocom, 1991), and metal rotation (Bejar, 1993). These types of items lend themselves to AIG approaches because the identified radicals are highly related to the construct and they can be manipulated in a way that facilitates the creation of many variant and isomorphic items. For example, abstract reasoning items, like those of the Advanced Progressive Matrix Test (APM; Raven, 1938), can easily be manipulated to increase cognitive processing load, and therefore item difficulty, by increasing the number of rules applied in the pattern and the level of abstraction of the shapes (e.g., overlays, fusions, and distortions) (Embretson, 1999). Embretson and Gorin (2001) utilized AIG methods when generating assembly of objects items. They manipulated a number of item level characteristics (i.e., radicals) that were hypothesized to be representative of the levels of processing required to solve the items, such as the number of pieces, the total number of edges in all pieces, and the number pieces with curved edges, among others. Results revealed modest, positive, significant correlations (.20 to .473 in absolute value) between all but one of the manipulated item characteristics and item difficulty. These results supported the proposed

cognitive processing model for the assembly of objects task and provided substantive evidence of construct validity.

There are very few examples of AIG methods in the domain of verbal reasoning (Embretson & Wetzel, 1987; Gorin, 2005; Gorin & Embretson, 2006; Holling, Bertling, & Zeuch, 2009), which may be due to the multidimensionality of language, or that the item radicals and incidentals are difficult to program into software, or both. For example, when identifying the generative components for Graduate Record Examination paragraph comprehension items, Gorin and Embretson (2006) used nine predictors of item difficulty related to text difficulty (modifier and predicate propositional density, text content word frequency, percent of content words, percent of relevant text, vocabulary level of the correct response and distractors, and reasoning of the correct response and the distractors) based on the cognitive processing model of reading comprehension of Embretson and Wetzel (1987). Additional predictors of item difficulty based on the reading comprehension model of Sheehan and Ginther (2000) were also included: passage length (short = 150 words, long = 450 words) and item format (regular format or special format that was hypothesized to require additional cognitive processing). Results indicated that text encoding and the level of vocabulary used in the response options accounted for significant variability in item difficulty ($R^2 = .62$, adjusted $R^2 = .34$). More recently, Holling et al (2009) reported results on the AIG of word problems to test probability theory knowledge in university students. The authors identified seven concepts of probability theory (e.g., intersection of independent events, set union for disjoint events) as generative item features. Twenty items were then generated using text templates that allowed for variation of specific text and numbers, but otherwise maintained

the same wording across different combinations of the generative item features. Results indicated that all but one of the item generative features reached statistical significance; amount of variability in item difficulty accounted for by the generative features was not reported.

Nonword Repetition AIG

In the speech and hearing sciences, a commonly measured construct is that of phonological short term memory (PSTM). Performance on measures of PSTM is one method for identifying children with language impairment (e.g., Dollaghan & Campbell, 1998; Gathercole & Baddeley, 1990). One of the tasks commonly used to measure PSTM is a nonword repetition (NWR) task. A NWR task requires a person to listen to a pseudo word, also known as a nonword (e.g., /b æ t ɛ r æ /), and then repeat it.

When used to identify language impairment in native English-speaking children, NWR tasks have good classification accuracy (Dollaghan & Campbell, 1998; Graf Estes, Evans, & Else-Quest, 2007); however, researchers have struggled to reproduce similar results in Spanish-English bilingual children (e.g., Calderon, 2003; Gutierrez-Clelle & Simon-Cerejido, 2010; Windsor, Kohnert, Lobitz, & Pham, 2010, but see Girbau & Schwartz, 2008). Using a researcher developed NWR task, Calderon (2003) observed significant mean differences between Spanish-English bilingual children with LI and with TD; however, the differences were not large enough to be clinically useful in the accurate identification of the children with LI. In contrast, using a different researcher developed NWR task, Girbau and Schwartz (2008) observed significant mean differences between Spanish-English bilingual children with LI and with TD. The

differences were large enough to be clinically useful (82% sensitivity and 91% specificity); however, the combination of a very small sample size (11 children with LI and 11 matched typical peers) and the severity of the deficits in the children with LI may have contributed to the favorable diagnostic outcomes.

One source of variation among studies of NWR is the parameters that were used when creating nonwords for the researcher developed NWR tasks. For example, Calderon (2003) created nonwords using the following constraints: contained one to four syllables; included only infrequently occurring syllables; always followed the canonical stress pattern for Spanish (penultimate stress); had limited occurrence of later developing consonants for Spanish; and did not contain consonant clusters. In contrast, Girbau and Schwartz (2007, 2008) created nonwords using a much different set of constraints: contained two to four syllables; consisted of only medium-low frequency syllables that contained one vowel; nonwords began only with consonants; some nonwords contained consonant clusters; and the nonwords followed a number of different stress patterns. Without a systematic analysis of how nonword characteristics impact the repeatability of a nonword (e.g., item difficulty), it will be difficult to create NWR tasks that maximize the differences between children with LI and with TD. With this in mind, the purpose of the current study was to investigate the impact of nonword characteristics on nonword item difficulty and to evaluate the feasibility of AIG for NWR. A review of the literature on PSTM highlighting those processes that are likely radicals serves as a starting point. Next the item features (radicals) examined in the current study are described in terms of an experimental design. Finally, an analytic approach to AIG based on item difficulty modeling is described.

Phonological Short-term Memory

Phonological short-term memory is a temporary memory storage mechanism in which phonological information can be maintained in a ready state for use (Baddeley & Hitch, 1974). PSTM plays a critical role in language acquisition, particularly in the area of vocabulary development (Baddeley, Gathercole, & Papagno, 1998; Gathercole & Baddely, 1989; Gathercole, Service, Hitch, Adams, & Martin, 1999). From a theoretical point of view, Baddeley's model posits that PSTM provides temporary storage of unfamiliar phonological memory traces while more robust representations are being constructed (Gathercole & Baddeley, 1993). This process has been shown to play a role in vocabulary acquisition as demonstrated by scores on PSTM measures significantly predicting later vocabulary scores in young children (Gathercole & Baddeley, 1989; Gathercole, *et al.*, 1999) and in teenagers (Gathercole, *et al.*, 1999).

A deficit in PSTM has been well documented as an indicator of language impairment in children (Botting & Conti-Ramsden, 2001; Coady & Evans, 2008; Dollaghan & Campbell, 1998, Edwards & Lahey, 1998; Gathercole & Baddeley, 1989, 1990; Girbau & Schwartz, 2007, 2008; Graf-Estes, Evans, & Else-Quest, 2007; Montgomery, 1995). Gathercole and Baddeley (1990) evaluated the PSTM skills of six children with LI as compared to two typical control groups, one group was younger and matched for verbal abilities and the other group was the same age and matched on nonverbal intelligence. The children with LI performed significantly poorer on measures of PSTM (NWR and recalling word lists) than both control groups. Subsequently, a variety of studies that used NWR tasks

have reported similar results. Twenty-three such studies were reviewed in a recent meta-analysis that included 549 children with LI and 942 children with typical language (Graf-Estes, *et al.*, 2007). Across studies, children with LI scored significantly lower, on average, by 1.27 standard deviations than same age typical peers. Thus, it is evident that measures of PSTM, in particular NWR, can provide valuable information when identifying children with LI.

Nonword Repetition Processing Model

Nonword repetition is a widely accepted measure of PSTM ability (Gathercole & Baddeley, 1990; Dollaghan & Campbell, 1998) and has been closely associated with Baddeley's multi-component model of working memory (Baddeley, 2000, 2003; Baddeley & Hitch, 1974). There are three components of Baddeley's (2000) model (See Figure 1): (1) an executive control system (*central executive*); (2) visual and verbal subsystems (visuo-spatial sketchpad and phonological loop respectively) that are slaves to the central executive and provide temporary storage of visual and verbal information respectively; and (3), a long-term storage system which contains stored information and is capable of interacting with the working memory system.

NWR tasks are theorized to measure the capacity of the phonological loop. Figure 2 illustrates the flow of verbal information through the phonological loop. During the input phase verbal information is coded into a phonological representation, also called a trace, and then held in the phonological store. The trace is subject to rapid decay unless it is refreshed by the sub-vocal rehearsal process; however, the longer the trace has to be maintained by the phonological

store and sub-vocal rehearsal process, the more degraded it becomes (Baddeley & Hitch, 1974).

An assumption of Baddeley's model is that PSTM capacity is limited and that capacity is a function of the speed at which a trace decays in the phonological store and the speed at which the trace can be refreshed by the sub-vocal rehearsal process (Baddeley, 2007). Existence of trace decay is demonstrated by the word length effect, where shorter words are repeated more accurately than longer words (Baddeley, Thompson, & Buchanan, 1975); longer words take more time to recall and therefore are more vulnerable than shorter words to trace decay or forgetting. Evidence of the rehearsal process is exhibited by a reduction of the word length effect when the rehearsal process is interrupted, such as requiring a participant to say unrelated sounds (e.g., repeating the word "the", between each word that they are required to recall). Such an interruption of the rehearsal process nullifies the superior recall of shorter words over longer words (Cowan, Day, Saults, Keller, Johnson, & Flores, 1992).

Baddeley's model also assumes a connection between the phonological loop and long-term memory, which allows information stored in long-term memory to be used as supportive resource by the phonological loop during recall. Evidence of this connection comes from the lexicality effect where real words are recalled with greater accuracy than nonwords (Hulme, Maughan, & Brown, 1991; Gathercole, Pickering, Hall, & Peaker, 2001); real words have a lexical entry in long-term memory, whereas nonwords do not. Further evidence stems from the language familiarity effect where bilinguals recall verbal stimuli in

their native language with greater accuracy than verbal stimuli in their less familiar second language, which suggests a relationship between the robustness of the lexical entry in long-term memory and verbal recall accuracy (Thorn & Gathercole, 1999, 2001).

The facilitative effect of knowledge stored in long-term memory on the accuracy of short-term recall has been attributed to redintegration. Redintegration is the process of rebuilding degraded phonological traces in PSTM with information stored in long-term memory (Brown & Hulme, 1995, 1996; Hulme, Roodenrys, Schweickert, Brown, Martin, & Stuart, 1997; Schweickert, 1993). During recall when the phonological trace is accessed, the redintegration process is activated, which attempts to rebuild the partially degraded sounds of the phonological trace using information stored in long-term memory. The recall advantage is made possible when information stored in long-term memory is accessed quickly and easily, which makes the process of rebuilding a degraded trace more likely to succeed. Successful rebuilding of degraded traces leads to greater accuracy in recall, which is evidenced by real words being recalled with greater accuracy than nonwords (Hulme *et al.*, 1991). Similarly, the language familiarity effect suggests that representations in the more familiar language are accessed more readily than those in the less familiar language, thus providing a recall advantage for stimuli in the more familiar language over stimuli in the less familiar language of bilinguals (Kohnert, Windsor & Yim, 2006; Thorn & Gathercole, 1999, 2001).

Research indicates that redintegration is facilitated by two levels of information stored in long-term memory, lexical and sub-lexical (Vitevitch, 2003).

Long-term memory support provided by the lexical level is dependent on the phonological similarity at the whole word level between the target word/nonword and other real words that stored in the lexicon. This is evidenced by the effects of phonological *neighborhood density* (ND) on PSTM recall performance. ND refers to “the number of words that resemble a given word [or nonword]...by adding, subtracting, or substituting a single phoneme in that word [or nonword]” (Vitevitch, 2003, p. 487-488). Lexical level support of PSTM recall performance has been evidence by the higher recall accuracy of real words over nonwords and by the higher recall accuracy of nonwords with higher NDs over nonwords with lower NDs (De Cara & Goswami, 2002, 2003; Gathercole, *et al.*, 1999; Roodenrys & Hinton, 2001; Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002; Thomson, Richardson, & Goswami, 2005; Thorn & Frankish, 2005; Vitevitch & Luce, 1998; Vitevitch, *et al.*, 1999).

Long-term memory support provided by the sub-lexical level has been evidenced by the effects of *phonotactic probability* (PP) on PSTM recall. PP refers to “the frequency with which phonological segments and sequences of phonological segments occur in [a particular] language” (Vitevitch, 2003, p. 488); in other words, PP represents the probability that a particular sequence of sounds would occur in a particular language. During recall, redintegration seems to be facilitated by nonwords with high PP, which is evidenced by the higher recall accuracy for nonwords with high PP vs. nonwords with low PP (Gathercole, *et al.*, 1999; Munson, Kurtz, & Windsor, 2005; Roodenrys & Hinton, 2001; Vitevitch & Luce, 1998; Vitevitch, *et al.*, 1999).

In summary, evidence supports the phonological loop component of Baddeley's model (2000) as a cognitive model for PSTM. The phonological loop is assumed to have a limited capacity to store verbal information, which is evidenced by the word length effect (Baddeley *et al.*, 1975). Further, the loop is assumed to be supported by long-term memory, which is evidenced by the lexicality effect (Hulme *et al.*, 1991; Gathercole *et al.*, 2001), and the facilitative effects of PP (Munson *et al.*, 2005), ND (Munson *et al.*, 2005b). Thus, the phonological loop is a possible explanation for the construct of PSTM processing, and provides a cognitive model to be used in the current study for item development.

NWR Radicals

Based on the theoretical and empirical evidence of the current research on nonword repetition, three nonword item features have emerged as strong candidates for AIG model radicals: (1) number of syllables in a nonword; (2) phonotactic probability; and (3), phonological neighborhood density. These three item features were identified based on research that suggests a relationship between levels of the item feature and item difficulty (Coady & Evans, 2008; Gathercole, Frankish, Pickering, & Peaker, 1999; Graf-Estes *et al.*, 2007; Munson, Kurtz, & Windsor, 2005; Roodenrys & Hinton, 2001).

Number of Syllables. Researchers have observed a strong significant negative relationship between the number of syllables in a nonword and NWR accuracy (e.g., Archibald & Gathercole, 2006; Dollaghan & Campbell, 1998; Gathercole, Willis, Emslie, & Baddeley, 1992; Girbau & Schwartz, 2007; Gray, 2003; Montgomery, 1995, 2004); such a relationship makes it a good candidate

as a radical. In terms of cognitive resources, the number of syllables in a nonword relates to the amount of short-term memory required to store it; therefore, the longer a nonword is, the more short-term memory capacity is required to store it and the greater the amount of cognitive resources are required by the sub-vocal rehearsal process to maintain it. Evidence of the relationship between the number of syllables and NWR accuracy was made most apparent by a meta-analysis of 23 nonword repetition studies which observed a main effect for number of syllables and an interaction between the number of syllables and LI status (Graf-Estes *et al.*, 2007). Across TL and LI groups, NWR accuracy decreased significantly as the number of syllables increased. Further, the significant interaction indicated larger between group differences for nonwords with three to four syllables than for nonwords with one to two syllables. Based on these relationships, syllable length is a logical radical such that increases in the number of syllables of a nonword should coincide with increases in item difficulty.

Phonological Neighborhood Density. Researchers have reported a significantly positive relationship between ND and NWR accuracy (De Cara & Goswami, 2002, 2003; Gathercole, *et al.*, 1999; Thomson, *et al.*, 2005; Roodenrys & Hinton, 2001; Roodenrys, *et al.*, 2002; Thorn & Frankish, 2005; Vitevitch & Luce, 1998; Vitevitch, *et al.*, 1999), which makes ND a good potential radical. Further, ND is related to cognitive processing because nonwords with few to no neighbors provide little to no lexical support during NWR and therefore, require more cognitive resources for maintaining the nonword in PSTM. This is evidenced by the higher accuracy in recall of nonwords with high vs. low ND

(Metsala & Chishold, 2009). Thus, ND is a logical candidate as a radical such that increases in ND correspond to decreases in item difficulty.

Phonotactic Probability. Researchers have documented a positive relationship between PP and NWR accuracy (Gathercole, *et al.*, 1999; Munson, Kurtz, & Windsor, 2005; Roodenrys & Hinton, 2001; Vitevitch & Luce, 1998; Vitevitch, *et al.*, 1999). PP is related to cognitive processing because of redintegration. That is, the lower the PP of a nonword, the less likely that sub-lexical information can be used to fill in the blanks of the degraded memory trace during redintegration. This translates into the expenditure of more resources in maintaining the trace in PSTM. The positive relationship between PP and NWR accuracy has been evidenced by greater accuracy in repeating nonwords with high vs. low PP (Munson *et al.*, 2005; but see Coady *et al.*, 2010). Thus, PP is a logical radical such that increases in the PP of a nonword should correspond with decreases in item difficulty.

Experimental studies of NWR and serial nonword recall found greater accuracy in recall for nonwords with high ND over nonwords with low ND (Thorn & Frankish, 2005; Vitevitch & Luce, 1998; Vitevitch, *et al.*, 1999); however, there is a positive relationship between ND and PP, which calls into question the unique contribution of ND and PP on NWR accuracy. The relationship between ND and PP was investigated by Thorn and Frankish (2005), who found that when holding PP constant participants were significantly more accurate at recalling nonwords with high ND than nonwords with low ND. Similarly, when holding ND constant, participants were significantly more accurate at recalling nonwords with high PP than nonwords with low PP. Their results suggest that ND and PP do

uniquely contribute to nonword recall accuracy and thus should uniquely contribute to item difficulty.

Item Difficulty Modeling

In AIG development, the item features (i.e., radicals and incidentals) are hypothesized to represent the underlying cognitive processes of the measured construct or latent ability; however, the utility of the item radicals must be empirically evaluated by applying statistical models that can estimate the impact of such variables on the psychometric properties of the items. Based on the presumption that AIG radicals determine the cognitive and therefore psychometric properties of an item, statistical techniques that estimate the relationship between radicals and item difficulty, discrimination, and response time are appropriate. Estimating the relationship between item features (i.e., radicals) and estimates of item difficulty is known as *item difficulty modeling* (IDM). A common classical test theory based IDM approach uses regression techniques in which the proportion of correct responses to an item is regressed on the item characteristics. Such an analysis will allow a researcher to estimate the percent of variance explained in the proportion of correct responses by the set of item radicals. Further, partial correlations of the regression parameter estimates can be used to assess the individual contributions of each radical.

The utility of regression techniques for IDM may be limited (Embretson & Daneil, 2008; Daniel & Embretson, 2010). As explained by Embretson and Daniel (2008), classical multiple regression techniques replace the participants item responses with item level statistics, which in some cases can drastically reduce the sample size. The sample size reduction can result in large standard errors.

Since standard errors are used in determining the significance of the predictors (e.g., $t = \text{estimate}/\text{standard error}$), large standard errors will result in reduced power, which can limit the interpretability of the impact of the manipulated variables in the model.

An alternative to modeling aggregated item level statistics is to model the raw item response data via *item response theory* (IRT) based methods. In IRT, individuals' responses to test items are used to simultaneously estimate their level of the latent trait and the items' psychometric properties. A benefit to this approach is that you can harness the power of the whole sample size instead of having to aggregate, as you do when using classical test theory methods (Embretson & Daniel, 2008). There are several other benefits to using IRT that relate more to measurement precision and the interpretation of estimates (see Embretson & Reise, 2000); however, in terms of IDM, the benefit of IRT methods is the power to explain what drives the psychometric properties (e.g., item difficulty). The linear logistic test model (LLTM: Fischer, 1973) comes from a branch of IRT models called explanatory item response models, which integrate item content into the prediction of responding to an item correctly. If suitable item content features can be identified for each item, then parameters that correspond to the impact on item difficulty can be estimated directly using the LLTM. This is an advantage over classical IDM, which estimates the impact on item difficulty indirectly by using a series of separate analyses and aggregated data.

The LLTM is an extension of the Rasch model, one of the most basic unidimensional IRT models (Rasch, 1960). The Rasch model, or the one

parameter logistic model, assumes a logistic distribution and predicts the probability of success for person j on item i (i.e., $P(X_{ij} = 1)$), as follows:

$$P(X_{ij} = 1 | \theta_j, \beta_i) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)} \quad (1)$$

where θ_j is the latent ability of person j , and β_i is the difficulty level of item i . The logit of equation 1, $\theta_j - \beta_i$, is the difference between the person's latent ability level and the item difficulty; further, the antilog of $\theta_j - \beta_i$ is equal to the probability of success.

In the LLTM, item difficulty (β_i) is substituted with a linear model of item difficulty. Items are scored on the product of their characteristics, q_{ik} , which is the score of item i on characteristic k in the cognitive model of the items, and an estimated weight of characteristic k , η_k . The probability that person j passes item i , $P(X_{ij} = 1)$ is given as follows:

$$P(X_{ij} = 1 | \theta_j, q, \eta) = \frac{\exp(\theta_j - \sum_{k=1}^K q_{ik}\eta_k)}{1 + \exp(\theta_j - \sum_{k=1}^K q_{ik}\eta_k)} \quad (2)$$

where q_{i1} is 1 and η_1 is an intercept. As mentioned, there is no parameter for item difficulty because it is predicted from a weighted combination of item characteristics that represent the cognitive complexity of the item.

There are several advantages to using the LLTM for test design. First, a test blueprint can be created based on the item characteristics that have empirical support for predicting the cognitive complexity of the construct (Embretson, 1998). Second, construct validity is explained at the item level such that the relative weights of the item characteristics represent the level of cognitive complexity that the item is measuring. In other words, the relative weights

describe the strength of relationship between the item's characteristics and the item's difficulty, thus presenting the opportunity to provide substantive support of construct validity for the measured construct (Messick, 1995). Third, IRT models measure item psychometric properties and person ability on a common scale, which allows for inferences about a person's performance on specific item types to be linked to score interpretations (Embretson & Reise, 2000, p. 27); however, LLTM models take this one step further because the probability of answering an item correctly is linked to and explained by the different sources of cognitive complexity in the items (Embretson & Daniel, 2008). Thus, it is evident that there are a number of advantages for selecting the LLTM as a measurement model to evaluate item characteristics as generative features for AIG. Though of lesser importance from a measurement development perspective, the LLTM model parameterization can also be viewed as an empirical test of a theoretical model of a construct.

Recent application of the LLTM model to measures of individual abilities has increased substantially (Daniel & Embretson, 2010; Embretson & Daniel, 2008; Embretson & Gorin, 2001; Gorin, 2005; Holling et al., 2009; Ivie & Embretson, 2010). Ivie and Embretson (2010) utilized the LLTM for IDM in the domain of spatial ability by expanding on the work of Embretson and Gorin (2001) with the assembly of objects task and an evaluation of a three-stage, top-down cognitive processing model: Encoding, Falsification, and Confirmation (Embretson & Gorin, 2001). A number of item characteristics that corresponded to each stage of the processing model were evaluated as to their impact on item difficulty. Results of the LLTM nested model statistics indicated that all three levels of the cognitive processing model contributed significantly to item difficulty.

Further, of the original 13 item characteristics, seven contributed significantly to item difficulty and represented all three levels of cognitive processing. Embretson and Daniel (2008) used the LLTM to understand and quantify the cognitive complexity of mathematical word problems on the Graduate Record Exam (GRE). Items were selected from an item bank of released GRE mathematical word problems and coded for 12 item features (e.g., number of knowledge principles or equations to be recalled, generating unique equations, number of sub-goals, and number of computations) that represented four stages of cognitive processing for mathematical word problem solving (problem translation and integration, solution planning, and decision). Results from the LLTM analysis indicated significant contribution of all item difficulty predictors which suggests positive support for the validity the proposed cognitive processing model. Meaning that, the manipulations of key variables in the cognitive processing model lead to changes in item difficulty. Thus, it is evident that the LLTM can be successfully used to model the impact of item characteristics on item difficulty using different types of items measuring different types of constructs.

Purpose

The purpose of the current study was to evaluate the extent to which the identified item radicals (number of syllables, PP, & ND) represented the theoretical mechanisms that underlie NWR accuracy. The primary research question evaluated the proposed cognitive processing model of nonword repetition as an accurate representation of the underlying mechanisms of correctly repeating a nonword. This question was addressed by examining the overall fit of the data to the cognitive model via the LLTM parameterization. In

addition, the effects of the individual AIG radicals were evaluated based on their contribution to the overall model. The practical implications of this research were to better understand PSTM as a construct, NWR as a task and its potential for the identification of LI in Spanish-English bilingual children.

Chapter 3

METHODS

Participants

This study was part of a larger study designed to develop a screening measure for LI in Spanish-speaking children. A sample of two-hundred and fifteen Spanish-English bilingual children was selected from the larger study. This sub-sample was selected because they were part of the first round of data collection when all items were administered; subsequent rounds of testing only collected data on subsets of the items. Ages ranged from five to seven years old with a mean age of 6.24 years ($SD = .67$). Nearly 50% of the sample was identified as language impaired by the larger study using the following measures: a parent report survey of language use and concern for LI (Restrepo, 1998); a standardized nonverbal scale - Kaufman Assessment Battery for Children, second edition (Kaufman & Kaufman, 2004); a Spanish-English language proficiency scale (Smyk, Restrepo, Gorin, & Gray, 2009); the Spanish Clinical Evaluation of Language Fundamentals, fourth edition (Wiig, Semel, & Secord, 2006); and the Structured Photographic Expressive Language Test, third edition (Dawson, Stout, & Eyer, 2003). All of the children who participated were recruited from elementary schools in a large metropolitan area in central Arizona. Parents reported that all children spoke a Mexican dialect of Spanish. Qualification for free or reduced lunch and mother's level of education were used as indirect measures of socio-economic status; ninety-six percent of the children in the sample qualified for free or reduced lunch, 7% of mother's had a college degree, 61% had a high school diploma, and 35% had only completed primary school.

Nonword Repetition Task

Nonword generation and AIG radicals. To create the Spanish NWR task a Spanish nonword generator (Morgan & Morgan, in preparation) was developed to randomly generate nonwords that adhered to the phonological rules of Spanish. Measures of PP and ND were calculated for each nonword in Spanish using LEXESP. LEXESP is a Spanish word frequency dictionary that includes a list of over 100,000 words and their corresponding frequency count from a five million word Spanish corpus (Sebastián-Gallés, Martí, Cuetos, & Carreiras, 2000). Davis and Perea (2005) modified the LEXESP by removing foreign words adopted by the Spanish language that do not conform to Spanish phonotactic rules. Further, duplicate entries and words with diacritics, such as hyphens, were also removed as these types of entries can influence phonotactic statistics such as PP and ND (Davies & Perea, 2005). Thus, using formulas of PP, measured by biphone frequency, and ND obtained in the literature (Storkel, 2004), PP and ND were calculated for each nonword using the modified LEXESP dictionary from Davis and Perea (2005). The nonword generator also checked the generated nonwords against a dictionary of real Spanish words (Davies & Perea, 2005) to ensure that no nonword was a real Spanish word. In addition, native Spanish-speakers with five different dialects (Mexican, Colombian, Peruvian, Venezuelan, & Castilian) reviewed all of the Spanish nonwords used in the experiments to determine if they sounded like real words. If they did, those nonwords were cut.

To create the Spanish NWR list a sample of 5,000 Spanish nonwords and their PP and ND were generated at each syllable length (3-5) to obtain stable estimates of the mean and standard deviation of the nonwords at each syllable

length. To control for and investigate the possible impact of PP and ND on item difficulty, four categories of nonwords at each syllable length were selected: (a) high PP-high ND, (b) high PP-low ND, (c) low PP-high ND, and (d) low PP-low ND. Since PP and ND are continuous variables the upper and lower quartiles of the PP and ND distributions were used as cutoffs for high and low PP and ND. Stricter cutoffs, such as the upper and lower 15%, may not have allowed for nonwords in the high PP-low ND, low PP-high ND categories as PP and ND are positively correlated (Storkel, 2004). Thus, PP was considered 'high' if the PP of the nonword fell above the 75th percentile of the PP distribution for the given syllable length of the nonword. Alternatively, PP was considered to be low if the nonword fell below the 25th percentile of the PP distribution for the given syllable length. The same cutoff rules that were used for PP were also used for ND. Three nonwords in each of the four categories at each syllable length were selected and reviewed by three native Spanish speakers to ensure that they were not real words in Spanish.

Nonword audio recording. Evidence shows that nonword duration time is negatively correlated with performance (Lipinski & Gupta, 2005); however, Spanish is a syllable-timed language, meaning that all syllables have similar duration (Whitley, 2002, pp. 71-72). Thus, nonwords of similar syllable length will have similar duration times. Penultimate stress is the most common stress pattern in Spanish (Whitley, 2002, pp. 69); however, penultimate stress in certain nonwords would sound unnatural. Therefore, the three native speakers were asked to repeat the nonword aloud and the stress pattern that was agreed upon by at least 2 out of 3 native speakers was used for each word when recording the nonwords.

All Spanish nonwords were recorded digitally as .wav files by a native Spanish-speaking female using a USB headset microphone (Cyber Acoustics AC-850) with Adobe® Audition 1.5. One-half second of silence was added to the beginning and 2.5 to 3.0 seconds of silence were added to the end of each nonword so that all nonword .wav files were five seconds in length. Silence at the beginning and end of the nonword .wav files equalized administration time for each nonword. Directions were recorded by the same native speaker who recorded the nonwords. The nonwords were put into playlists using iTunes and then downloaded onto iPod Nanos for administration.

Administration design. A planned missing data design (Appendix A) was used to create three forms of the NWR task where each participant repeated 24 of the 36 nonwords. Each form contained 24 nonwords with two items per category (high PP high ND, high PP low ND, low PP high ND, low PP low ND) per syllable length (3, 4, & 5). Within each syllable length the nonwords were pseudo-randomized to reduce any order effects and each list was presented in an order of increasing syllable length; an example form is provided in Appendix B. All forms were presented equally across the entire sample and forms were randomly assigned to participants.

Procedures

The nonwords were presented using iPod Nanos and the participant's responses were audio recorded using digital voice recorders with headset microphones. During administration, the participants were told that they were going to hear a funny language and the tester wanted to see how well they could repeat the words in that language. The participants listened to a set of

instructions in Spanish which was followed by a set of prompts that trained the participant to the task by practicing to repeat three nonwords. During the practice session, a trained administrator gave the participant feedback such as the level and clarity of the participant's voice and the speed at which the child was repeating the nonwords.

NWR scoring. Audio recordings of the participants' repetitions were scored by trained Spanish speakers. Items were scored dichotomously where a 1 indicated that the participant repeated the nonword with 100% accuracy and a 0 indicated any amount of errors in repetition. Ten percent of the audio recordings were scored twice and checked for inter-scorer reliability by the author who is Spanish-English bilingual proficient. Percent agreement was .963.

Analyses

Three phases of analyses were conducted. *Phase I* consisted of classical item-level statistics such as item descriptive statistics (i.e., means and variances) and inter-item correlations. Additionally, *Coefficient Alpha*, a measure of internal consistency or reliability, was estimated for each of the NWR administration forms using SPSS 19. The missing data that was created by the planned missing data design prohibited the estimation of reliability with internal consistency estimates across all administration forms; however, the different types of nonwords had equal representation across all forms and each form had the same number of items.

Phase II consisted of a dimensionality assessment of the items, fitting the items to a Rasch model and evaluating the items fit to the model. An assumption

of the unidimensional dichotomous Rasch model is that the items measure a single common construct (Embretson & Reise, 2000); therefore, the unidimensionality assumption was assessed prior to item parameter estimation. The unidimensionality of the set of items was assessed by conducting a confirmatory factor analysis (CFA) using full information maximum likelihood estimation in MPLUS 6.0; sample code is presented in Appendix C. It is suggested that the fit of CFA models be assessed using at least one fit index from each class (parsimony, absolute, and comparative; Brown, 2006; Yu & Muthén, 2002). Thus, the CFA model fit was assessed using the following methods: model chi-square where a $p > .05$ would indicate that the model estimates adequately reproduced the sample variances and covariances; the weighted standardized root mean squared residual (WSRMR) where values of less than 1.00 are considered adequate; the root mean squared error of approximation (RMSEA) where a value of less than 0.05 is considered adequate; and the comparative fit index (CFI) where a value of greater than 0.95 is considered adequate (Brown, 2006; Yu & Muthén, 2002). In addition, localized areas of strain in the model were evaluated using the standardized item residuals where values of less than the absolute value of 1.96 are considered adequate; items with standardized residuals larger than 1.96 were considered for removal.

Rasch Model Estimation. The remaining items from the CFA were fit to a Rasch model using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996); sample code is presented in Appendix C. BILOG uses marginal maximum likelihood (Bock & Aitkin, 1981) estimation via the EM algorithm (McLachlan & Krishnan, 1997) and incorporates a Bayesian framework (Mislevy, 1986). The analysis was conducted using the RASCH calibration routine in BILOG-MG,

which fixed the discrimination parameter for all items to one and scaled the item difficulty parameters to a mean of zero and standard deviation of 1.

Item Fit. Items were evaluated as to their fit to the Rasch model using the weighted mean square fit statistics (Wright & Masters, 1991) called infit, which is the ratio of the observed residuals to the expected residuals. When the ratio is close to one, the observed residuals are varying similarly to what is expected; thus, a weighted mean square that is close to one is desired. Mean square values above or below one indicate that the items are varying more or less than expected. Items with a mean square greater than one contribute less in terms of the overall estimation of the latent variable and tend to be questionable. Adams & Khoo (1996) suggested that .75 and 1.33 are reasonable lower and upper bounds of the weighted mean fit statistic for infit. In addition, the fit of the data to the Rasch model was visually evaluated comparing empirical versus model implied item characteristic curves (ICC). An ICC is the plot of the probability of answering an item correct as a function of θ . ICCs of the empirical data from the examinee's responses can be plotted against ICCs of the model implied parameters. The extent to which the empirically derived probability values fall within the 95% confidence interval of the model implied probability values indicates better or worse item fit. Items that do not demonstrate adequate fit to the Rasch model were considered for removal before proceeding to Phase III.

Item Distribution and Information Curves. For the current study, the items were constructed to represent a distribution of items that spanned from easy to very hard. Information curves can be used to evaluate the distribution of the items with respect to the amount of psychometric information that they provide along points on the latent continuum. More specifically, an *item* information curve

is the plot of the psychometric information that an item provides at any given value of θ_i , where the peak of the item information curve can be interpreted as the point on the latent continuum where that item provides the most information or the best discrimination between latent abilities. *Test* information is the sum of all of the items information at a given value of θ and the height of the test information function at each level of θ_i indicates the level of reliability of the items at that level of θ_i . The inverse of test information is the standard error of measurement (Embretson & Reise, 2000, p. 15) and is a measure of reliability in IRT. Since reliability in IRT is a function of θ we can be more or less confident about a person's test result based on their latent ability θ and the level of test information or standard error on that point in the latent continuum. In the current study it was desirable to observe a tall but wide test information function, because items were generated to target a wide range of participant abilities on the latent continuum.

Phase III provides the results that are central to the research questions and consist of evaluating the empirical model of item difficulty to the theoretical cognitive model. Further, the individual contribution of the item attributes (syllable length, phonotactic probability, and neighborhood density) as predictors of item difficulty were also evaluated. Descriptive statistics of the item attributes were estimated, followed by a series of regression models where the item characteristics and the two interactions (PP by number of syllables and ND by number of syllables) were used to predict item difficulty (proportion correct per item for the total sample). The regression models evaluated all combinations of adding the three item characteristics and the two interactions incrementally as predictors of item difficulty. The results of the regressions were used to get an

initial sense of the relationships between the item characteristics and item difficulty. Less attention was paid to tests of statistical significance and more was paid to the strength of the relationships between the predictors and the dependent variable.

LLTM Estimation. Conditional maximum likelihood estimates of the LLTM parameters from the raw item data were estimated using the eRm package for R (Mair, Hatzinger, & Maier, 2010); syntax is presented in Appendix C. A model comparison approach was used where the item characteristics were added one at a time thus creating nested models. These nested models were evaluated using a chi-square difference test of the -2 log likelihoods and the Akaike Information Criterion (AIC; smaller values indicate better fit). A correlation (r) and multiple-correlation squared (R^2) was calculated for each LLTM model to evaluate the relationship of the LLTM predictions of item difficulty to the Rasch model item difficulties. The R^2 can be interpreted similarly to regression where a predictor or set of predictors is said to explain a percent of variance in the dependent variable. The R^2 was used to evaluate the explanatory power of the overall model and to evaluate the individual contribution of each predictor.

Next the cognitive model coefficients (η_j) were evaluated based on their p -value, magnitude, and direction; the p -values and η_j weights can be interpreted as they would in regression where the p -values denote a statistically significant relationship between the predictor and the dependent variable and the η_j weights denote the direction (positive vs. negative) and the strength of that relationship when all other variables in the model are present. As with multiple regression,

multi-collinearity among the predictors may make the interpretation of the direction and magnitude of the parameter estimates less feasible.

Chapter 4

RESULTS

Descriptive Statistics

Thirty-six Spanish nonword items were administered to 215 Spanish-English bilingual children with TD and with LI. As expected, the children with TD had a significantly higher mean score than the children with LI, $F(1, 214) = 19.90$, $p < .01$, $\eta^2 = .01$. Item level proportion correct for the sample ranged from .13 (SD = .33) to .92 (SD = .27); 64% of the items had proportion correct values that fell within .30 to .70 accuracy. In addition, average proportion correct was calculated for each form: Form 1 ($\mu = .55$, SD = .20, $n = 65$); Form 2 ($\mu = .59$, SD = .22, $n = 77$); Form 3 ($\mu = .48$, SD = .20, $n = 78$). There were no perfect scores or zero scores for Form 1, two perfect scores and no zero scores for Form 2, and one perfect score and no zero scores for Form 3; descriptive statistics for all 36 items are reported in Table 1. Inter-item correlations were calculated using the raw dichotomous data and ranged from -.28 to .55 (See Table 2); 66% of the items had acceptable biserial correlations with the total score ($r_{\text{bis}} > .20$; See Table 1). Coefficient alpha was computed for each of the three administered forms and ranged from .78 to .85; coefficient alpha was not calculated across form due to the planned missing data design.

Dimensionality

Results from the unidimensional confirmatory factor analysis indicated that the data fit a single factor model well given the following fit statistics: chi-square test of model fit, $\chi^2(594) = 614.74$, $p = .27$; RMSEA = .01 with a 90% confidence

interval of 0 to .018; CFI = .988; and WRMSR = .90. Thirty-five items had statistically significant factor loadings at an alpha level of .01; standardized factor loadings ranged from .30 to .75; results did not indicate any areas of localized strain.

Rasch Modeling

The data were assessed as their fit to a Rasch model using BILOG-MG, -2 log likelihood = 9808.07, AIC = 9897.92; sample code is provided in Appendix C. The difficulty parameter estimates ranged from -2.56 to 2.58; item parameter estimates and fit statistics are reported in Table 3. Visual inspection of the ICCs indicated that seventeen items had three or more empirical values falling within the 95% confidence interval of the Rasch estimates; fifteen items had two empirical values falling within the 95% confidence interval of the Rasch estimates; four items had only one empirical value falling within the 95% confidence interval of the Rasch estimates; and item x15 had no empirical values falling within the 95% confidence interval of the Rasch estimates (See Figure 3 item ICCs). Based on a visual inspection of the ICCs, it seemed that the misfit was due to the constraint of the slope parameter to one. Many of the items that had only two empirical values fall within the 95% confidence interval appeared to have much flatter slopes than the defined slope of 1.00 in the Rasch model. Across all items, it did not appear that the misfit was due to guessing or the lower asymptote. A person-item map plot presents the distribution of person thetas and item difficulty parameters on the same latent scale. The person-item map in Figure 4 shows that the distribution of item difficulty parameters and the person thetas are slightly positively skewed; however, there seems to be a good spread

of item difficulties and person abilities ranging from negative two to positive two. As a result of perfect or zero scores, a few person theta parameters were estimated as higher or lower than the difficulty of all of the items which can create estimation problems; however, perfect and zero scores can be handled the marginal maximum likelihood estimation procedure used by BILOG-MG and were given finite theta estimates. Figure 5 depicts the total test information function and the standard error of measurement. The standard error of measurement ranged from around .05 to .2 over the middle quartiles of the information distribution.

Regression Analyses

A series of multiple regression analyses were conducted to initially evaluate the relationships between item features and item difficulty. Although the each item was created based on a set of categorical features (e.g., high-PP and low-ND), the actual values of the item features were used for the analyses (See Appendix D)The correlations between the item features ranged from -.36 to .55 and ND had the highest correlation (.53) with the dependent variable (See Tables 4 and 5). The first set of regression models included the three hypothesized radicals only; the second set of models included main effects and two 2-way interactions. Table 6 lists the results of the model comparisons and Table 7 lists the results of the best fitting model. Predictors in the best fitting model included ND and PP and accounted for 32% of the variance in item difficulty. Coefficients for the best fitting model indicated a significant positive main effect for ND; PP was retained in the model ($p = .06$) because it substantially increased the R^2 .

Surprisingly, the radical *number of syllables* was not a significant predictor ($p = .15$).

Additional Regression Analyses. The initial regression analyses left a substantial amount of variance unexplained. Therefore, additional features were considered in an attempt to explain more variance in item difficulty. The first additional item feature was the presence or absence of consonant clusters or consonant blends in the nonword. A consonant cluster is two or more consecutive consonants (e.g., *cr* or *str*); only consonant clusters that occurred within a syllable were considered. In Spanish, there are only 13 legal consonant clusters, the maximum number of consonants in row is two, and they only occur in the initial and medial positions of a word (Whitley, 2002). The presence of consonant clusters may add an additional level of difficulty that could be explained within the cognitive model because consonant clusters are an additional source of phonological complexity in words. Consonant clusters may also be a source of construct irrelevant variance if the participants have yet to acquire them; however, this is unlikely as the acquisition of consonant clusters in Spanish starts as early as 1;1 (years; months) in the initial position and 1;5 in the medial position. In addition, evidence suggests that the rate of occurrence of cluster reductions – a common type of error whereby the consonant cluster is reduced to one of the consonants – drops below 10% by five years of age in typically developing English-speaking children (Roberts, Burchinal, & Footo, 1990). In the current set of nonwords, thirteen of the 36 nonword had at least one consonant cluster and the pearson correlation between the presence of consonant clusters and item difficulty was moderate to strong, $r = -.62$, and significant $p < .01$.

The second additional coded item feature was the number of phonemes or sounds in a nonword. The item feature number of phonemes was examined in that it could provide a finer grain measure of nonword length than number of syllables. Counting the number of phonemes also accounts for the contribution of additional sounds provided sound blends, such as consonant clusters, which the variable number of syllables does not (e.g., *pato* vs. *plato*). The Pearson correlation between the number of phonemes and item difficulty was strong, $r = -.74$, and significant, $p < .01$; the average number of phonemes per nonword was 8.56 (SD = .49).

A second series of regression analyses was conducted to evaluate the explanatory contribution of the additional item features. Similar to the previous regression analyses, a model building approach was used where independent variables were entered into the regression equation one at a time and evaluated for their contribution to the model. The best fitting model from the previous set of regression analyses was used as the base model upon which the additional predictors were added; model comparison results are reported in Table 6 and regression estimates of the new best fitting model are reported in Table 7. Results of the final model indicated that the additional predictor, number of phonemes, significantly contributed to increased explained variance in item difficulty, adjusted $R^2 = .60$, change in $F(1,32) = 23.44$, $p < .01$. Though consonant clusters did not significantly explain more variance as compared to the previous model; its small p -value (.10) and moderately strong strength of relationship (-.619) with the dependent variable suggests that it should be retained for LLTM analyses. Overall, the final model, which included the variables

ND, PP, and number of phonemes, accounted for 60% of the variance in item difficulty.

LLTM Analyses

LLTM Estimation. Results of the Rasch and LLTM models are reported in Table 8. The results indicated that the Rasch model was the best fitting model, -2 log likelihood = 2501.39, AIC = 2591.24 and the best fitting LLTM model was Model 6, -2 log likelihood = 2691.19, AIC = 2697.30. The three predictors in the best fitting LLTM model included, number of phonemes, PP, and consonant clusters; all were significant with p -values less than .01. The parameter estimates of LLTM Model 6 and their respective statistics are reported in Table 9.

The contribution of each radical was evaluated by its R^2 when it was the only predictor in the model; in addition, its squared partial correlation was also calculated and describes the individual contribution of the predictor when controlling for all other predictors (radicals) in the LLTM model. Results indicated that the item radical *number of phonemes* accounted of the greatest amount of variance in item difficulty ($R^2 = .62$) when it was the only predictor in the model and it had the largest squared partial correlation LLTM model 6 (*partial* $R^2 = .67$).

LLTM Model Fit. Results indicated strong fit of the LLTM model predicted values when compared to the Rasch model. The LLTM predicted item difficulty parameter estimates of the best fitting LLTM model (Model 6) were highly correlated with the Rasch item difficulty parameter estimates ($r = .83$). Further, an R^2 of .70 indicated that the set of item radicals for Model 6 accounted for 70% of the variance item difficulty. Correlations and R^2 s between the Rasch item

difficulty parameter estimates and the LLTM predicted item difficulty parameter estimates are reported in Table 8.

The relationship of the LLTM predicted item difficulties to the Rasch estimated item difficulties was also evaluated visually. After rescaling both sets of item difficulties to have a mean of zero and standard deviation of one, the LLTM difficulty parameter predictions were plotted against the Rasch difficulty parameters estimates (See Figure 6); see Table 10 for the rescaled item difficulties for LLTM Model 6 and the Rasch model. Each of the numbers on the scatter plot represents an item and the proximity of the number to the diagonal line indicates the precision at which the LLTM was able reproduce the item difficulty parameter estimates of the Rasch model. As can be seen from the scatter plot, many of the numbers are falling on or near the diagonal line, which corroborates the numerical results presented above.

Evaluation of Incidentals. Three incidental item characteristics were identified and coded: vowel as beginning sound, vowel as ending sound, and the inclusion of a late acquiring sound. These incidentals were chosen because within the assumptions of the cognitive model they did not have a strong association with item difficulty; see Table 4 for their descriptive statistics and Table 5 for correlations among all of the item features and their correlations with the proportion correct. Twelve items had a vowel as a beginning sound, twenty-one items had a vowel as an ending sound, and eight items had vowels as both beginning and ending sounds. All of the items had late acquiring sounds for Spanish, so that incidental was not analyzed. The remaining two incidentals were added as a set to the final LLTM model to evaluate their impact on item difficulty

above and beyond that of the item radicals. Results indicated that the LLTM model with incidentals (LLTM Model 7) fit significantly better than the LLTM model without them (LLTM Model 6), $\chi^2(2) = 48.82$, $p < .01$, $R^2 = .78$; model statistics for LLTM Model 7 are reported in Table 8. Both incidentals were significantly negatively associated with item difficulty; parameter estimates for LLTM Model 7 and their respective statistics are reported in Table 9. After rescaling the item difficulties to have a mean of zero and standard deviation of 1, the predicted item difficulties from the LLTM model with the incidentals were plotted against the Rasch model (see Figure 7); the rescaled item difficulties for the LLTM model with the incidentals are in Table 10. In comparison to the scatter plot in Figure 6, it is seemed that the predicted item difficulty parameter estimates from the LLTM with the incidentals fit tighter to the diagonal line. The visual inspection was corroborated by a slightly higher correlation between the predicted item difficulties for the LLTM model with the incidentals and the Rasch model (.89) than the correlation between the Rasch and the LLTM without the incidentals (.84); the correlation between the two LLTM models was .95.

Chapter 5

DISCUSSION AND LIMITATIONS

A Cognitive Model for Nonword Repetition

The current study evaluated the phonological loop component of Baddeley's (2000) model as cognitive processing model for NWR. The model hypothesizes that both PSTM and long-term memory contribute to variation in nonword repetition ability. Previous studies of nonword repetition tasks have primarily focused on the diagnostic accuracy of the task when used to identify children with language impairments (Dollaghan & Campbell, 1998; Gutierrez-Clellen & Simon-Cerejido, 2010); however, the cognitive processing analysis of nonword repetition items in the current study provides important information directed at understanding the substantive meaning of the construct underlying nonword repetition tasks. The results provide construct validity evidence in support of a cognitive processing model for NWR and a list of construct-relevant item characteristics for the future development of nonword repetition items.

The developed item difficulty model suggests that PSTM capacity is primarily responsible for the variation in NWR accuracy. Specifically, the length of the nonwords had the largest impact on item difficulty as evidenced by the largest R^2 value (.62) when the item radical number of phonemes was added as the lone predictor in the LLTM model. Further, the predictor number of phonemes had the largest partial squared correlation (.74) in the final LLTM model, which indicates that when controlling for all other radicals and incidentals in the model, it explained the most variance in item difficulty.

Also consistent with the predictions of Baddeley's model and the results of previous studies (Graf-Estes, et al, 2007), the relationship between the predictor number of phonemes and item difficulty is positive. The positive direction of the regression and structural LLTM weight suggests that repeating longer nonwords requires more PSTM resources in terms of the capacity of the phonological short-term store and the maintenance of the phonological memory trace by the sub-vocal rehearsal process. Thus, increases in the number of phonemes were associated with increases in item difficulty.

In terms of the contributions made by long-term memory, the item difficulty model supports the predictions of Baddeley's model that information stored in long-term memory contributes positively to NWR accuracy. In addition, in comparison to PSTM, information stored in long-term memory plays a lesser role in NWR. Two radicals operationalized the potential lexical (ND) and sub-lexical (PP) support provided by long-term memory during NWR. As individual predictors, both PP and ND explained small to moderate amounts of variance in item difficulty ($R^2 = .1$ and $.26$ respectively); however, when modeled with other predictors (e.g., number of phonemes), only PP was a significant predictor. The results of the final LLTM model suggest that PP is significantly negatively associated with item difficulty and that it explains a moderate amount of variance (partial $R^2 = .31$). This result supports the prediction that redintegration is a *support* mechanism, as opposed to the primary mechanism of PSTM, that is used during NWR.

Nonword Item Generation

After evaluating the cognitive model, the radical structure was evaluated for its potential to generate nonword items. The results provided support for using the proposed cognitive model as an item blueprint for cognitive complexity when generating nonword repetition items. That is, the estimated LLTM parameter weights corresponded to the hypothesized representativeness of the cognitive processes for the construct NWR ability. For example, Baddeley's model assumes a limited capacity of PSTM, therefore it was hypothesized that longer nonwords would require more resources in terms of capacity of temporarily storing the phonological trace and maintenance by the sub-vocal rehearsal process; therefore, longer nonwords were predicted to be more difficult than shorter nonwords. The LLTM results supported this hypothesis such that the number of phonemes in a nonword was significantly positively related to item difficulty (unstandardized $\eta = .43$).

In addition to the contribution of PSTM, it was predicted by Baddeley's model that sub-lexical information stored in long-term memory would support PSTM recall; results indicated that PP was significantly negatively associated with item difficulty (unstandardized $\eta = -4.36$). Although consonant clusters were not initially included as an item radical, they were included as an indicator of phonological complexity and were found to be significantly positively associated with item difficulty (unstandardized $\eta = .53$). Furthermore, consonant clusters had the second largest partial R^2 (.63), larger than that of PP, which suggests that the phonological complexity of the nonword is important to consider when creating

nonwords. Other measures of phonological complexity, such as dip- and trip-
thongs should be explored.

Neighborhood density, on the other hand, was predicted to negatively impact item difficulty; however, the LLTM results indicated that it was not a significant predictor when combined with the other predictors. In retrospect, the null result of ND is not surprising because of the resources that were used to calculate ND. ND was calculated by first generating a nonword and then checking it against the LEXESP dictionary to see if the nonword matched any of the real words in the dictionary when adding, deleting or substituting each of the phonemes in the nonword. At best, the LEXESP dictionary is an exaggeration of an adult's lexicon and it certainly largely over estimates a child's lexicon. Therefore, many of the phonological neighbors that were considered in the calculation of ND for a particular nonword would not be in a child's lexicon. Thus, even though some of the nonwords in the current study were calculated to have phonological neighbors, it is more likely that the ND for many of these nonwords was zero for a child. Therefore, the children were likely unable to benefit from the lexical support provided by nonwords with high ND. Future studies should take this into consideration by either creating or finding a word frequency dictionary that is calculated using a corpus of children's language samples.

In addition to the item radicals, the impact of potential item incidentals on item difficulty was evaluated. Results indicated that the item incidentals were significantly related to item difficulty above and beyond that of the item radicals. In particular the incidental *begins with a vowel* was highly statistically significant and negatively associated with item difficulty; this suggests that items that began

with a vowel were easier than items that began with a consonant. Further examination revealed that the variable *begins with a vowel* was significantly negatively correlated with the radical PP and the radical number of phonemes. These correlations suggest that words that started with a vowel tended to be shorter and have lower PP. While the other incidental, ends with a vowel, was a significant predictor in the LLTM model, it was not significantly correlated with any of the other predictors. Incidentals have the potential to introduce construct irrelevant variance as evidenced by the current results; therefore, it is just as important to explore item incidentals as it is radicals. The item incidental *late acquired sounds* was unable to be explored in current study because all of the items included at least one of these sounds. Future studies may want evaluate this incidental as children with a phonological or articulation impairment may struggle more with producing these sounds than typically developing children or even children with language impairment.

Item Decomposition. One benefit of item difficulty modeling is the ability to decompose the items into its representative components of processing. That is, the value of the item characteristics can be multiplied by the structural weights of the item radicals and incidentals. For example, in equation 4 for item 21 with seven (7) phonemes, 0.08 PP, zero (0) consonant clusters, a (1) beginning vowel, and an (1) ending vowel, item difficulty is decomposed as follows, using the weights given in Table 9:

$$\begin{aligned}
 b'_i &= .39(7) + -4.36(0.08) + .53(0) + -.63(1) + -.25(1) & (4) \\
 &= 2.72 + -.36 + 0 + -.63 + -.25 \\
 &= 1.48
 \end{aligned}$$

Thus, item 21 is predicted to have moderate difficulty (1.48), and the primary source of difficulty is PSTM load. In fact, it is difficult to conceive of an example where PSTM load would not be the primary source of item difficulty given the current set of items; however, item decomposition can help to distinguish between two nonwords with the same number of phonemes but different values on the other predictors. The contributions of long-term memory would be more apparent if an additional set of nonwords based on the phonological rules of English were administered to these children. An additional dichotomous item radical that was coded to indicate the base language of the nonwords (Spanish or English) could then be included in the LLTM model. If coded one for Spanish and zero for English, the structural weight would represent the impact of native language knowledge stored in long-term memory on item difficulty. Language by item feature interactions would also warrant further investigation.

In addition to being able to identify sources of item difficulty, the structural weights estimated by the LLTM could be used for programming an automatic item generator. As alluded to earlier, the viability of automatic item generation is dependent upon a number of factors including the identification of a set of radicals and incidentals, ease of programming the algorithms into software, among others. In the current study, a nonword generator was developed to create nonwords that adhered to the constraints of Spanish phonology. With some additional programming, the same generator could utilize the identified radicals and incidentals and their estimated structural weights to create a new set of nonwords. These new nonwords could then be administered to another group of children for cross validation. Further, the accuracy of item difficulty prediction could be achieved through the exploration of more item radicals and the tighter

control of incidentals. Future studies could cross validate the item blueprint and structural model by using the estimated structural weights from the LLTM model in conjunction with nonword generator to create a new set of items. Such studies would be warranted before moving forward with automatic item generation.

Limitations

Several limitations of the study design and data structure affected the interpretability and generalizability of the results and conclusions drawn here. First, the predictions of the LLTM were limited by the data model fit to the Rasch model. In the current study, the data seemed to fit the Rasch model moderately well, but there were some items that had less than desirable item fit statistics. That being said, the purpose of the study was not to create Rasch fitting items, but to evaluate a set of item radicals. When including the incidentals, the results of the current study were able to explain nearly 80% of the variance in item difficulty which is quite substantial.

If future studies were concerned with the fit of the data to the Rasch model and wanted to try to improve the overall fit, a next step would be to estimate the fit of the data to a 2 parameter-logistic (2-PL) IRT model. A 2-PL model allows both the item difficulties and discrimination parameters to vary across items. Using BILOG-MG, the data were fit to a Rasch and 2-PL model for comparison. Results indicated that the data fit the 2-PL model better as indicated by a lower AIC statistic and a significant nested model chi-square, $\chi^2(34) = 936.60, p < .01$. Correlations between the item features and the item discrimination parameters would reflect the relationship and potential impact that the item features have on item discrimination. The correlations between the item

features and the discrimination parameters ranged from -0.21 to 0.24, but none were significant. If some of the correlations had been stronger and significant, then a constrained 2-PL, such as in Daniel and Embretson (2010), could be used to predict the impact of the item features on item difficulty and discrimination parameters. Currently, the eRm package for R does not handle constrained 2-PL models, so future studies may want to consider using other software packages to explore the constrained 2-PL model with NWR data.

In terms of limitations to the interpretability of the results, the multicollinearity of the model predictors was most problematic. Though the collinearity diagnostics for the regression model were within acceptable limits, most of the item radicals were inter-correlated at .40 or greater which can produce unreliable results. As a result, the individual contribution of any one predictor in a multiple-predictor model is difficult to interpret. For example, the sign of the estimate for ND changed from positive to negative in the final regression model, suggesting a possible suppressor effect (Cohen, Cohen, West, & Aiken, 2003). Upon further examination, ND is moderately negatively correlated with consonant clusters (-.41) and moderately negatively correlated with the number of phonemes (-.60). These relationships seem to clash as ND is moderately negatively correlated with item difficulty whereas the number of phonemes and consonant clusters are moderately positively correlated with item difficulty. These correlations make practical sense such that as nonwords get longer they have fewer phonological neighbors; a similar but more complex argument could be made for consonant clusters. Though not specifically tested here, it is likely that the effect of ND on item difficulty was suppressed by the variables number of phonemes and consonant clusters. It is unclear how multi-

collinearity is affecting the outcomes of the LLTM models, but some possible options for handling multi-collinearity would be to create composite variables or principle components of the predictors. In a principal components analysis, the correlated predictors undergo a linear transformation and are reduced to a smaller set of uncorrelated variables or components. These components can then replace the original predictors in the regression or LLTM model while still representing an associated cognitive process.

The generalizability of the results was limited by the chosen radicals and the specific sample of items generated for the current study. The experimental items were systematically constructed by manipulating the original set of item radicals. In doing so, however, there were limitations to what could be created. For example, it was very difficult to create nonwords with phonological neighbors at the upper extremes of nonword length. This meant that some of the longer nonwords that were classified as having high ND only had one phonological neighbor, while the shorter nonwords that were classified as having high ND had many neighbors, in some cases upwards of ten with the highest being 13; however, such a limitation was difficult to avoid. At least all of the statistical analyses all used the actual values of the item radicals and not their item descriptor categories, which largely mitigated the impact of this limitation.

Finally, the generalizability of the results was also limited by the lack of a cross validation. Future studies should aim to cross validate the proposed model with other sets of items from existing measures, specifically those that have shown good diagnostic power for identifying children with LI. Perhaps NWR tasks can vary in terms of all of the radicals in the current study, but it is also possible

that the best discriminators of LI are a subset of items that vary in terms of only specific radicals, or even ones that have yet to be thought of. Additionally, an alternative analysis to the one in current study could model the effect size differences between LI and TD groups on items in lieu of item difficulty. Such an analysis has the potential to identify which design features are related to diagnostic discriminatory power.

Conclusion

Nonword repetition tasks have been used extensively as a diagnostic tool to identify children with language impairment; however, there is debate as to whether valid inferences about a child's language impairment status can be drawn from the results of a NWR task. Current theories on language impairment suggest two primary sources for language impairment, (1) a language deficit mostly in the area of grammar and syntax (Leonard, 1998) and (2) a processing deficit which has been investigated as a general processing deficit (Kail, 1994; Kail & Leonard, 1986) and as a processing deficit in specific areas such as PSTM (Gathercole & Baddley, 1989, 1990). The results of the current study suggest that nonword repetition primarily taps into PSTM, with some variance in item difficulty being attributed to support from long-term memory. Thus, we can then infer that children who "fail" NWR tasks are demonstrating a PSTM deficit. Surprisingly, however, despite observing significant group mean differences, researchers have struggled to observe adequate levels of diagnostic accuracy when using NWR tasks to identify bilingual children with language impairments (Gutierrez-Clellen & Simon-Cerejido, 2010; Morgan, 2010, but see Girbau & Schwartz, 2008). Could this mean that bilingual children with LI do not

demonstrate PSTM deficits on the same magnitude as their monolingual peers? Perhaps being bilingual reduces the likelihood of a PSTM deficit because bilingual children are able to develop their PSTM system to greater degree than monolingual children by the virtue of having to learn two or more languages. While the goal of this study was not to evaluate the diagnostic capabilities of these nonword items, the children with TD did significantly outperform the children with LI. That said, the results of this study do provide a list of item radicals and incidentals that will help future efforts to generate different types of nonwords and investigate their impact on diagnostic accuracy.

Human Subjects

This research was conducted with the expressed permission of Arizona State University. Appendix E contains the IRB approval documents for this study.

REFERENCES

- Archibald, L. M. D., & Gathercole, S. E. (2006). Short-term and working memory in specific language impairment. *International Journal of Language & Communication Disorders, 41*(6), 675-693.
- Arendasy, M. (2005). Automatic generation of Rasch-calibrated items: Figural Matrices Test GEOM and Endless Loops Test EC. *International Journal of Testing, 5*, 197-224.
- Arendasy, M., & Sommer, M. (2007). Using psychometric technology in educational assessment: The case of a schema-based isomorphic approach to the automatic generation of quantitative reasoning items, *Learning and Individual Differences, Vol. 17*, no. 4, pp. 366-383.
- Baddeley, A. (2003). Working memory and language: an overview. *Journal of Communication Disorders, 36*(3), 189-208.
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417-423.
- Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158-173.
- Baddeley, A., & Hitch, (1974). Working memory. In *The Psychology of Learning and Motivation* (Bower, G.A., ed.), pp. 47-89, Academic Press
- Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575-589.
- Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Fredriksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-359). Hillsdale, NJ: Lawrence Erlbaum.
- Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Mahwah, NJ: Lawrence Erlbaum.
- Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15*, 129-137.
- Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

- Botting, N., & Conti-Ramsden, G. (2001). Non-word repetition and language development in children with specific language impairment (LI). *International Journal of Language & Communication Disorders, 36*(4), 421-432.
- Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research*. The Guildford Press. New York, NY.
- Brown, G. D. A., & Hulme, C. (1995). Modeling item length effects in memory span: No rehearsal needed? *Journal of Memory and Language, 34*(5), 594-621.
- Brown, G. D. A., & Hulme, C. (1996). Nonword repetition, STM, and word age-of-acquisition: A computational model. *Models of Short-Term Memory, 129-148*.
- Coady, J. A., & Evans, J. L. (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (LI). *International Journal of Language & Communication Disorders, 43*(1), 1-40.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time and the effects of word-length on immediate memory. *Journal of Memory and Language, 31*, 1–17.
- Dawson, J. I., Stout, C. E., & Eyer, J. A. (2003). *The Structured Photographic Expressive Language Test* Third Edition. Dekalb, IL: Janelle Publications.
- Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods, 37*(4), 665-671.
- De Cara, B., & Goswami, U. (2002). Similarity relations among spoken words: The special status of rimes in English. *Behavior Research Methods Instruments and Computers, 34*(3), 416-423.
- De Cara, B., & Goswami, U. (2003). Phonological neighbourhood density: Effects in a rhyme awareness task in five-year-old children. *Journal of Child Language, 30*(03), 695-710.

- Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41*(5), 1136.
- Edwards, J., & Lahey, M. (1998). Nonword repetitions of children with specific language impairment: Exploration of some explanations for their inaccuracies. *Applied Psycholinguistics, 19*(2), 279-309.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.
- Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum Press.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.
- Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*, 407-433.
- Embretson, S. E. (2000). Multidimensional measurement from dynamic tests: Abstract reasoning under stress. *Multivariate Behavioral Research, 35*, 505-543.
- Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychological Science Quarterly, 50*(3), 328-344.
- Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension. *Applied Psychological Measurement, 11*, 175-193.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.
- Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language, 42*(4), 481-496.

- Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study *Journal of Memory and Language*, 28(2), 200-213.
- Gathercole, S. E., & Baddeley, A. D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language*, 29(3), 336-360.
- Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hillsdale, NJ: Erlbaum.
- Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology Learning Memory and Cognition*, 25, 84-95.
- Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology A*, 54(1), 1-30.
- Gathercole, S.E., Service, E., Hitch, G.J., Adams, A.M., & Martin, A.J. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology*, 13, 65–77.
- Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology*, 28(5), 887-898.
- Girbau, D., & Schwartz, R. (2007). Non-word repetition in Spanish-speaking children with specific language impairment (LI). *International Journal of Language & Communication Disorders*, 42(1), 59-75.
- Girbau, D., & Schwartz, R. G. (2008). Phonological working memory in Spanish–English bilingual children with and without specific language impairment. *Journal of Communication Disorders*, 41, 124-145.
- Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42, 351-373.
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25, 21–35.
- Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30, 394-411.
- Graf-Estes, G. K., Evans, J. L., & Else-Quest, M. (2007). Differences in nonword repetition performance of children with and without specific language

- impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, 50, 177–195.
- Gray, S. (2003). Word-learning by preschoolers with specific language impairment: What predicts success? *Journal of Speech, Language, and Hearing Research*, 46(1), 56-67.
- Gray, S. (2004). Word learning by preschoolers with specific language impairment predictors and poor learners. *Journal of Speech, Language and Hearing Research*, 47(5), 1117-1132.
- Holling, H., Bertling, J., & Zeuch, N. (2008). Automatic item generation of probability word problems. *Studies in Educational Evaluation*, 35, 71–76.
- Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, 30, 685–701.
- Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a reintegration process in immediate serial recall. *Journal of Experimental Psychology-Learning Memory and Cognition*, 23(5), 1217-1232.
- Irvine, S. H. (2002). The foundations for item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3–24). Mahwah, NJ: Lawrence Erlbaum.
- Kail, R. (1994). A method for studying the generalized slowing hypothesis in children with specific language impairment. *Journal of Speech, Language and Hearing Research*, 37(2), 418.
- Kail, R., & Leonard, L. B. (1986). Word-finding abilities in language-impaired children. *ASHA Monographs*, (25)(25), 1-39.
- Kaufman, A., & Kaufman, N. (2004). Kaufman assessment battery for children 2nd edition.
- Kohnert, K., Windsor, J., & Yim, D. (2006). Do language-based processing tasks separate children with language impairment from typical bilinguals? *Learning Disabilities Research and Practice*, 21(1), 19.
- Leonard, L. B. (1998). *Children with specific language impairment*. Cambridge, MA: The MIT Press.
- Lipinski, J., & Gupta, P. (2005). Does neighborhood density influence repetition latency for nonwords? separating the effects of density and duration. *Journal of Memory and Language*, 52(2), 171-192.

- Mair, P., Hatzinger, R., & Maier, M. (2010). eRm: Extended Rasch Modeling. R package version 0.13-0. <http://CRAN.R-project.org/package=eRm>
- McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: John Wiley & Sons.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 117-195.
- Montgomery, J. W. (1995). Sentence comprehension in children with specific language impairment: The role of phonological working memory. *Journal of Speech, Language and Hearing Research*, *38*(1), 187.
- Montgomery, J. W. (2004). Sentence comprehension in children with specific language impairment: Effects of input rate and phonological working memory. *International Journal of Language & Communication Disorders*, *39*(1), 115-133.
- Munson, B., Kurtz, B. A., & Windsor, J. (2005). The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *48*(5), 1033.
- Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal*, *40*, 905-928.
- Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. 1938 individual form. London: Lewis.
- Restrepo, M. A. (1998). Identifiers of predominantly spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(6), 1398.
- Roberts, J. E., Burchinal, M., & Footo, M. M. (1990). Phonological process decline from 2;6 to 8 years. *Journal of Communication Disorders*, *23*, 205–217.
- Roodenrys, S. & Hinton, M. (2001). Sublexical or lexical effects on serial recall of nonwords? *Journal of Experimental Psychology Learning Memory and Cognition*, *28*(1), 29-33.
- Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-

- term memory. *Journal of Experimental Psychology Learning Memory and Cognition*, 28(6), 1019-1034.
- Sebastián-Gallés, N., Martí, M. A., Carreiras, M., & Cuetos, F. (2000). LEXESP: Una base de datos informatizada del español. *Primer Informe [LEXESP: A Computerized Database of Spanish]. Spain: Universitat De Barcelona.*
- Semel, E., Wiig, E. H., & Secord, W. H. (2009). *Clinical evaluation of language fundamentals preschool—Spanish edition (CELF-P2)*. San Antonio, TX: Psychological Corporation.
- Schweickert, R. (1993). A multinomial processing tree model for degradation and reintegration in immediate recall. *Memory and Cognition*, 21, 168-168.
- Sheehan, K. M., & Ginther, A. (2000, April). What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section. Paper presented at the 2000 Annual Meeting of the National Council of Measurement in Education, New Orleans, LA. In Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42, 351-373.
- Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language and Hearing Research*, 47(6), 1454-1468.
- Thomson, J. M., Richardson, U., & Goswami, U. (2005). Phonological similarity neighborhoods and children's short-term memory: Typical development and dyslexia. *Memory and Cognition*, 33(7), 1210.
- Thorn, A. S. C., & Frankish, C. R. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology Learning Memory and Cognition*, 31(4), 729-735.
- Thorn, A. S. C., & Gathercole, S. E. (1999). Language-specific knowledge and short-term memory in bilingual and non-bilingual children. *The Quarterly Journal of Experimental Psychology A*, 52(2), 303-324.
- Thorn, A. S. C., & Gathercole, S. E. (2001). Language differences in verbal short-term memory do not exclusively originate in the process of subvocal rehearsal. *Psychonomic Bulletin and Review*, 8(2), 357-364.
- Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistics and Phonetics*, 17, 487-499.

- Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3), 374-408.
- Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68(1-2), 306-311.
- Whitley, M. S. (2002). *Spanish/English Contrasts: A Course in Spanish Linguistics*. Georgetown University Press.
- Yu, C., & Muthén. B. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Unpublished Dissertation. University of California at Los Angeles.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BLOG-MG: Multiplegroup IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International.

Table 1. Nonword Item Descriptive Statistics

| Item Number | Proportion Correct | Standard Deviation | <i>r</i> biserial |
|-------------|--------------------|--------------------|-------------------|
| x1 | 0.81 | 0.39 | 0.25 |
| x2 | 0.74 | 0.44 | 0.29 |
| x3 | 0.51 | 0.50 | 0.42 |
| x4 | 0.80 | 0.40 | 0.26 |
| x5 | 0.78 | 0.42 | 0.05 |
| x6 | 0.85 | 0.36 | 0.08 |
| x7 | 0.56 | 0.50 | 0.26 |
| x8 | 0.54 | 0.50 | 0.29 |
| x9 | 0.63 | 0.48 | 0.01 |
| x10 | 0.92 | 0.27 | -0.19 |
| x11 | 0.30 | 0.46 | 0.15 |
| x12 | 0.35 | 0.48 | 0.13 |
| x13 | 0.61 | 0.49 | 0.37 |
| x14 | 0.82 | 0.38 | 0.34 |
| x15 | 0.28 | 0.45 | 0.37 |
| x16 | 0.34 | 0.47 | 0.38 |
| x17 | 0.74 | 0.44 | 0.19 |
| x18 | 0.68 | 0.47 | 0.17 |
| x19 | 0.40 | 0.49 | 0.37 |
| x20 | 0.26 | 0.44 | 0.23 |
| x21 | 0.75 | 0.44 | 0.01 |
| x22 | 0.58 | 0.49 | 0.13 |
| x23 | 0.31 | 0.46 | 0.22 |
| x24 | 0.52 | 0.50 | 0.15 |
| x25 | 0.63 | 0.49 | 0.36 |
| x26 | 0.63 | 0.49 | 0.34 |
| x27 | 0.58 | 0.50 | 0.37 |
| x28 | 0.63 | 0.49 | 0.42 |
| x29 | 0.66 | 0.48 | 0.20 |
| x30 | 0.51 | 0.50 | 0.19 |
| x31 | 0.29 | 0.45 | 0.24 |
| x32 | 0.25 | 0.43 | 0.37 |
| x33 | 0.35 | 0.48 | 0.07 |
| x34 | 0.13 | 0.33 | 0.19 |
| x35 | 0.26 | 0.44 | 0.18 |
| x36 | 0.39 | 0.49 | 0.16 |

Table 2. Inter-Item Correlations

| item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|------|-------|-------|-------|-------|-------|-------|-------|-------|------|------|------|------|-------|-------|-------|-------|-------|-------|-------|
| 1 | 1.00 | | | | | | | | | | | | | | | | | | |
| 2 | .13 | 1.00 | | | | | | | | | | | | | | | | | |
| 3 | .08 | .24** | 1.00 | | | | | | | | | | | | | | | | |
| 4 | .30** | .14 | .24** | 1.00 | | | | | | | | | | | | | | | |
| 5 | .28* | .28* | .11 | .08 | 1.00 | | | | | | | | | | | | | | |
| 6 | .21** | .17 | -.03 | .08 | .13 | 1.00 | | | | | | | | | | | | | |
| 7 | .29* | -.04 | .20 | .28* | .16 | .15 | 1.00 | | | | | | | | | | | | |
| 8 | .23 | .15 | .55** | .54** | .09 | .22* | .23** | 1.00 | | | | | | | | | | | |
| 9 | .10 | .05 | -.01 | -.07 | .04 | .05 | .23 | .29* | 1.00 | | | | | | | | | | |
| 10 | .03 | -.15 | -.28* | .12 | .22 | .31* | .08 | .08 | .15 | 1.00 | | | | | | | | | |
| 11 | .15 | -.01 | .13 | -.08 | .05 | .11 | .15 | .42** | .16 | .02 | 1.00 | | | | | | | | |
| 12 | .07 | .10 | .09 | .12 | .09 | .00 | .06 | .00 | .19* | -.02 | .23* | 1.00 | | | | | | | |
| 13 | .12 | .27** | .16 | .17 | .05 | .10 | .09 | .22 | .16 | .03 | -.01 | .17 | 1.00 | | | | | | |
| 14 | .25** | .21* | .28** | .16 | .34** | .10 | .28* | .30* | .15 | -.11 | .16 | .17 | .15* | 1.00 | | | | | |
| 15 | .10 | .07 | .29** | .13 | .03 | .01 | .23 | .35** | .11 | .09 | .27* | .17 | .05 | .11 | 1.00 | | | | |
| 16 | .10 | .21* | .20* | .13 | .34** | .11 | .20 | .33** | .07 | .10 | .24 | .06 | .13* | .22** | .25** | 1.00 | | | |
| 17 | .12 | .20 | .39** | .25* | .10 | .14* | .20* | .29** | -.03 | .12 | -.15 | .16 | .24** | .32** | .19** | .24** | 1.00 | | |
| 18 | .28** | .16 | .24 | .17 | -.02 | .16** | .11 | .22* | -.02 | .20 | .02 | -.04 | .01 | .22** | .15* | .13 | .27** | 1.00 | |
| 19 | .13 | .13 | .25* | .15 | .21* | .17 | .24** | .18* | .20 | .10 | .03 | .20 | .26* | .15 | .26* | .23 | .21* | .26** | 1.00 |
| 20 | .03 | .09 | .09 | -.05 | .16 | .15* | .21* | .17 | .23 | .00 | .12 | .05 | .09 | .15* | .16* | .24** | .21** | .07 | .20* |
| 21 | .12 | .10 | -.03 | .02 | .10 | .27** | -.06 | .30* | .22* | .06 | .20* | .08 | .14 | .13 | .14* | -.01 | .18* | .03 | .11 |
| 22 | -.06 | .22 | .23 | -.16 | .17 | .04 | .27* | .27* | .19* | .08 | .15 | .21* | .11 | .04 | .29* | .14 | .17 | .02 | .50** |

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

Note - .00 correlations go beyond 2 decimal places

| item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|------|------------------|------------------|------------------|------------------|------|------------------|------------------|------------------|------------------|------------------|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 23 | .11 | .18 | .35** | .18 | .12 | .21** | .37** | .23 | .14 | .07 | .06 | .26** | .13 | .19** | .18 ^ˆ | .15 ^ˆ | .31** | .18 ^ˆ | .31 ^ˆ |
| 24 | .11 | -.03 | .17 | .12 | .18 | .21 | .47** | .19 | .23 ^ˆ | .16 | .27** | .08 | .11 | .17 | .24 | .19 | .24 | -.04 | .41** |
| 25 | .06 | .22 ^ˆ | .20 ^ˆ | .19 ^ˆ | .07 | .07 | .39** | .38** | .12 | .17 | -.27 ^ˆ | -.17 | .23** | .18** | .19** | .08 | .19** | .22** | .30 ^ˆ |
| 26 | .13 | .08 | .11 | .02 | -.04 | .16 | .00 | .26 ^ˆ | .38** | .00 | .12 | .08 | .20 ^ˆ | .19 ^ˆ | .25** | .17 | .25 ^ˆ | .35** | .16 |
| 27 | .11 | .31** | .25** | .02 | .19 | .05 | .10 | .20 | .12 | -.09 | .19 | .14 | .13 ^ˆ | .26** | .07 | .18** | .18 ^ˆ | .16 ^ˆ | .41** |
| 28 | .04 | .11 | .15 | .04 | .14 | .06 | .28 ^ˆ | .27 ^ˆ | .07 | .19 | .16 | .11 | .18** | .21** | .25** | .15 ^ˆ | .28** | .29** | .27 ^ˆ |
| 29 | .15 ^ˆ | .23 | .24 | .01 | .09 | .15 ^ˆ | .18 ^ˆ | .05 | .07 | .05 | -.02 | .21 | .15 ^ˆ | .27** | .26** | .17 ^ˆ | .27** | .19** | .29** |
| 30 | .04 | .09 | .27 ^ˆ | .12 | .14 | .24** | .03 | .20 ^ˆ | .23 | .25 ^ˆ | .21 | -.04 | .15 ^ˆ | .16 ^ˆ | .16 ^ˆ | .17 ^ˆ | .11 | .23** | .17 ^ˆ |
| 31 | .09 | .09 | .21 | .03 | .11 | .06 | .24** | .10 | -.06 | -.08 | -.07 | .07 | .15 ^ˆ | .06 | .22** | .13 | .13 ^ˆ | .09 | .26** |
| 32 | .08 | .23 | .21 | .11 | .09 | .16 | .24** | .23** | .19 | .16 | .24 | .15 | .16 | .27 ^ˆ | .14 | .22 | .15 | .19 ^ˆ | .41** |
| 33 | .00 | .23 | .01 | -.10 | .08 | .15 | .04 | -.03 | -.05 | .05 | .11 | .00 | .07 | .19 | -.04 | .01 | .21 | .21 | .16 |
| 34 | .17 | .18 | .23 | .08 | .16 | .19 | .10 | .01 | .09 | -.06 | .04 | .26** | .28 ^ˆ | .13 | .30 ^ˆ | .06 | .09 | .16 | .16 |
| 35 | .06 | .19 | .17 | -.03 | .13 | .19 | .19 | .12 | .15 | .05 | .09 | .22 ^ˆ | .14 | .09 | .21 | .27 ^ˆ | .16 | .00 | .20 |
| 36 | .10 | .17 | .36** | .19 | -.01 | .22** | .36** | .23 | .05 | -.13 | .12 | .25** | .11 | .19** | .23** | .01 | .27** | .30** | .30 ^ˆ |

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

Note - .00 correlations go beyond 2 decimal places

| Item | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
|------|-------|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|-------|------|-------|------|------|
| 20 | 1.00 | | | | | | | | | | | | | | | | |
| 21 | .24** | 1.00 | | | | | | | | | | | | | | | |
| 22 | .15 | .03 | 1.00 | | | | | | | | | | | | | | |
| 23 | .16* | .09 | .27** | 1.00 | | | | | | | | | | | | | |
| 24 | -.02 | .01 | .20* | .30** | 1.00 | | | | | | | | | | | | |
| 25 | .10 | .17* | -.12 | .24** | .22 | 1.00 | | | | | | | | | | | |
| 26 | .25* | -.15 | .14 | .20 | .04 | .15 | 1.00 | | | | | | | | | | |
| 27 | .16* | .19** | .21 | .05 | .14 | .12 | .14 | 1.00 | | | | | | | | | |
| 28 | .20** | .23** | .33* | .18* | .18 | .27** | .22* | .19** | 1.00 | | | | | | | | |
| 29 | .21** | .36** | .29* | .17* | .11 | .29** | .14 | .27** | .29** | 1.00 | | | | | | | |
| 30 | .14* | .20** | .01 | .07 | .25* | .23** | .00 | .07 | .23** | .19** | 1.00 | | | | | | |
| 31 | .14* | .17* | .01 | .17* | .14 | .12 | .31* | .16* | .10 | .29** | .14* | 1.00 | | | | | |
| 32 | .16 | .21 | .35** | .28* | .47** | .28* | .18 | .39** | .35** | .31** | .26** | .10 | 1.00 | | | | |
| 33 | .15 | -.07 | .08 | .17 | .01 | -.04 | .21 | .11 | .31* | .12 | -.08 | .17 | .02 | 1.00 | | | |
| 34 | .14 | .09 | .05 | .21* | .03 | .16 | .19 | .16 | .26 | .18 | .00 | .35** | .10 | .10 | 1.00 | | |
| 35 | -.03 | .10 | .17 | .27** | .11 | -.05 | .10 | .12 | .27* | .10 | -.10 | .03 | .34** | .10 | .26** | 1.00 | |
| 36 | .26** | .15* | .14 | .25** | .16 | .23** | .12 | .05 | .29** | .24** | .15* | .20** | .41** | .14 | .08 | .04 | 1.00 |

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

Note - .00 correlations go beyond 2 decimal places

Table 3. Rasch Model Item Difficulty Estimates and Fit Statistics.

| Item | <i>b</i> | S. E. | Lower CI | Upper CI | Infit |
|------|----------|-------|----------|----------|-------|
| 1 | -0.42 | 0.25 | -0.90 | 0.06 | 1.07 |
| 2 | -0.30 | 0.25 | -0.79 | 0.18 | 1.02 |
| 3 | 0.09 | 0.27 | -0.44 | 0.62 | 0.89 |
| 4 | -0.40 | 0.25 | -0.88 | 0.09 | 1.05 |
| 5 | -0.44 | 0.23 | -0.90 | 0.01 | 1.08 |
| 6 | -0.58 | 0.24 | -1.05 | -0.11 | 0.88 |
| 7 | -0.02 | 0.25 | -0.52 | 0.47 | 0.96 |
| 8 | -0.04 | 0.25 | -0.54 | 0.46 | 0.88 |
| 9 | -0.09 | 0.23 | -0.55 | 0.37 | 1.08 |
| 10 | -0.55 | 0.22 | -0.99 | -0.11 | 0.95 |
| 11 | 0.54 | 0.30 | -0.04 | 1.12 | 0.96 |
| 12 | 0.45 | 0.28 | -0.10 | 1.01 | 1.01 |
| 13 | -0.09 | 0.26 | -0.59 | 0.42 | 0.95 |
| 14 | -0.44 | 0.25 | -0.94 | 0.05 | 0.82 |
| 15 | 0.60 | 0.32 | -0.03 | 1.22 | 0.90 |
| 16 | 0.45 | 0.30 | -0.14 | 1.04 | 0.93 |
| 17 | -0.37 | 0.24 | -0.85 | 0.10 | 0.88 |
| 18 | -0.28 | 0.24 | -0.75 | 0.19 | 1.00 |
| 19 | 0.27 | 0.28 | -0.28 | 0.82 | 0.84 |
| 20 | 0.60 | 0.31 | -0.01 | 1.21 | 0.99 |
| 21 | -0.27 | 0.23 | -0.72 | 0.19 | 0.91 |
| 22 | 0.00 | 0.25 | -0.48 | 0.48 | 0.91 |
| 23 | 0.58 | 0.31 | -0.02 | 1.18 | 0.90 |
| 24 | 0.13 | 0.25 | -0.37 | 0.63 | 0.91 |
| 25 | -0.11 | 0.26 | -0.61 | 0.39 | 0.97 |
| 26 | -0.11 | 0.25 | -0.61 | 0.39 | 1.02 |
| 27 | -0.02 | 0.26 | -0.53 | 0.48 | 0.94 |
| 28 | -0.11 | 0.26 | -0.62 | 0.40 | 0.87 |
| 29 | -0.22 | 0.24 | -0.69 | 0.26 | 0.95 |
| 30 | 0.06 | 0.25 | -0.43 | 0.56 | 1.11 |
| 31 | 0.58 | 0.31 | -0.03 | 1.18 | 1.00 |
| 32 | 0.65 | 0.33 | 0.01 | 1.30 | 0.83 |
| 33 | 0.47 | 0.28 | -0.08 | 1.02 | 1.11 |
| 34 | 1.23 | 0.45 | 0.35 | 2.12 | 0.88 |
| 35 | 0.70 | 0.32 | 0.07 | 1.33 | 0.98 |
| 36 | 0.37 | 0.28 | -0.17 | 0.91 | 0.97 |

CI – 95% Confidence Interval

Table 4. Descriptive Statistics of Item Features.

| Item feature | N | Min | Max | Mean | Std. Deviation |
|-------------------------|----|-------|-------|------|----------------|
| Phonotactic Probability | 36 | -1.44 | 1.54 | 0.20 | 0.65 |
| Neighborhood Density | 36 | 0.00 | 13.00 | 1.94 | 3.33 |
| Number of Syllables | 36 | 3.00 | 5.00 | 4.00 | 0.83 |
| Consonant Clusters | 36 | 0.00 | 1.00 | 0.36 | 0.49 |
| Number of Phonemes | 36 | 5.00 | 11.00 | 8.44 | 1.98 |
| Begins with a Vowel | 36 | 0.00 | 1.00 | 0.28 | 0.46 |
| Ends with a Vowel | 36 | 0.00 | 1.00 | 0.69 | 0.47 |

Table 5. Inter-Feature Correlations and their Correlations with Proportion Correct.

| | PP | ND | Number Syllables | Number Phonemes | Consonant Clusters | Begin Vowel | End Vowel | Proportion Correct |
|-----------------------------|--------|--------|---------------------|--------------------|-----------------------|----------------|--------------|-----------------------|
| Phonotactic Probability(PP) | 1.00 | | | | | | | |
| Neighborhood Density(ND) | 0.03 | 1.00 | | | | | | |
| Number of Syllables | 0.55** | -0.36* | 1.00 | | | | | |
| Number of Phonemes | 0.62** | 0.60** | 0.72** | 1.00 | | | | |
| Consonant Clusters | 0.49** | 0.49** | -0.01 | 0.49** | 1.00 | | | |
| Begins with Vowel | -0.40* | 0.06 | 0.01 | 0.40* | -0.35* | 1.00 | | |
| Ends with Vowel | -0.01 | 0.32* | 0.11 | -0.19 | -0.20 | -0.8 | 1.00 | |
| Proportion Correct | -0.26* | 0.53** | -0.42** | -0.74** | -0.62** | 0.27 | 0.35* | 1.00 |

* $p < .05$, ** $p < .01$

Table 6. Regression Model Statistics.

| Model | R ² | Adj. R ² | p-value | Model Comparison |
|-------|----------------|---------------------|---------|------------------|
| 1 | 0.29 | 0.27 | | |
| 2 | 0.17 | 0.15 | | |
| 3 | 0.07 | 0.04 | | |
| 4 | 0.33 | 0.29 | 0.15 | 4,1 |
| 5 | 0.36 | 0.32 | 0.06 | 5,1 |
| 6 | 0.40 | 0.35 | 0.14 | 6,5 |
| 7 | 0.38 | 0.33 | 0.29 | 7,5 |
| 8 | 0.63 | 0.60 | 0.00 | 8,5 |
| 9 | 0.66 | 0.62 | 0.11 | 9,8 |

1 – Neighborhood Density
2 – Number of Syllables
3 – Phonotactic Probability
4 – Neighborhood Density, Number of Syllables
5 – Neighborhood Density, Phonotactic Probability
6 – Neighborhood Density, Phonotactic Probability, Neighborhood Density by Number of Syllables
7 – Neighborhood Density, Phonotactic Probability, Phonotactic Probability by Number of Syllables
8 – Neighborhood Density, Phonotactic Probability, Number of Phonemes
9 – Neighborhood Density, Phonotactic Probability, Number of Phonemes, Consonant Clusters

Table 7. Regression Model Parameter Estimates.

| Model | Parameters | B | S.E. | Beta | Partial Correlations |
|-------|-------------------------|-------|------|-------|----------------------|
| 3 | Constant | 0.58 | 0.07 | | |
| | Neighborhood Density | 0.03 | 0.01 | 0.54 | 0.56 |
| | Phonotactic Probability | -1.33 | 0.69 | -0.27 | -0.32 |
| 8 | Constant | 0.46 | 0.04 | | |
| | Neighborhood Density | -0.02 | 0.01 | -0.26 | -0.22 |
| | Phonotactic Probability | 2.44 | 0.94 | 0.50 | 0.42 |
| | Number of Phonemes | -0.13 | 0.03 | -1.21 | -0.65 |

Table 8. Rasch and LLTM Model Summary Statistics and Final Model Results.

| Model | -2 <i>ln</i> L | AIC | Parameters | Model Comparison | <i>r</i> | <i>R</i> ² |
|-------|----------------|---------|------------|------------------|----------|-----------------------|
| Rasch | 2501.39 | 2591.24 | 35 | | | |
| LLTM1 | 3067.01 | 3069.03 | 1 | | .32 | .10 |
| LLTM2 | 2954.59 | 2956.61 | 1 | | .51 | .26 |
| LLTM3 | 2747.43 | 2749.45 | 1 | | .78 | .62 |
| LLTM4 | 2714.99 | 2719.05 | 2 | 4*,3 | .81 | .65 |
| LLTM5 | 2743.97 | 2748.02 | 2 | 5,3 | .78 | .62 |
| LLTM6 | 2691.19 | 2697.30 | 3 | 6*, 4 | .83 | .70 |
| LLTM7 | 2642.37 | 2652.65 | 5 | 7*, 6 | .89 | .78 |

*R*² and *r* – correlations between LLTM and Rasch *b* parameters

* - model chi-square difference test was significant, *p*-value < .00001

LLTM Model Predictors

1 – Phonotactic Probability

2 – Neighborhood Density

3 – Number of Phonemes

4 – Number of Phonemes, Phonotactic Probability

5 – Number of Phonemes, Neighborhood Density

6 – Number of Phonemes, Phonotactic Probability, Consonant Clusters

7 – Number of Phonemes, Phonotactic Probability, Consonant Clusters, Begins with a Vowel, Ends with a Vowel

Table 9. Final LLTM and LLTM with Incidentals Model Summary Statistics and Final Model Results.

| Model | Parameters | eta ^a | Std. Error | t statistic | Partial R ² |
|-------|-------------------------|------------------|------------|-------------|------------------------|
| 6 | Number of Phonemes | 0.43 | 0.03 | 16.38* | .67 |
| | Phonotactic Probability | -4.36 | 0.94 | 4.65* | .25 |
| | Consonant Clusters | 0.53 | 0.08 | 6.90* | .56 |
| 7 | Number of Phonemes | 0.39 | 0.03 | 16.38* | .73 |
| | Phonotactic Probability | -4.36 | 0.96 | 4.65* | .31 |
| | Consonant Clusters | 0.53 | 0.08 | 6.90* | .63 |
| | Begins with a Vowel | -0.63 | 0.07 | 9.14* | .42 |
| | Ends with a Vowel | -0.25 | 0.07 | 3.73* | .19 |

a - eta values are unstandardized

* $p < .01$

Table 10. LLTM and Rasch Item Parameters Estimates.

| Item | LLTM <i>b</i> | Rasch <i>b</i> | LLTM* <i>b</i> |
|------|---------------|----------------|----------------|
| 1 | 1.16 | 1.32 | 0.87 |
| 2 | 1.36 | 0.85 | 1.21 |
| 3 | -0.55 | -0.34 | 0.03 |
| 4 | 1.01 | 1.22 | 0.67 |
| 5 | 1.10 | 1.28 | 1.24 |
| 6 | 1.33 | 1.83 | 1.19 |
| 7 | -1.17 | 0.00 | -0.53 |
| 8 | 0.97 | 0.03 | 0.63 |
| 9 | 1.07 | 0.44 | 0.84 |
| 10 | 1.36 | 2.56 | 1.85 |
| 11 | -0.79 | -1.15 | -1.04 |
| 12 | -0.68 | -0.96 | -0.96 |
| 13 | 1.20 | 0.15 | 0.76 |
| 14 | 0.45 | 1.43 | 0.70 |
| 15 | -0.91 | -1.54 | -0.97 |
| 16 | -1.09 | -1.22 | -1.35 |
| 17 | 0.44 | 1.03 | 0.92 |
| 18 | 0.46 | 0.72 | 0.94 |
| 19 | -0.99 | -0.73 | -1.28 |
| 20 | -0.84 | -1.47 | -0.84 |
| 21 | 0.83 | 1.04 | 1.30 |
| 22 | -0.40 | 0.18 | -0.50 |
| 23 | -1.06 | -1.24 | -0.45 |
| 24 | -0.74 | -0.17 | -0.76 |
| 25 | -0.42 | 0.22 | -0.51 |
| 26 | 0.07 | 0.22 | -0.06 |
| 27 | -0.21 | -0.03 | 0.06 |
| 28 | -0.35 | 0.22 | -0.05 |
| 29 | -0.52 | 0.54 | -0.66 |
| 30 | -0.28 | -0.24 | -0.32 |
| 31 | -0.20 | -1.43 | -0.34 |
| 32 | -1.38 | -1.60 | -1.65 |
| 33 | -0.32 | -1.00 | -0.43 |
| 34 | -1.32 | -2.58 | -1.61 |
| 35 | -0.55 | -1.49 | -0.67 |
| 36 | -0.41 | -0.77 | -0.73 |

b – IRT difficulty parameter estimates, rescaled to mean of 0 and standard deviation of 1

LLTM – Final LLTM model

LLTM* - Final LLTM model + incidentals

Figure 1. Baddeley's (2000) Multicomponent Model of Working Memory

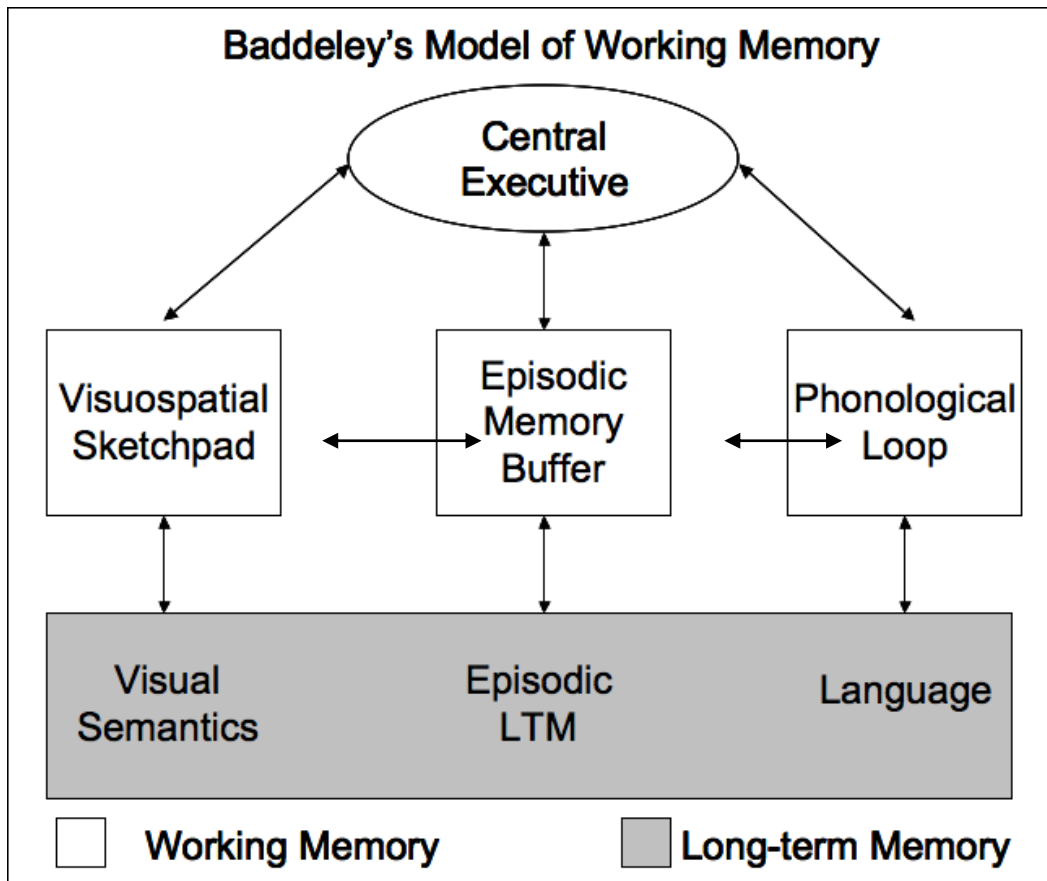


Figure 2. The Phonological Loop (Baddeley *et al.* 1998).

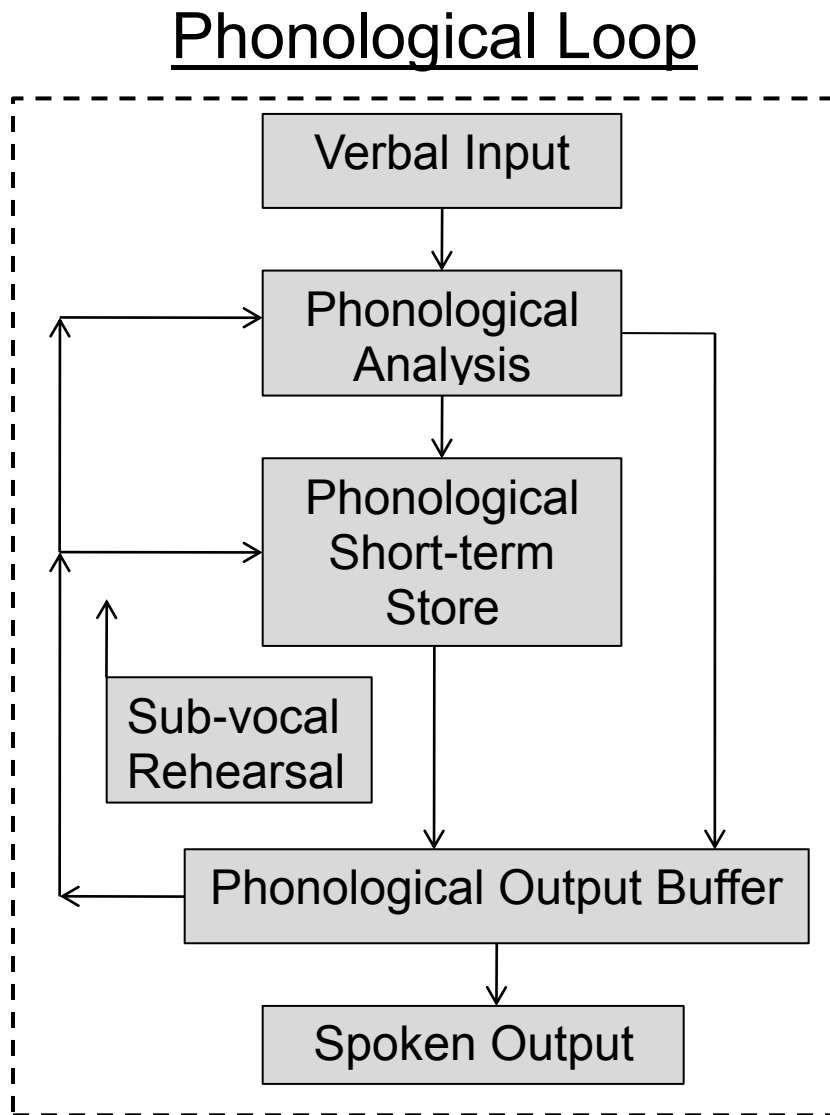
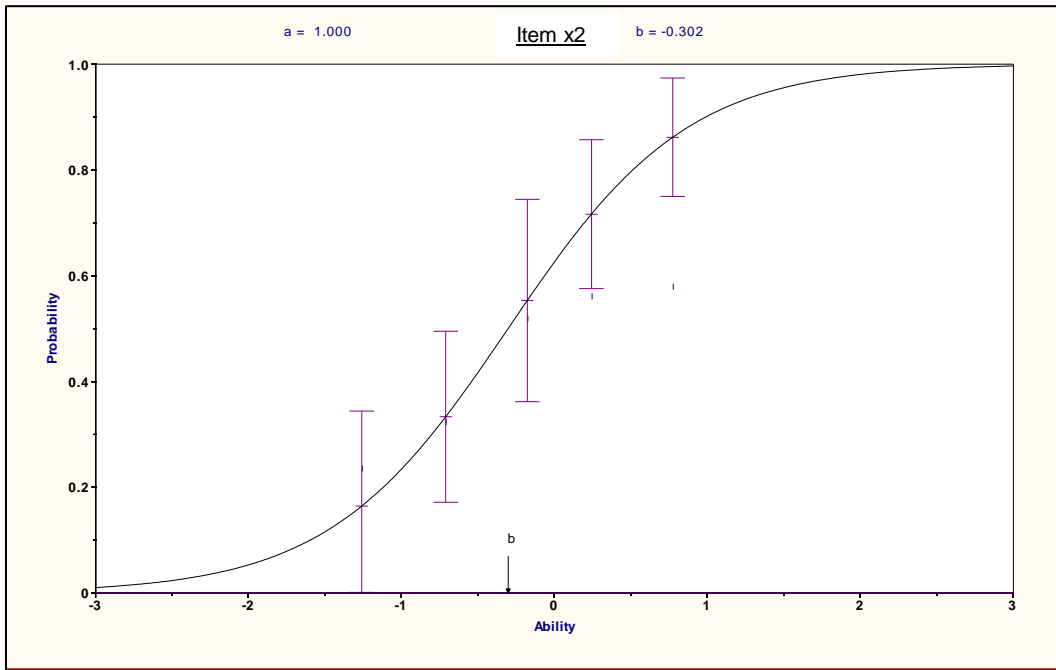
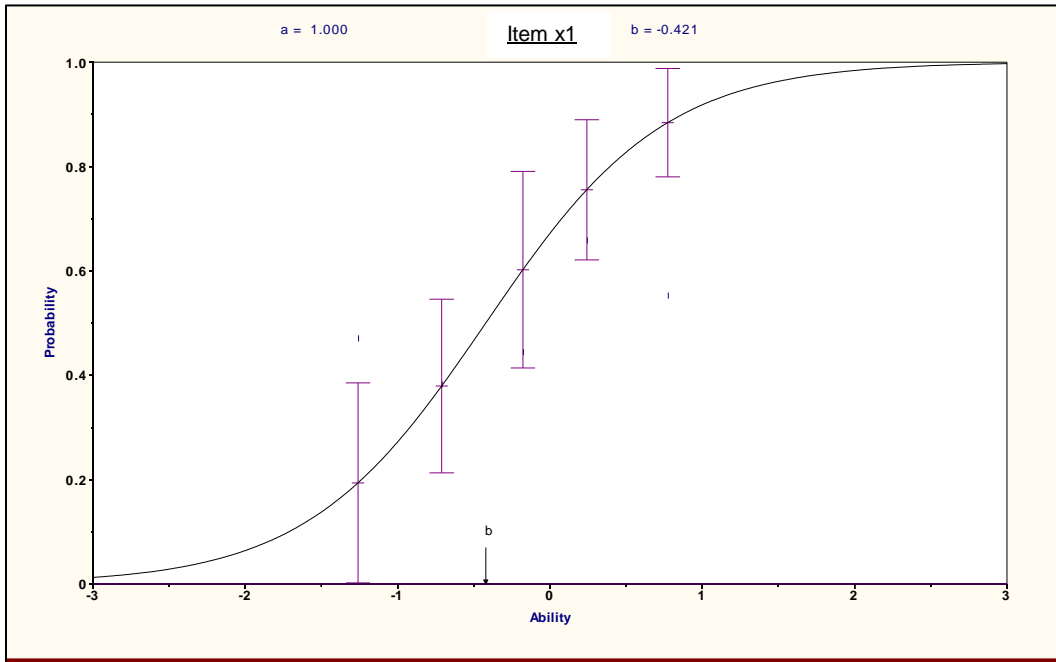
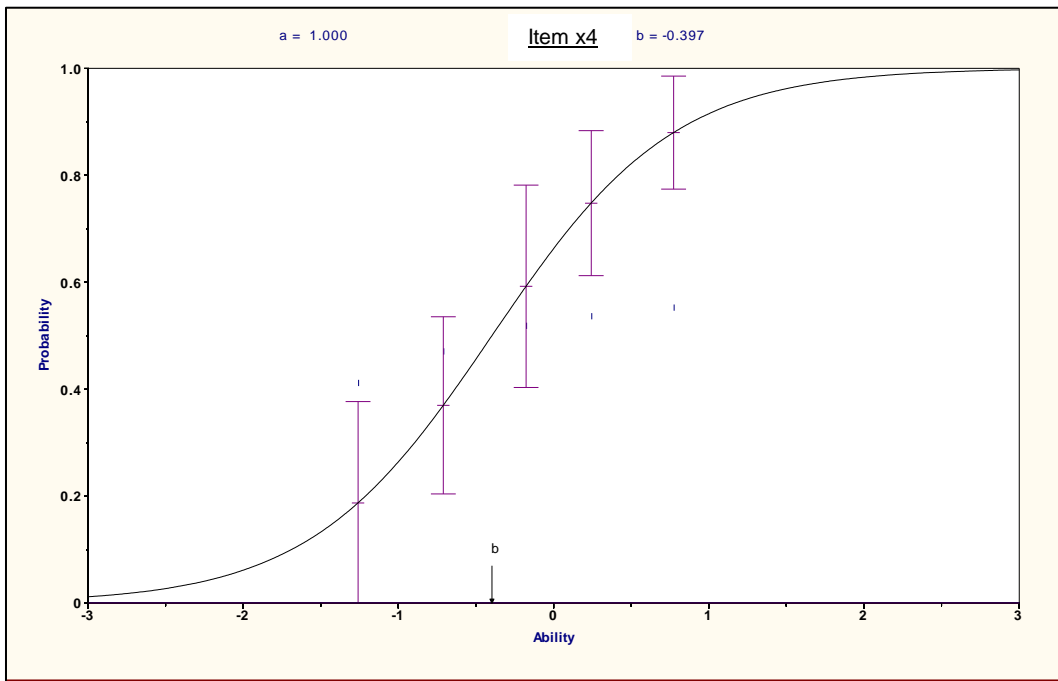
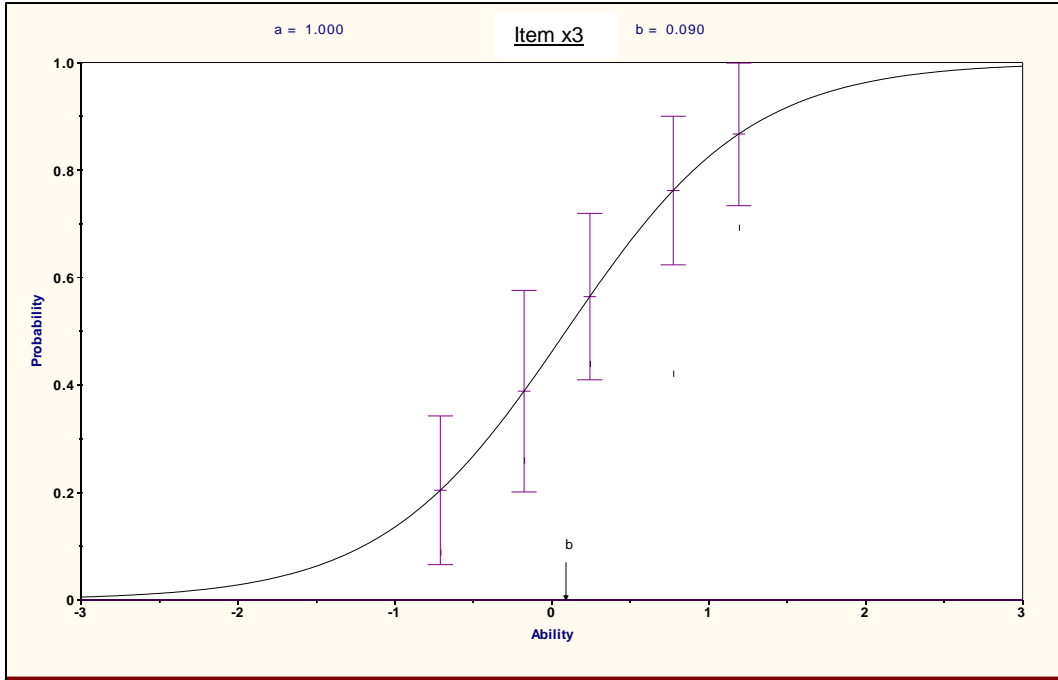
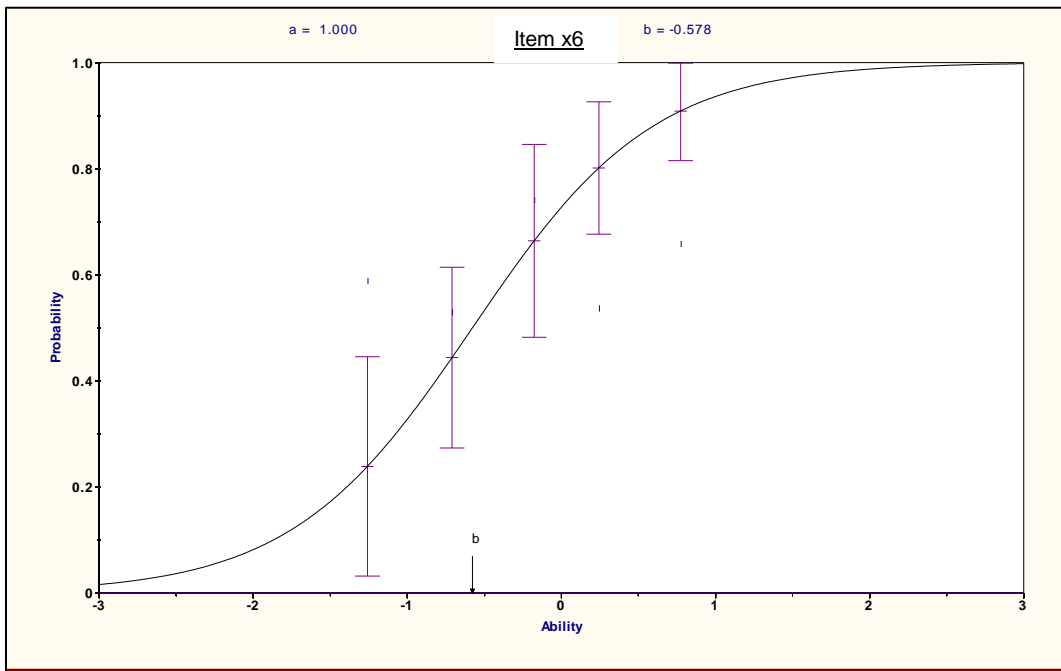
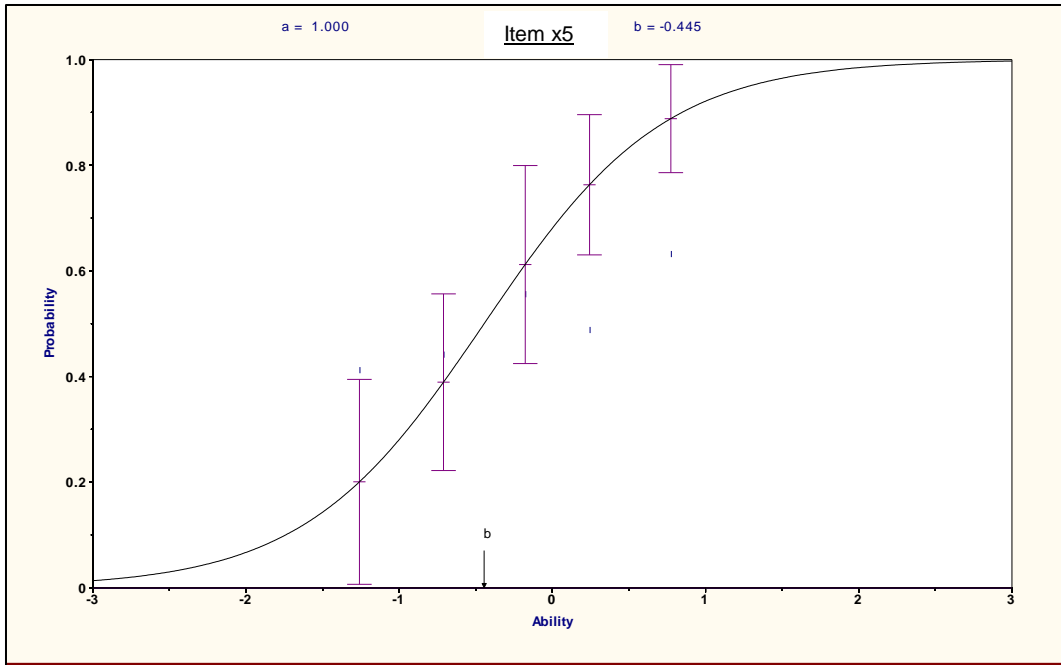
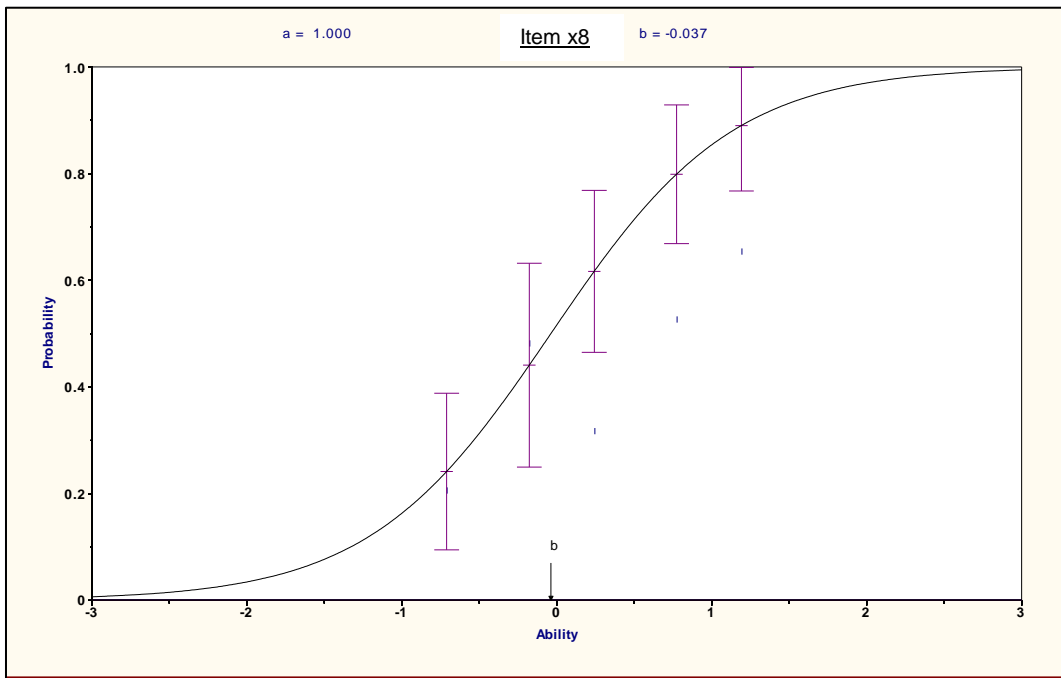
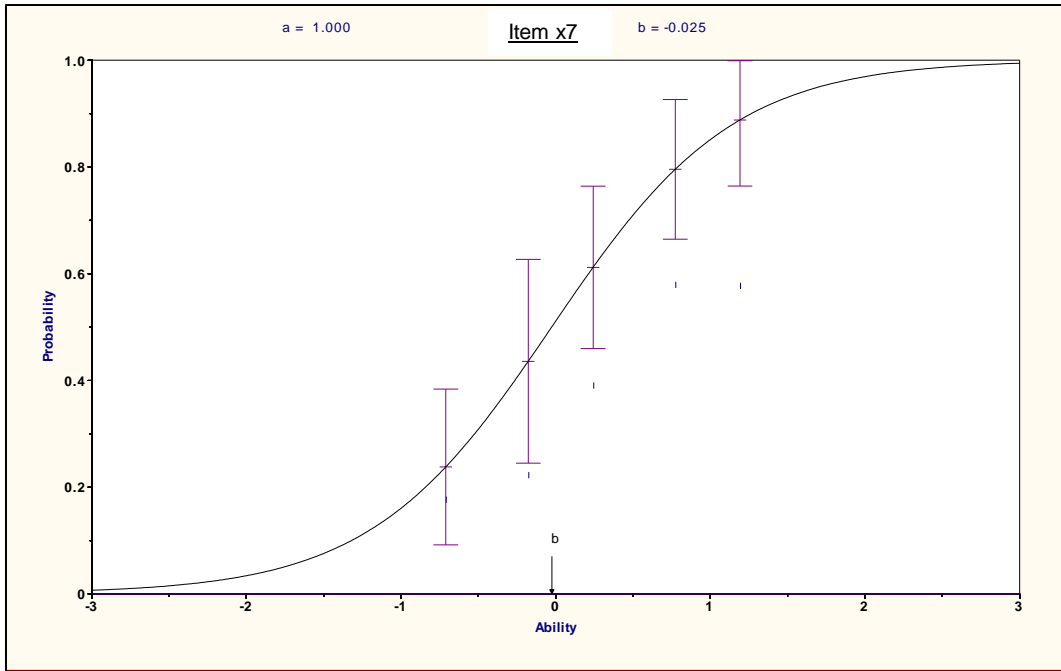


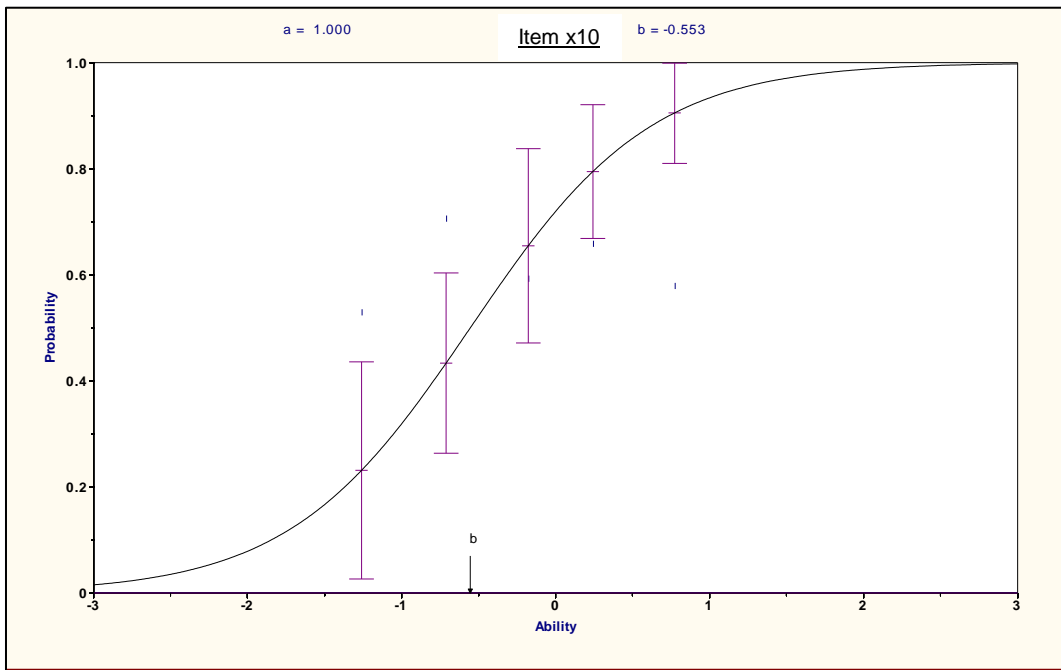
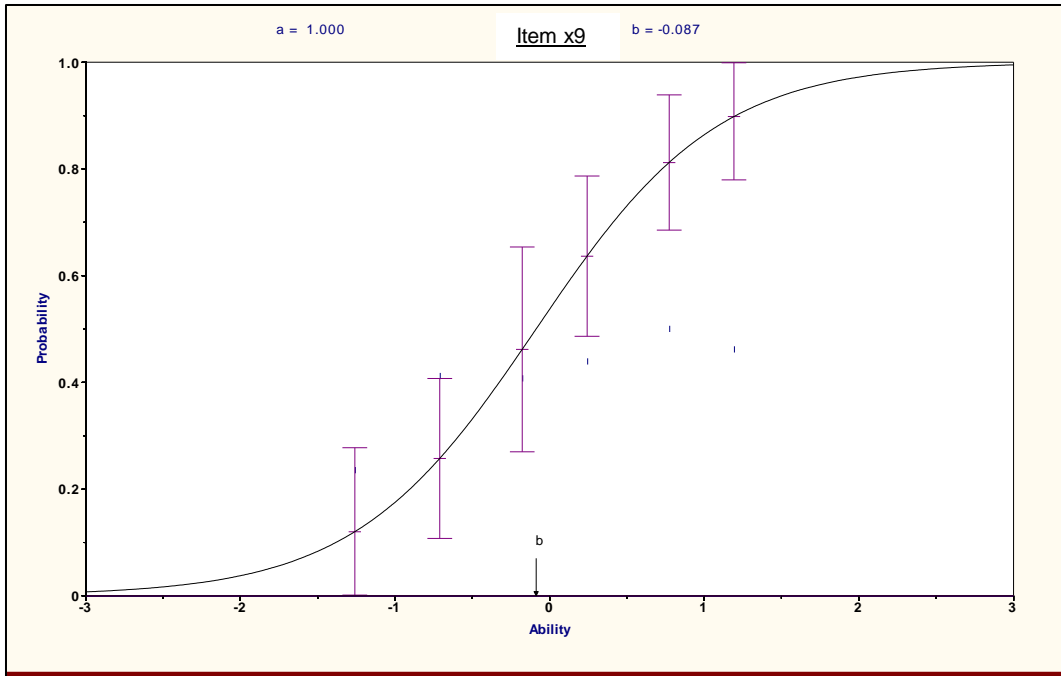
Figure 3. Empirical Versus Model Implied Item Characteristic Curves.

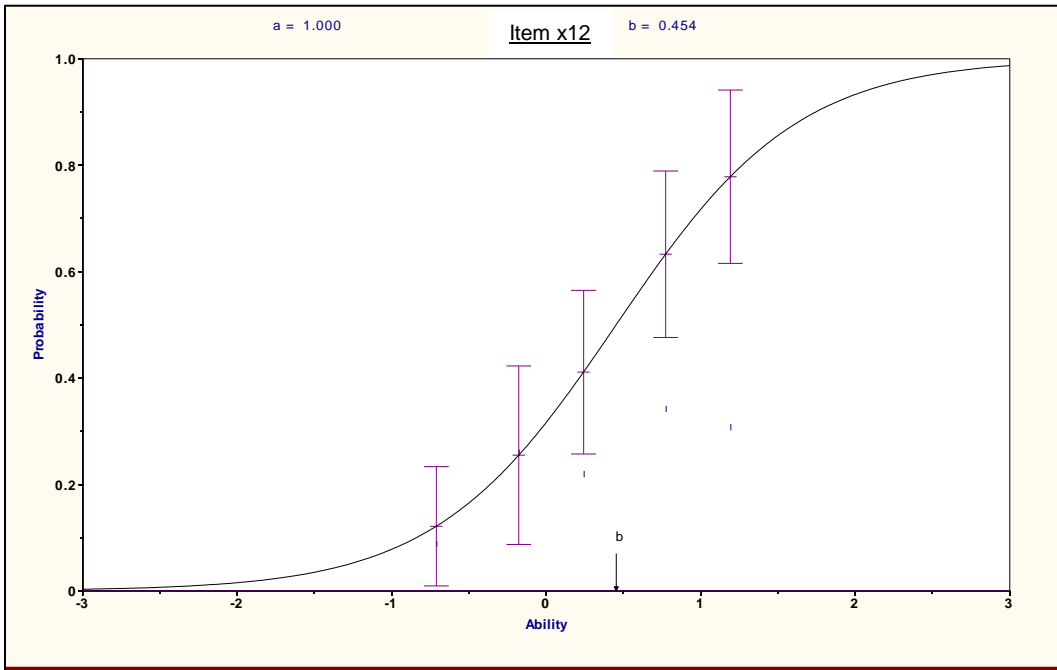
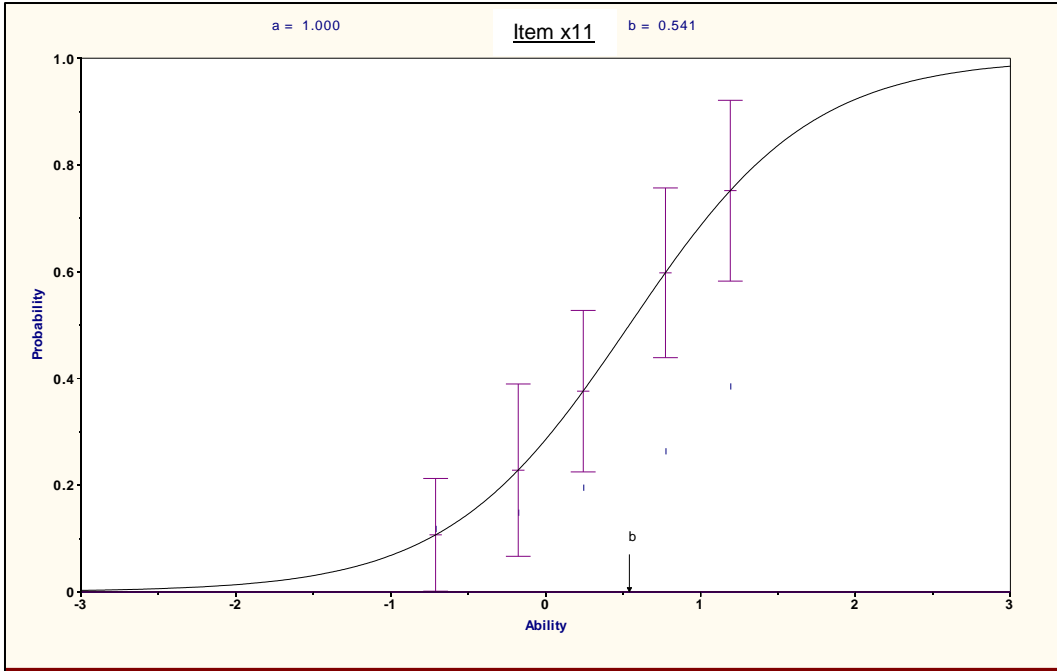


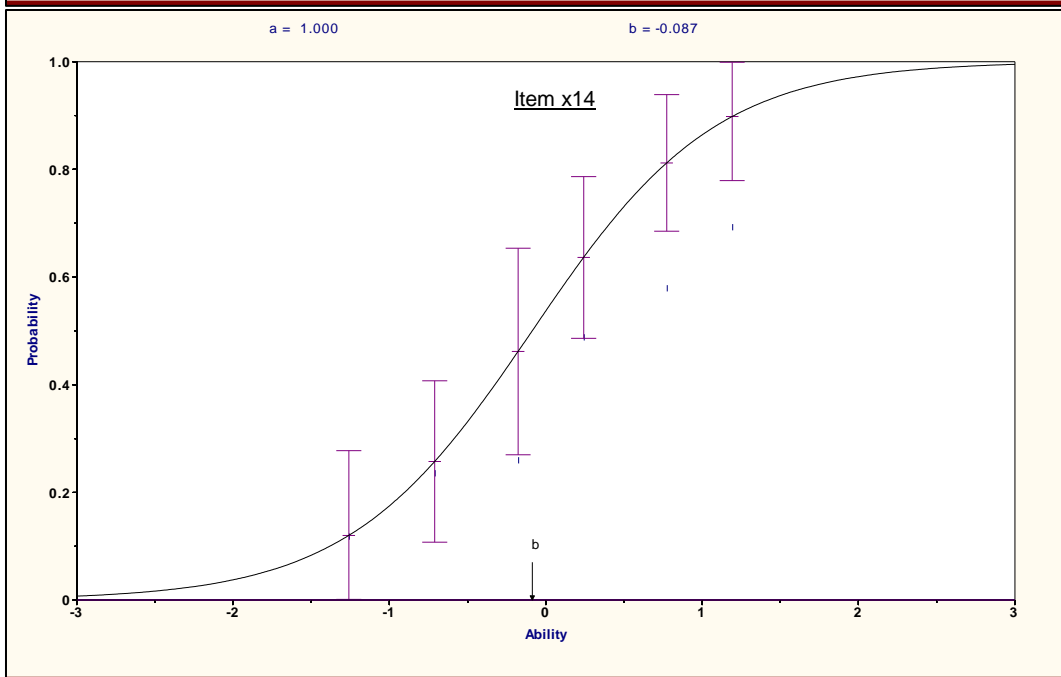
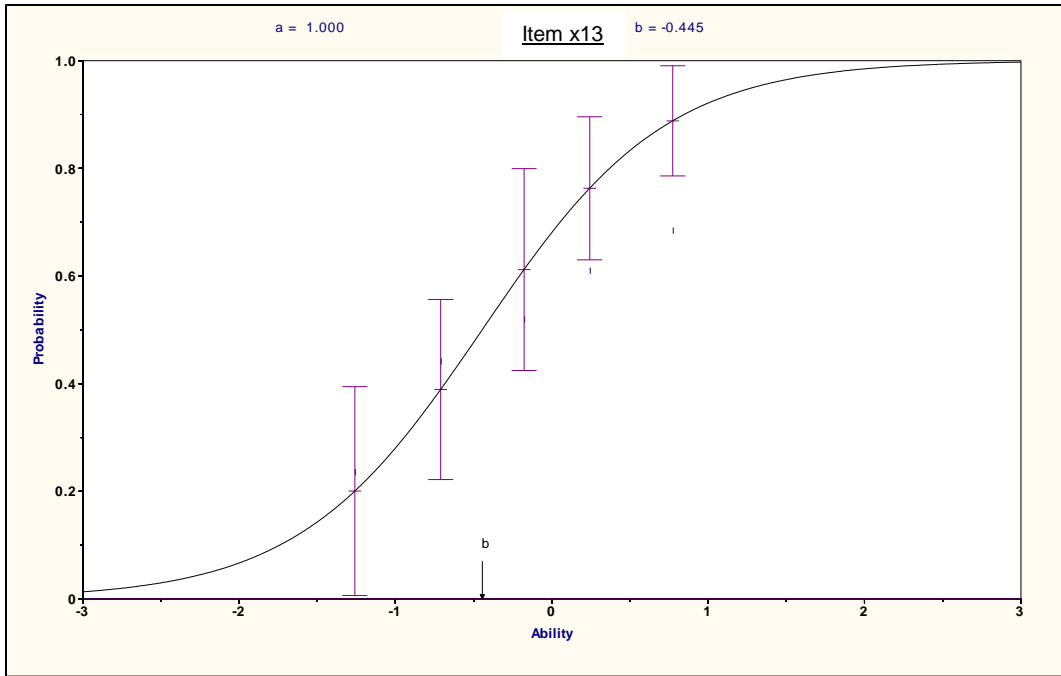


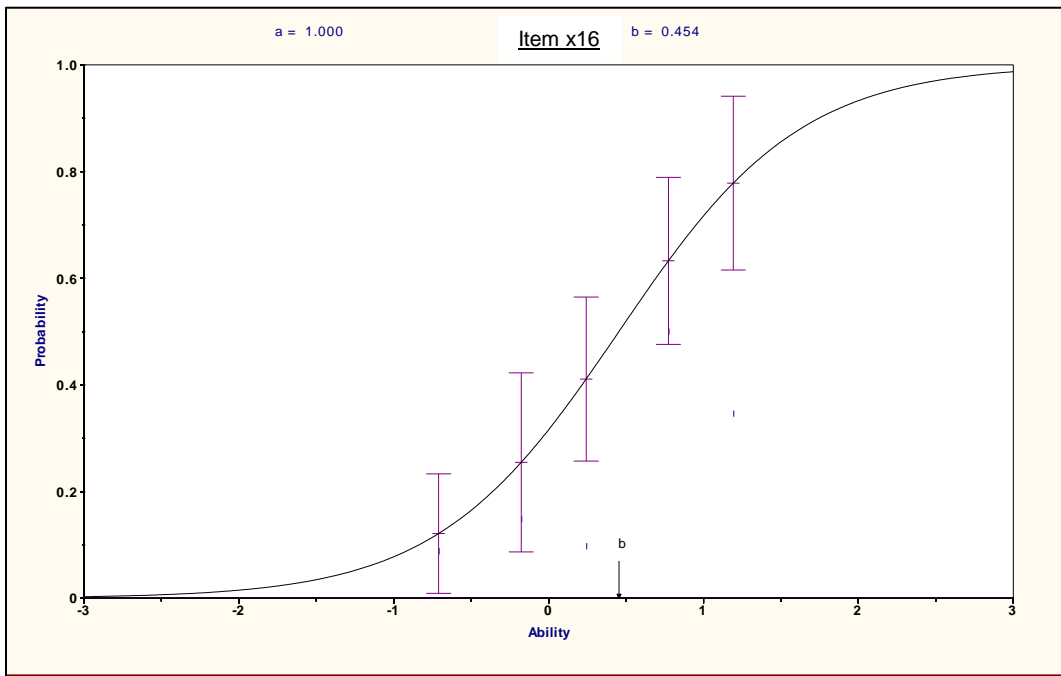
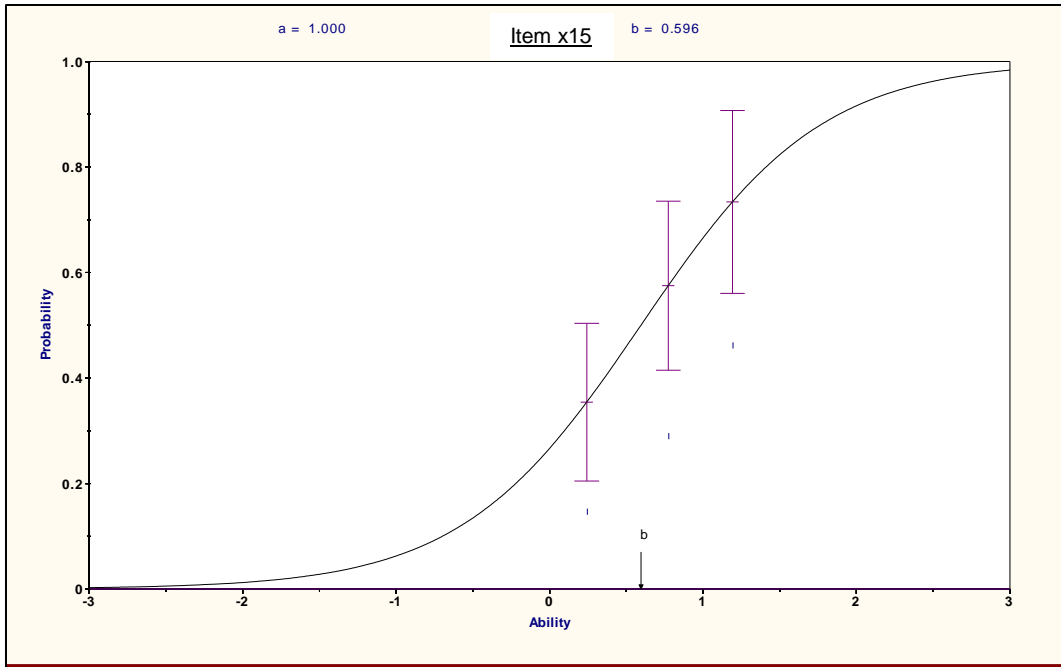


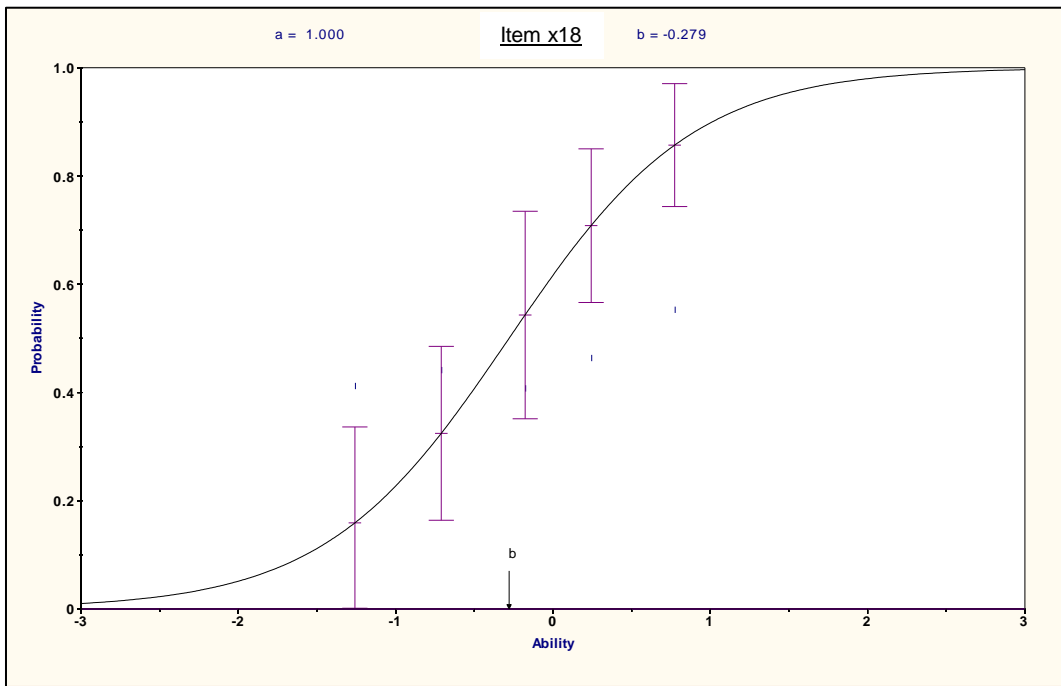
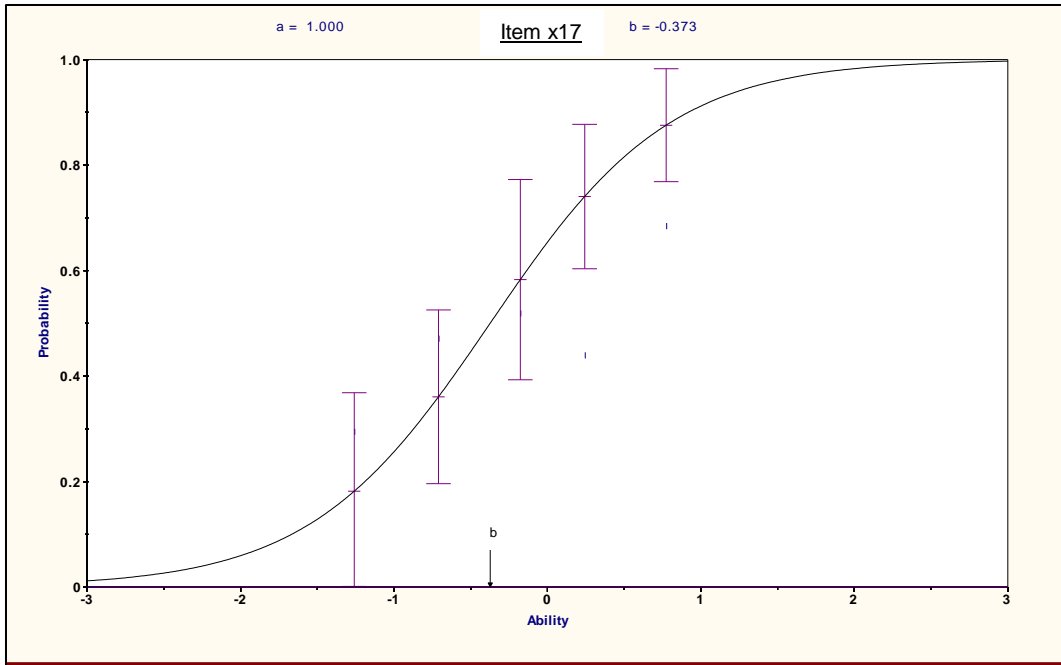


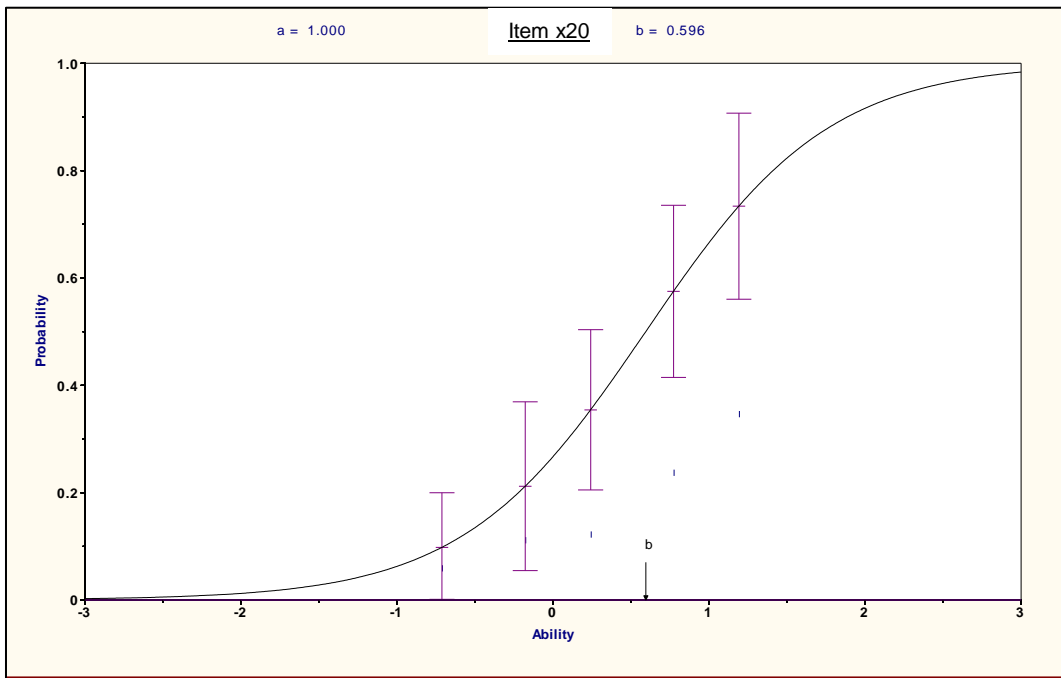
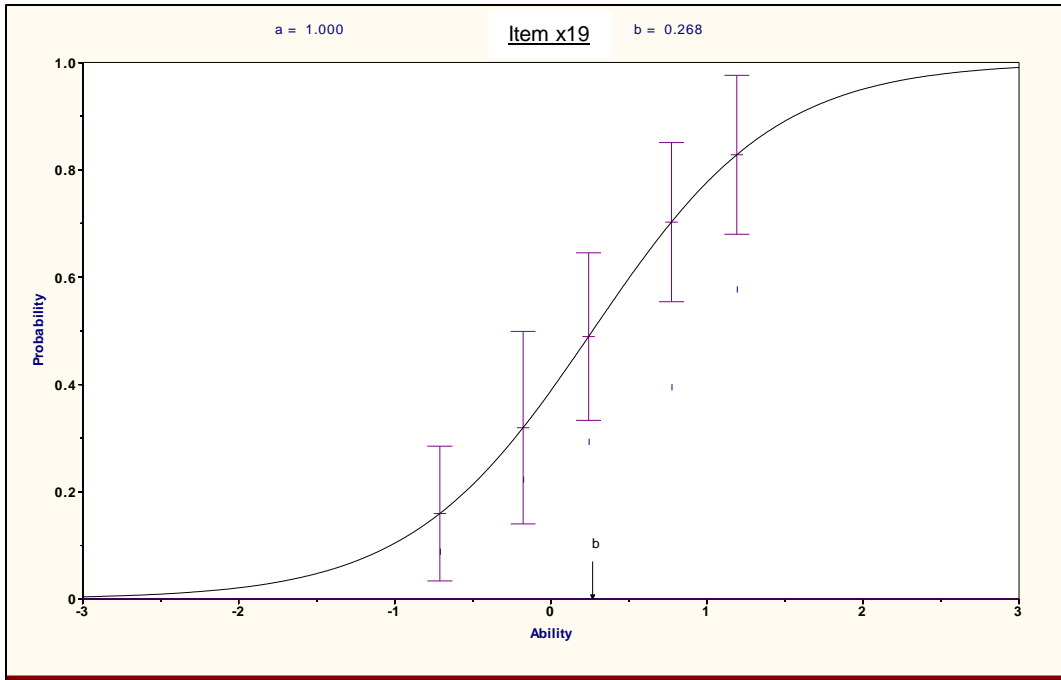


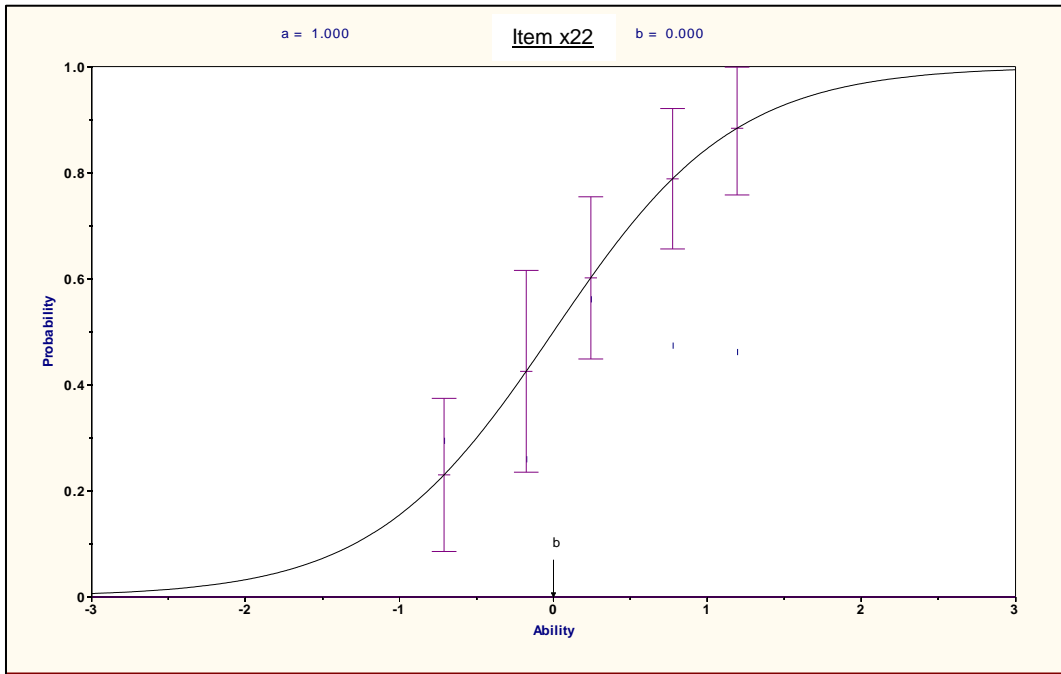
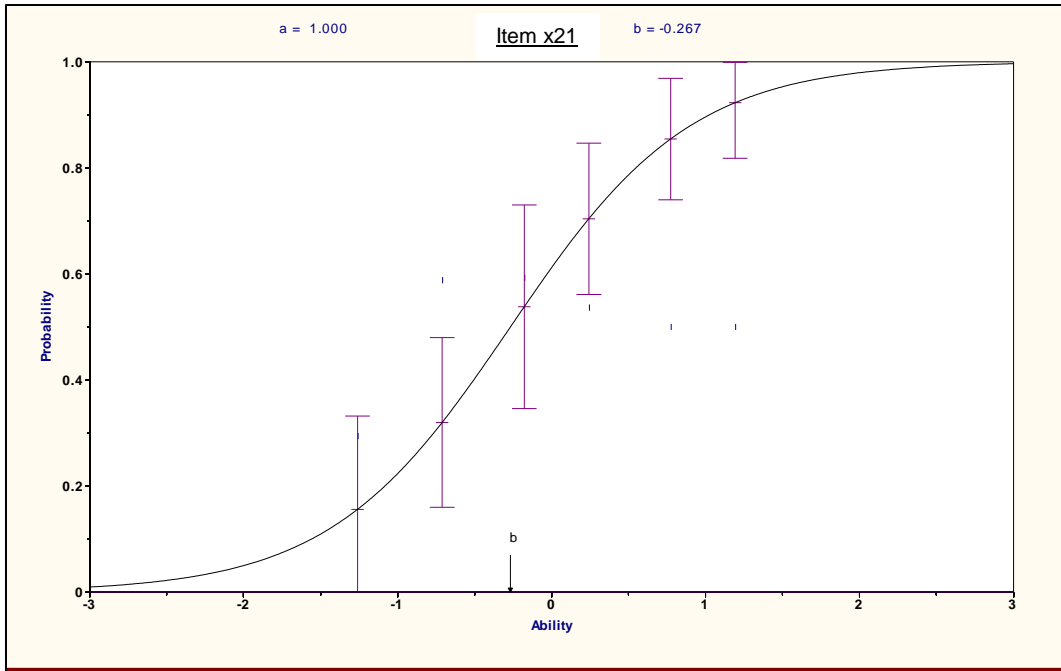


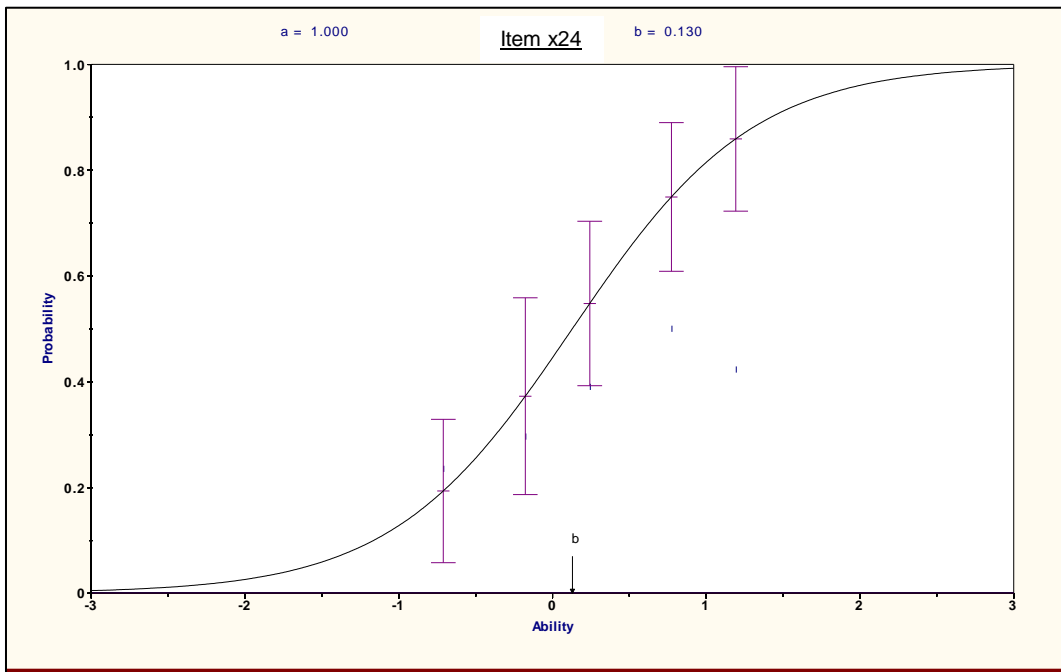
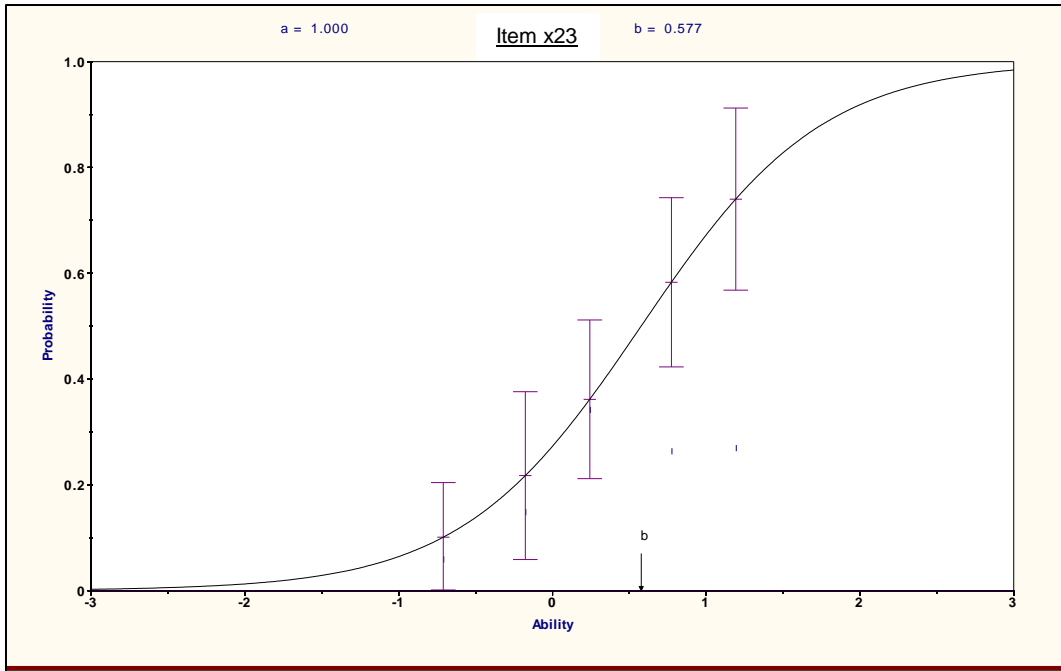


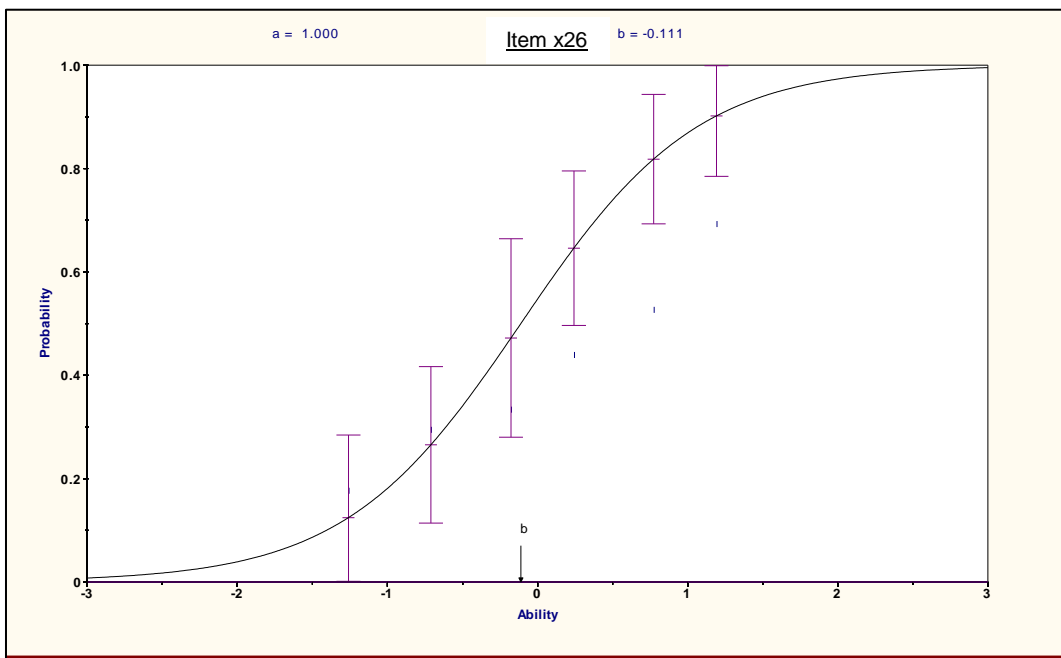
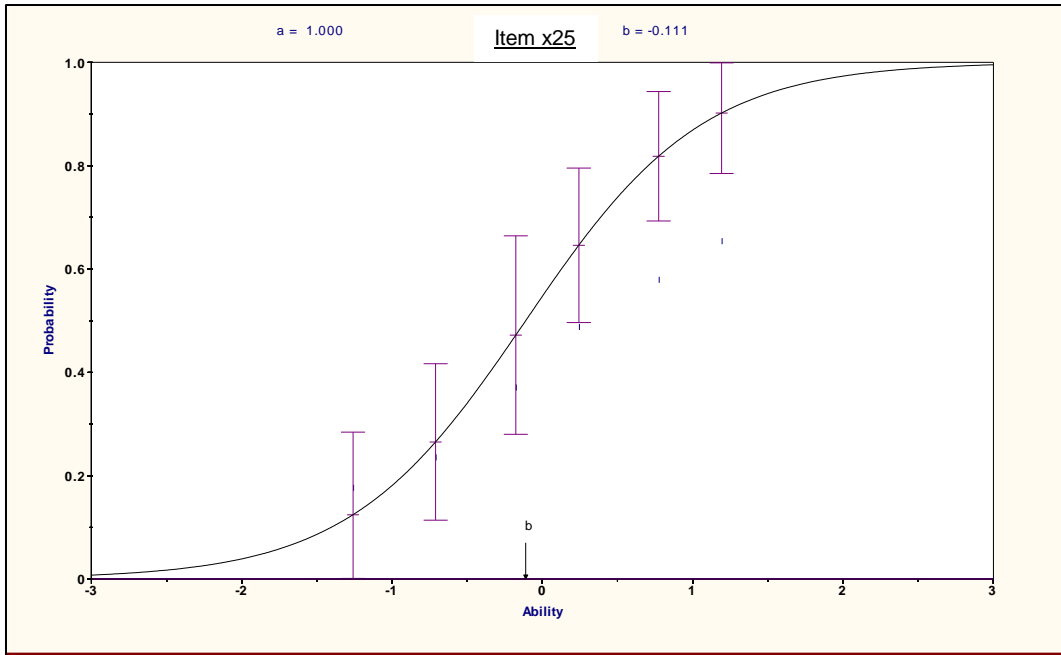


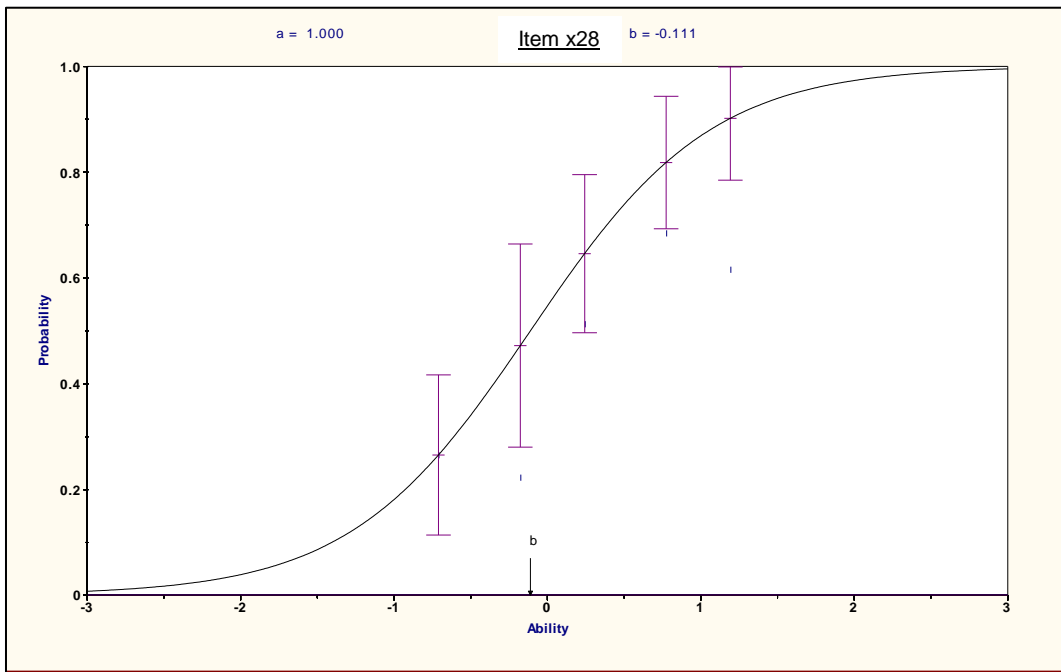
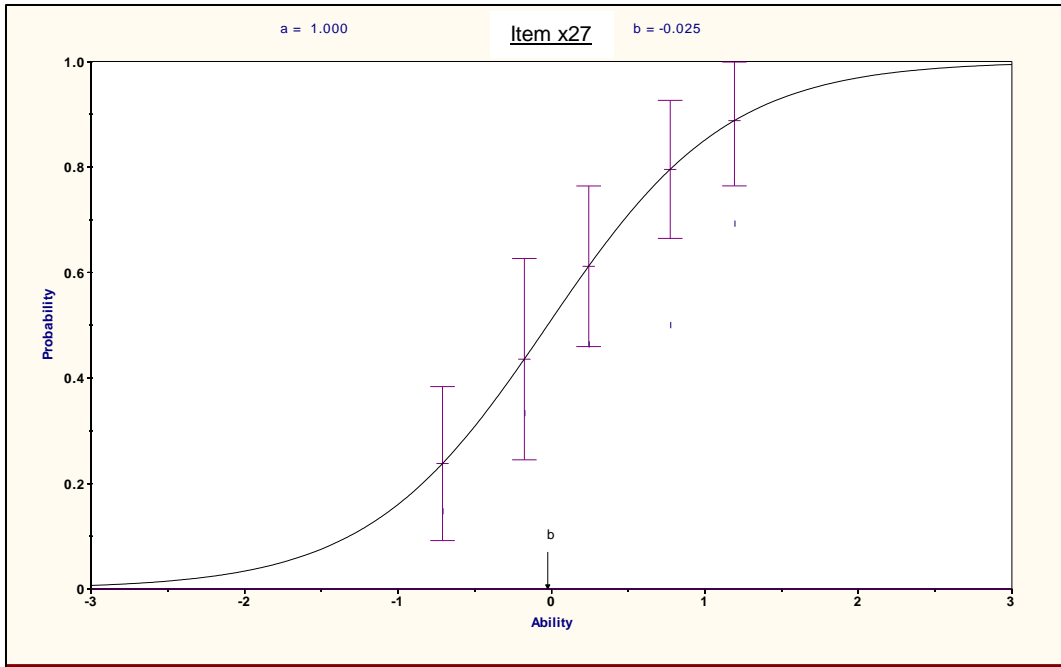


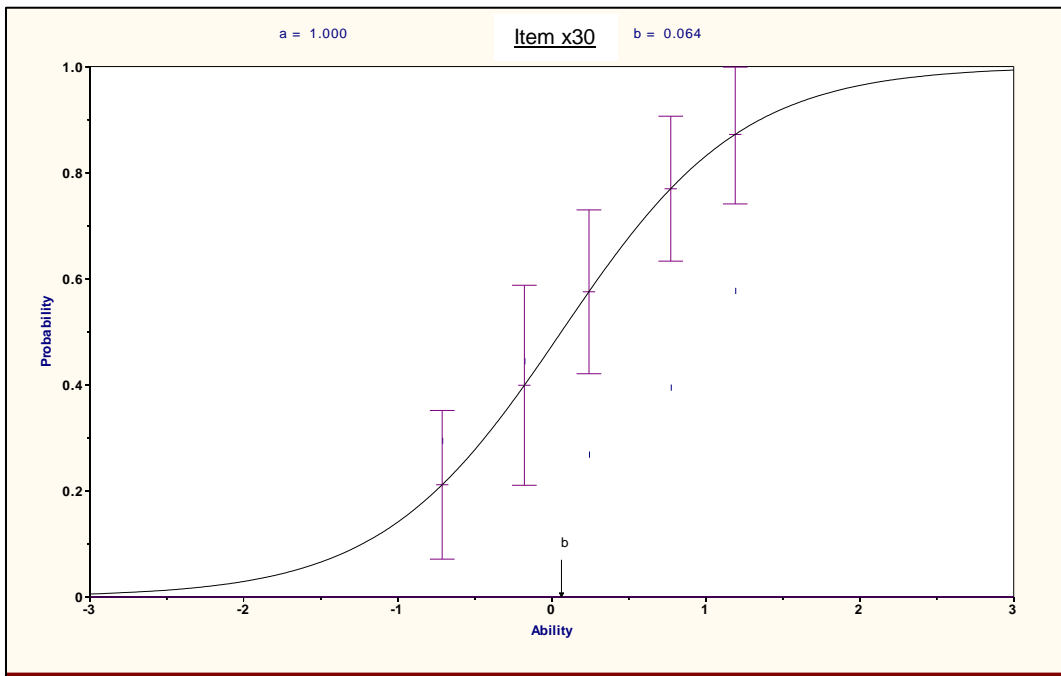
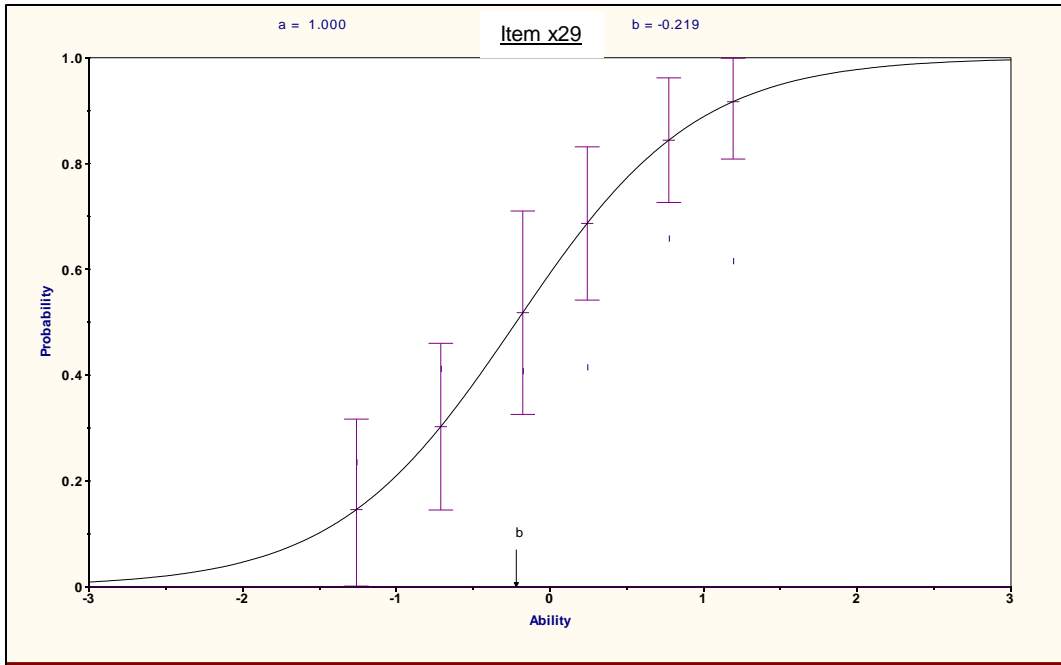


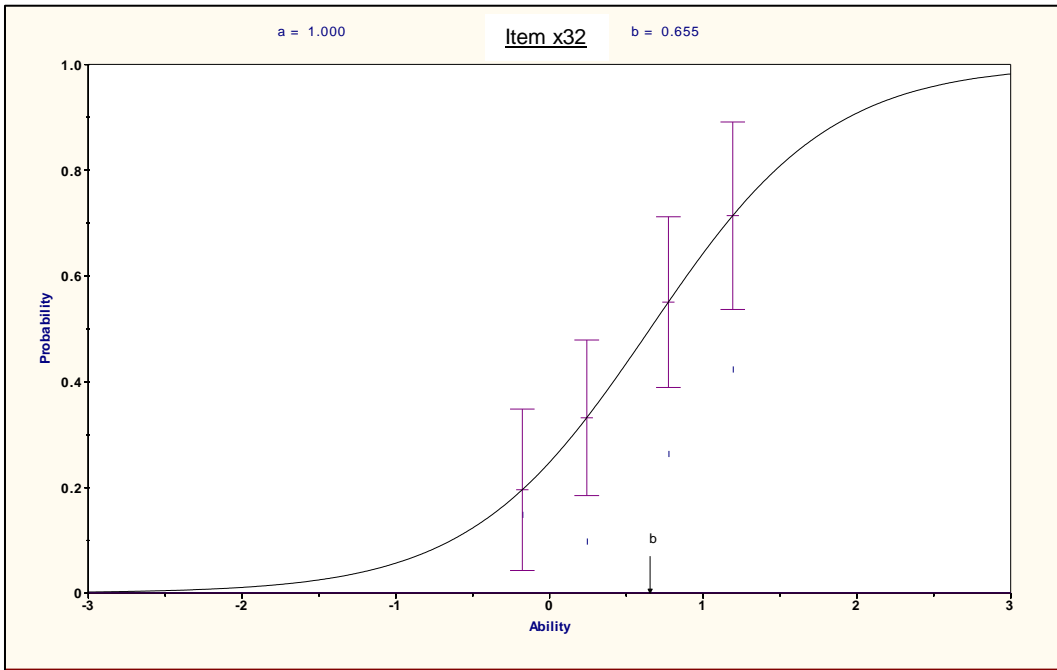
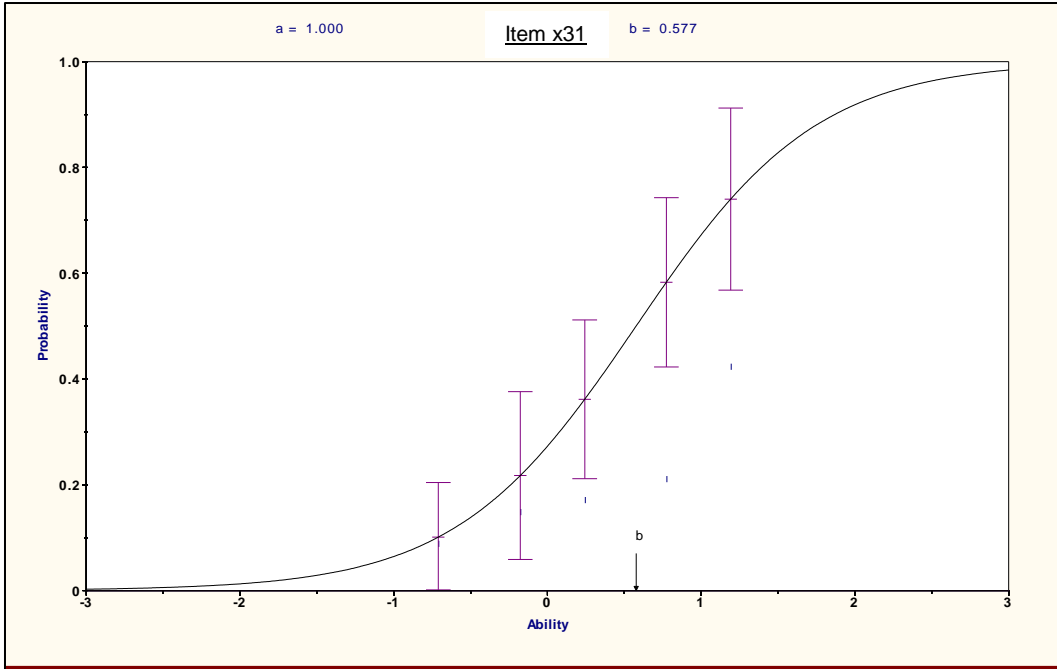


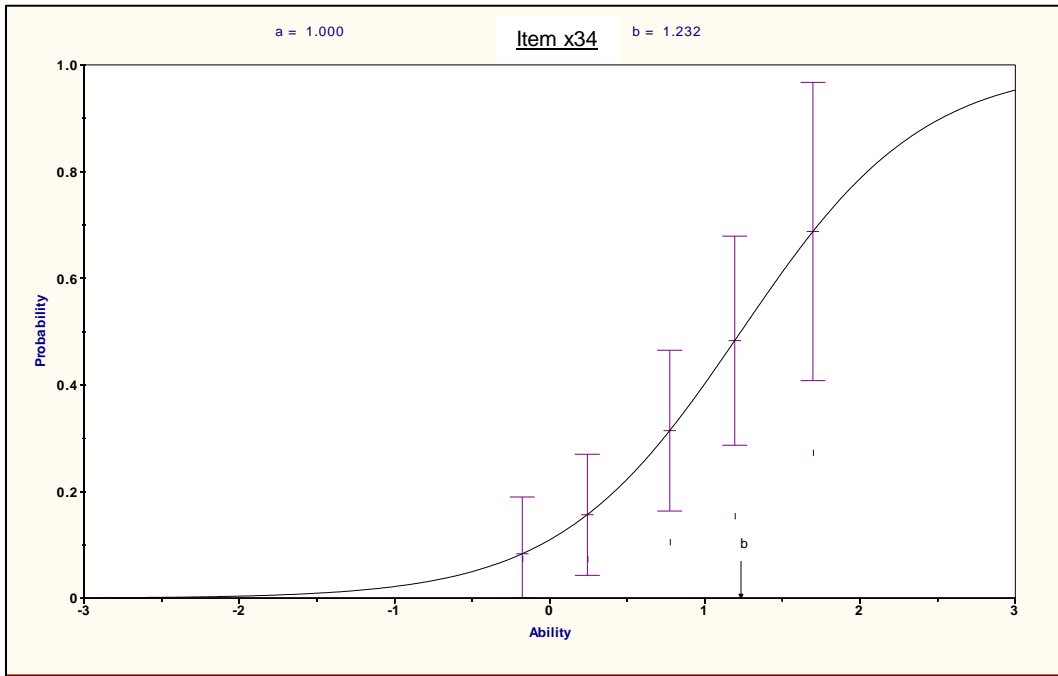
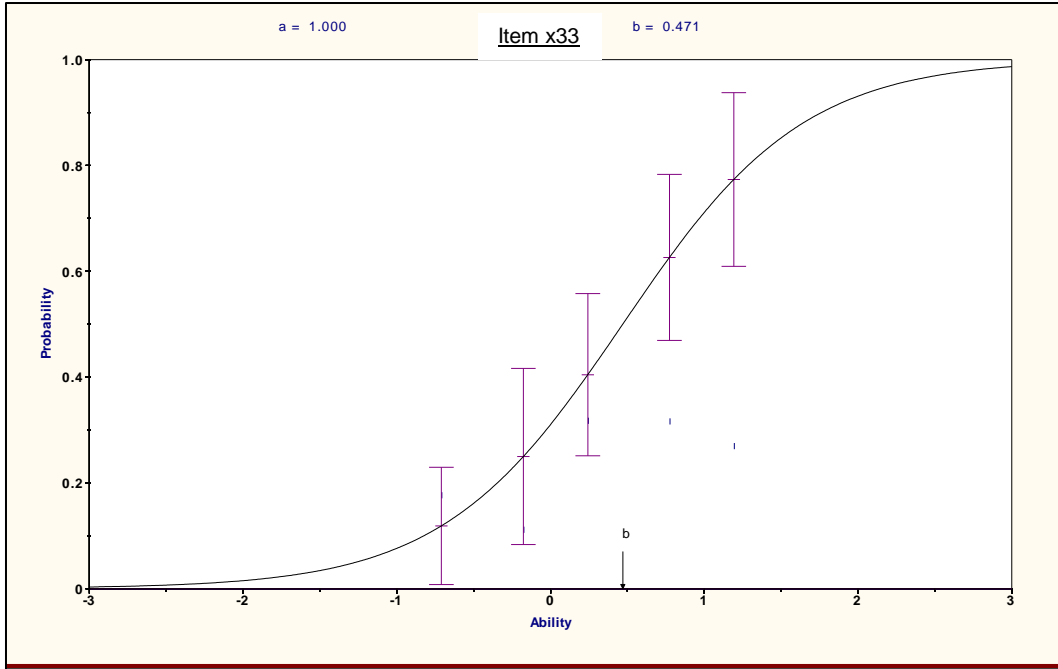












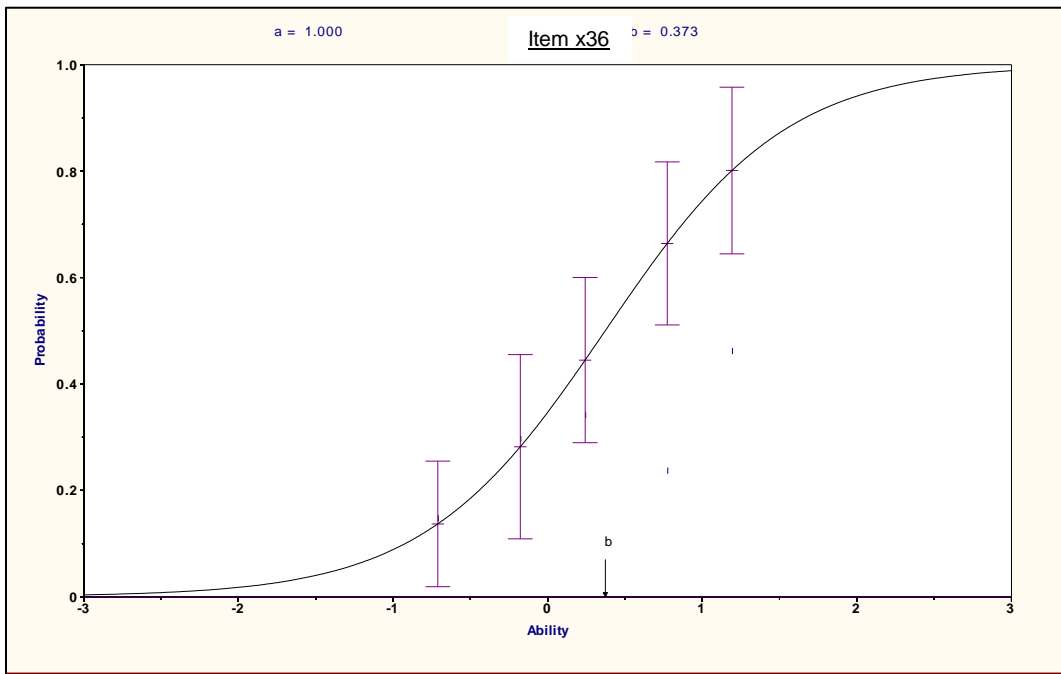
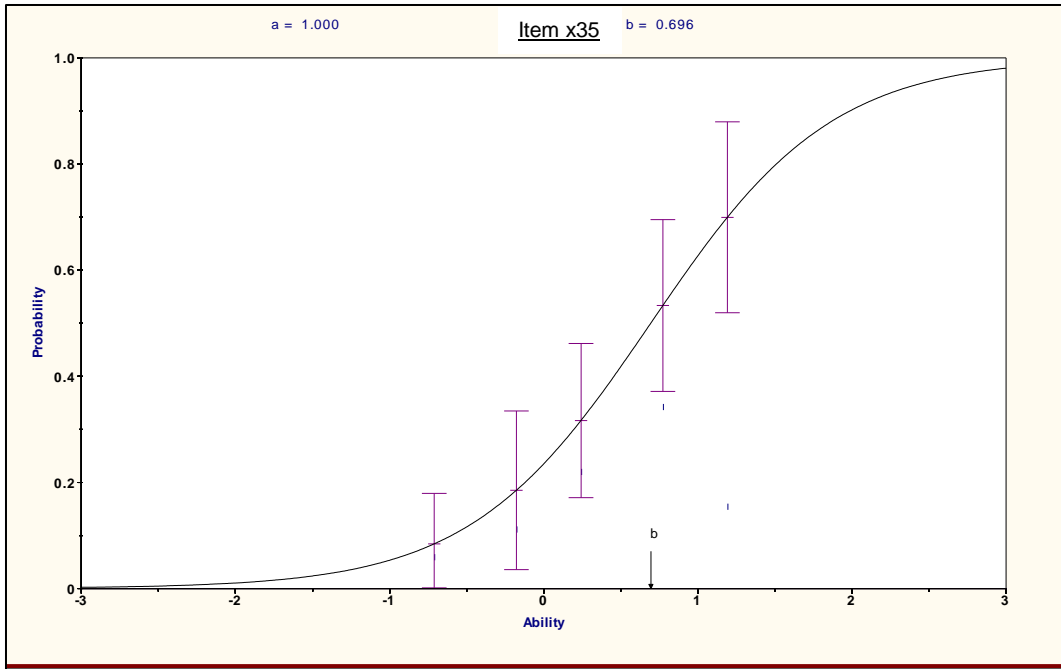


Figure 4. Rasch Person by Item Map.

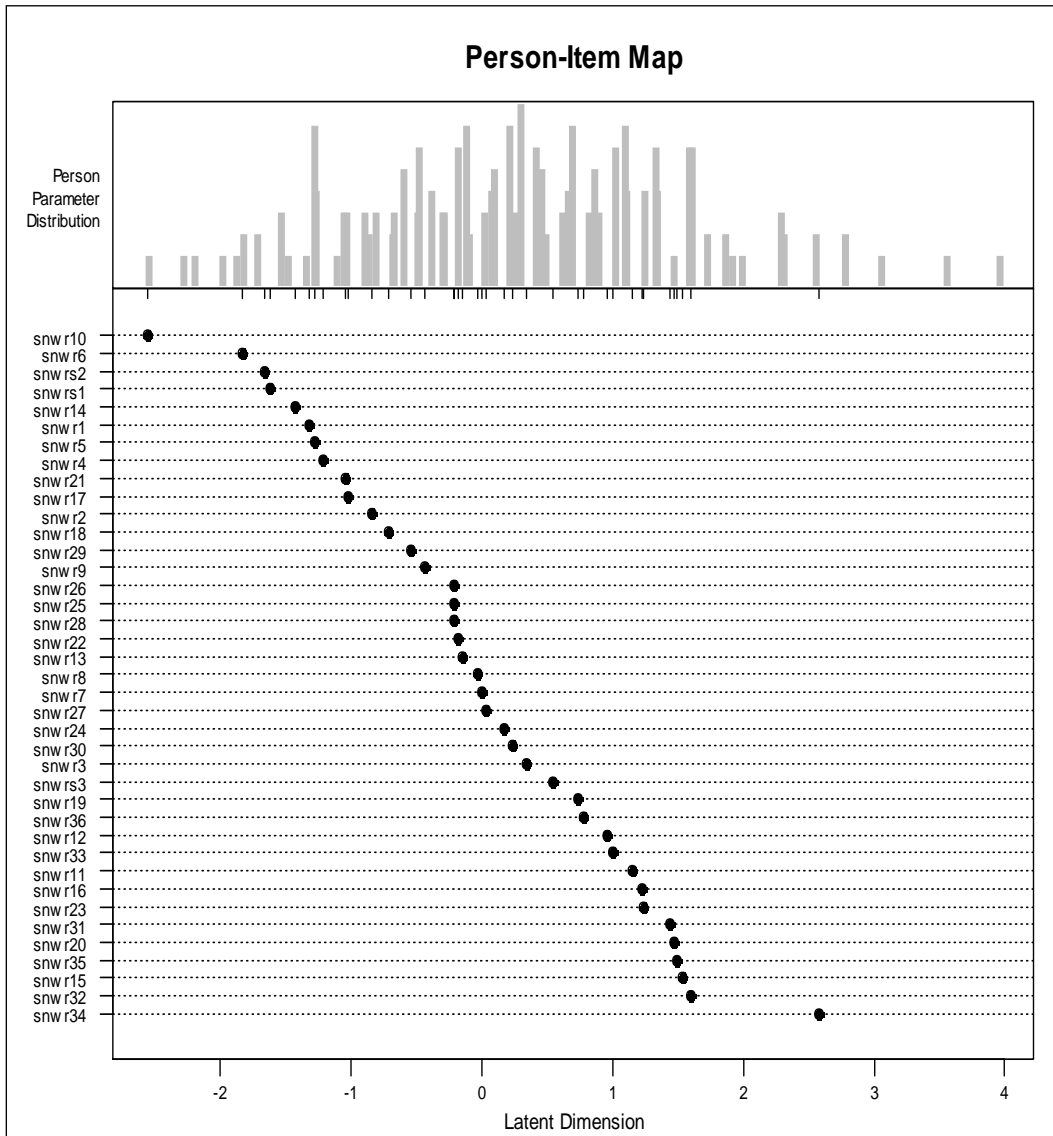


Figure 5. Plot of Total Information Function and Standard Error.

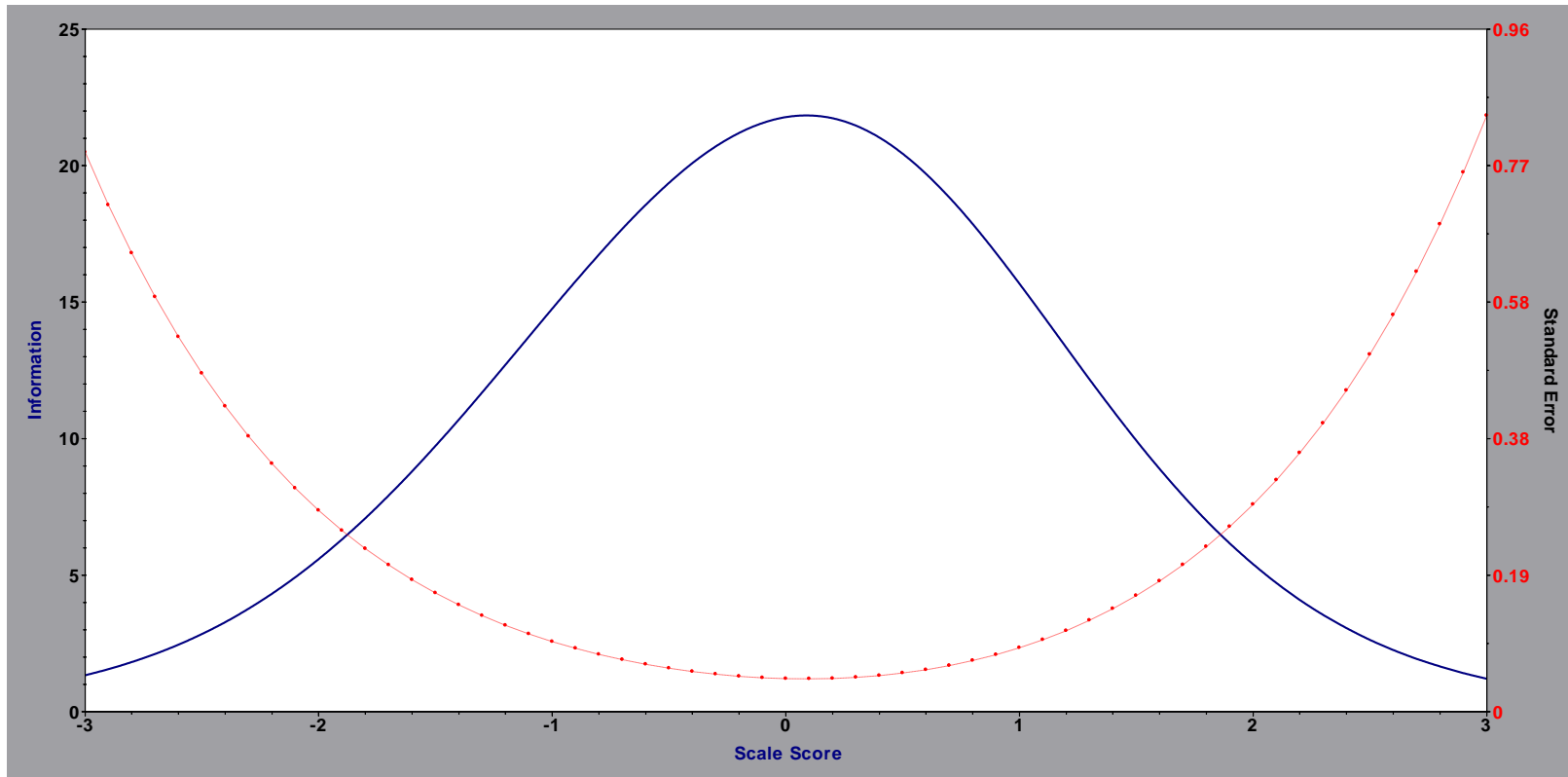


Figure 6. Plot of b Parameters for Rasch and Final LLTM Model.

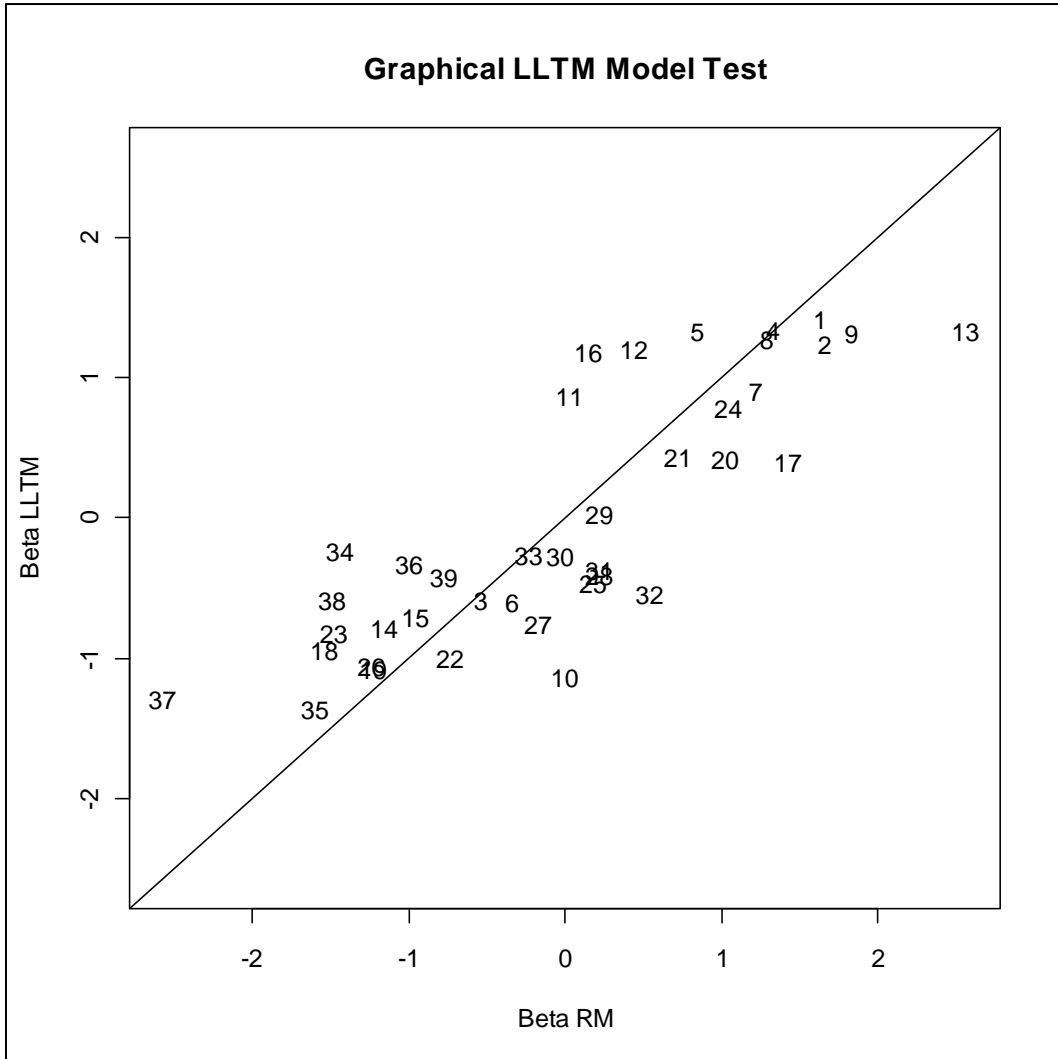
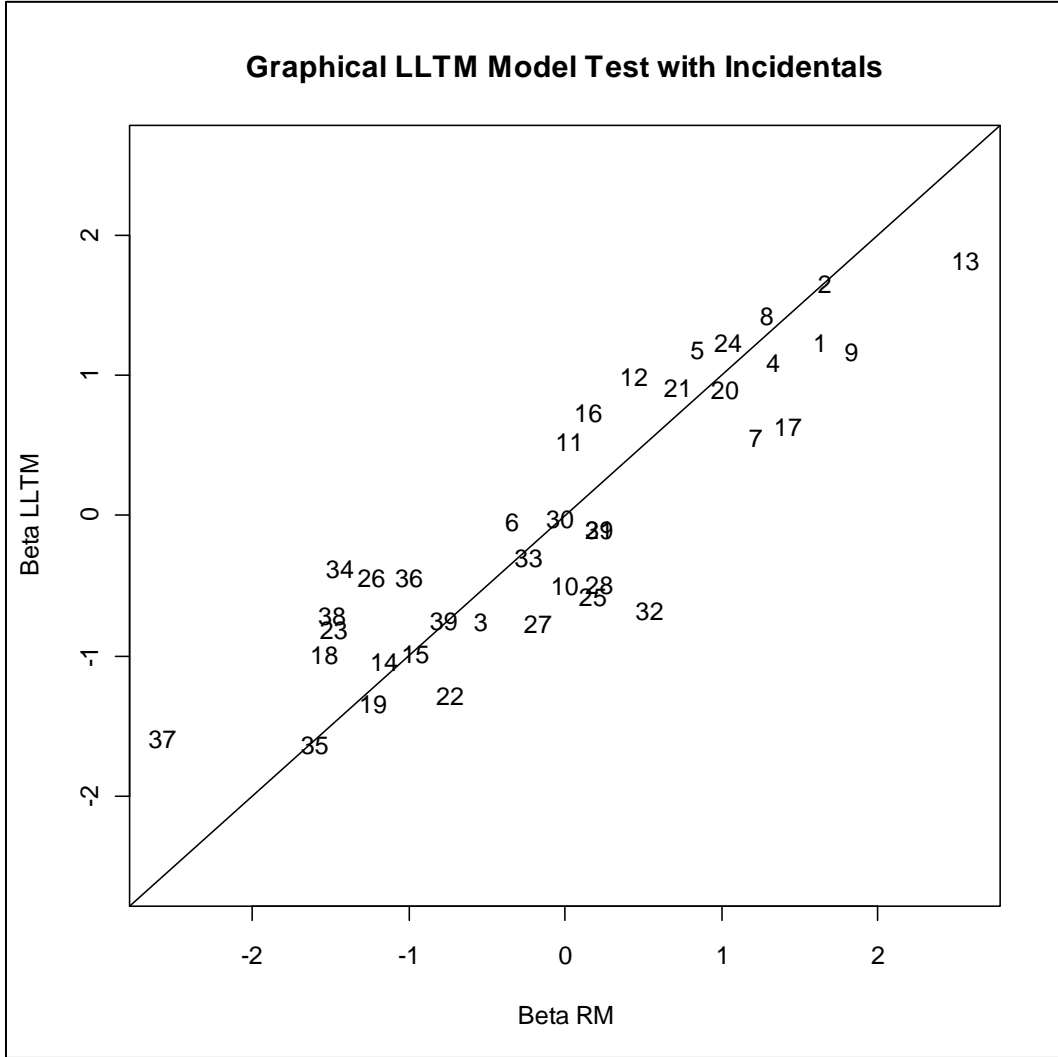


Figure 7. Plot of b Parameters for Rasch and LLTM Model with Incidentals.



APPENDIX A

PLANNED MISSING DATA DESIGN

| | syllables | pp | nd | A | B | C |
|----|-----------|-----|-----|---|---|---|
| 1 | 3 | hi | hi | X | X | |
| 2 | 3 | low | hi | X | X | |
| 3 | 3 | hi | low | X | X | |
| 4 | 3 | low | low | X | X | |
| 5 | 3 | hi | hi | | X | X |
| 6 | 3 | low | hi | | X | X |
| 7 | 3 | hi | low | | X | X |
| 8 | 3 | low | low | | X | X |
| 9 | 3 | hi | hi | X | | X |
| 10 | 3 | low | hi | X | | X |
| 11 | 3 | hi | low | X | | X |
| 12 | 3 | low | low | X | | X |
| 13 | 4 | hi | hi | X | | X |
| 14 | 4 | low | hi | X | | X |
| 15 | 4 | hi | low | X | | X |
| 16 | 4 | low | low | X | | X |
| 17 | 4 | hi | hi | X | X | |
| 18 | 4 | low | hi | X | X | |
| 19 | 4 | hi | low | X | X | |
| 20 | 4 | low | low | X | X | |
| 21 | 4 | hi | hi | | X | X |
| 22 | 4 | low | hi | | X | X |
| 23 | 4 | hi | low | | X | X |
| 24 | 4 | low | low | | X | X |
| 25 | 5 | hi | hi | X | X | |
| 26 | 5 | low | hi | X | X | |
| 27 | 5 | low | hi | X | | X |
| 28 | 5 | low | low | X | | X |
| 29 | 5 | hi | hi | | X | X |
| 30 | 5 | low | hi | | X | X |
| 31 | 5 | hi | low | X | X | |
| 32 | 5 | low | low | X | X | |
| 33 | 5 | hi | hi | X | | X |
| 34 | 5 | hi | low | | X | X |
| 35 | 5 | hi | low | X | | X |
| 36 | 5 | low | low | | X | X |

APPENDIX B

EXAMPLE NONWORD REPETITION TASK



SSLIC Spanish NWR Form B ID

Cohort

49093

Form **B**

A B C D

- Setup: Put headphones with microphone on child.
 - Plug the white audio splitter into the ipod.
 - Plug the black (not gray) cable on the child's headphones into the white splitter.
 - Plug your headphones into the white splitter also,
 - Plug the gray cable on the child's headphones into the holes labeled mic on the digital recorder.
 Put your headphones on so that one ear can hear the child speaking and the other can hear the recording playing on the ipod.
- Testing: Press record on the digital recorder and say (the mic from the child's headphones will pick up your voice but you must speak with a clear strong voice): ID: _____ SSLIC Spanish Nonword repetition form F, Tester: _____, Date: _____.
 - On the ipod go to Music ->playlists-> List F SPAN NWR SSLIC and hit the play button.
 - All of the directions for the child are included in the recording.
- Scoring: When you are testing the child you only need to indicate if the child says the whole word correctly or not by circling 1 or 0 in the final column. All other scoring will be completed in the lab.

Date (MM/DD/YY)

| | | | | | |
|--|--|--|--|--|--|
| | | | | | |
|--|--|--|--|--|--|

Shade Circles Like This--> ●
 Not Like This--> ⊗

| Sample Items | Transcription | Child's production | Phon # / total | Correct 1 0 |
|--------------|-------------------------------|--------------------|---|---|
| snwrs1 | mi/por /mipor/ | | <input type="checkbox"/> __/5 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwrs2 | ba/te/ra /bæteræ/ | | <input type="checkbox"/> __/6 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwrs3 | de/sa/gri/dar /dɛsægridær/ | | <input type="checkbox"/> <input type="checkbox"/> __/10 | <input type="radio"/> 1 <input type="radio"/> 0 |
| | Transcription | Child's production | Phon # / total | Correct 1 0 |
| snwr_9 | re/ga/do /rɛgædo/ | | <input type="checkbox"/> __/6 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_10 | o/sa/ko /osæko/ | | <input type="checkbox"/> __/5 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_11 | gle/fli/tom /glɛflitom/ | | <input type="checkbox"/> __/9 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_12 | fli/mes/gra /flimɛsgræ/ | | <input type="checkbox"/> __/9 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_1 | ka/te/ra /kæteræ/ | | <input type="checkbox"/> __/6 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_2 | a/me/bo /æmɛbo/ | | <input type="checkbox"/> __/5 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_3 | ba/tro/kria /bætrokrjæ/ | | <input type="checkbox"/> __/9 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_4 | cha/jo/ve /tʃæhobɛ/ | | <input type="checkbox"/> __/6 | <input type="radio"/> 1 <input type="radio"/> 0 |

SSLIC Spanish NWR Form B



49093

| | | | | |
|---------|---|--|--|---|
| snwr_21 | a/ra/ ka /da /æ r æ k æ d æ/ | | <input type="checkbox"/> _/7 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_22 | de/ma/ kro /do /d ε m æ k r o d o/ | | <input type="checkbox"/> _/9 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_23 | je/tos/ ma /bra /h ε t o s m æ b r æ/ | | <input type="checkbox"/> <input type="checkbox"/> _/10 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_24 | es/lo/ mi /gre /ε s l o m i g r ε/ | | <input type="checkbox"/> _/9 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_13 | e/a/ te /ra /ε æ t ε r æ/ | | <input type="checkbox"/> _/6 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_14 | cha/ta/ te /ro /tʃ æ t æ t ε r o/ | | <input type="checkbox"/> _/8 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_15 | gre/bra/ fe /ar /g r ε b r æ f ε æ r/ | | <input type="checkbox"/> <input type="checkbox"/> _/10 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_16 | va/pla/ lle /ron /b æ p l æ dʒ ε r o n/ | | <input type="checkbox"/> <input type="checkbox"/> _/10 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_33 | ma/sa/ ji / si /ta /m æ s æ h i s i t æ/ | | <input type="checkbox"/> <input type="checkbox"/> _/10 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_34 | tes/no/ kra / wi a /t ε s n o k r æ w j æ/ | | <input type="checkbox"/> <input type="checkbox"/> _/11 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_35 | da/ne/ ni / bo /do r /d æ n ε n i b o d o r/ | | <input type="checkbox"/> <input type="checkbox"/> _/11 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_36 | wi/ba/ ja / do /se /w i b æ h æ d o s ε/ | | <input type="checkbox"/> <input type="checkbox"/> _/10 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_25 | ma/jo/ me / ta / na /m æ h o m ε t æ n æ/ | | <input type="checkbox"/> <input type="checkbox"/> _/10 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_26 | e/si/ ri / to /ra /ε s i r i t o r æ/ | | <input type="checkbox"/> _/9 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_27 | ta/ra/ ba / te /ra /t æ r æ b æ t ε r æ/ | | <input type="checkbox"/> <input type="checkbox"/> _/10 | <input type="radio"/> 1 <input type="radio"/> 0 |
| snwr_28 | se/va/ ti / ro / si /s ε b æ t i r o s i/ | | <input type="checkbox"/> <input type="checkbox"/> _/10 | <input type="radio"/> 1 <input type="radio"/> 0 |

APPENDIX C
SAMPLE SYNTAX

MPLUS Confirmatory Factor Analysis: Test of Unidimensionality

TITLE: CFA - Dimensionality Test of the items;

DATA: FILE IS "C:\DATA\AIG\AIG-Data.csv";

VARIABLE: NAMES ID snwr1-snwr36;

USEV = snwr1-snwr36;

CATEGORICAL = snwr1-snwr36;

MISSING ARE ALL (-99);

ANALYSIS:

MODEL: F1 by snwr1-snwr36;

BILOG-MG: Rasch Model

Rasch model

```
>GLOBAL DFName = 'C:\Data\AIG\AIG-Data.dat',  
  NPArm = 1,  
  SAVe;  
>SAVE MASTer = 'SNWR.MAS',  
  CALib = 'SNWR.CAL',  
  PARm = 'SNWR.PAR',  
  SCORe = 'SNWR.SCO',  
  COVariance = 'SNWR.COV',  
  TSTat = 'SNWR.TST',  
  POSt = 'SNWR.POS',  
  EXPEcted = 'SNWR.EXP',  
  ISTat = 'SNWR.IST';  
>LENGTH NITems = (36);  
>INPUT NTOtal = 36,  
  NALt = 2,  
  NIDchar = 3;  
>ITEMS ;  
>TEST1 TNAmE = 'SNWR',  
  INUmber = (1(1)36);  
(3A1, 36A1)  
>CALIB PLOt = 1.0000,  
  ACCel = 1.0000,  
  RASch;  
>SCORE ;
```

R Syntax for eRm Package

```
#####  
# Set the working directory to where the data files are  
#  
#####  
setwd("C:/Data/AIG/R/")  
library(eRm) #Load eRm Package  
  
#####  
# Set up array "NW" to store item responses  
# Row= 1 person, Columns=items1-36  
# Missing data = NA -> required by eRm package  
#####  
  
NW <- array(NA, c(215, 36))  
NW<- read.table("AIG-Data.txt", header=TRUE)  
  
#####  
# Set up array "Char" to store item characteristics  
# In other words the W matrix  
# Char Elements[PP, ND, PP, ND, Syl, Phon, Syl_PP, Syl_ND, ConClust]  
#####  
  
Char <- array(NA, c(40, 10))  
Char <- read.table("Item_Characteristics.txt", header=TRUE)  
  
#####  
#Rasch Code  
#####  
  
resRM <- RM(NW) #Runs Rasch model  
summary(resRM) #Summary statistics  
  
#####  
# LLTM Code  
#####  
  
#Design Matrix – Can choose any from Char Martrix  
W <- matrix(c(Char$ND, Char$PP, Char$Phon), ncol=3)  
  
res1 <- LLTM(NW, W = W) #Run LLTM  
summary(res1) #Summary statistics
```



APPENDIX D

Q MATRIX OF ITEM FEATURES

| | Syllable | Phone | PP | ND | CC | BegVo | EndVo |
|----|----------|-------|------|----|----|-------|-------|
| 1 | 3 | 6 | 0.11 | 13 | 0 | 0 | 1 |
| 2 | 3 | 5 | 0.01 | 4 | 0 | 1 | 1 |
| 3 | 3 | 9 | 0.08 | 0 | 1 | 0 | 1 |
| 4 | 3 | 6 | 0.01 | 0 | 0 | 0 | 1 |
| 5 | 3 | 6 | 0.10 | 12 | 0 | 0 | 1 |
| 6 | 3 | 5 | 0.01 | 4 | 0 | 1 | 1 |
| 7 | 3 | 10 | 0.06 | 0 | 1 | 0 | 0 |
| 8 | 3 | 6 | 0.01 | 0 | 0 | 1 | 0 |
| 9 | 3 | 6 | 0.08 | 10 | 0 | 0 | 1 |
| 10 | 3 | 5 | 0.01 | 4 | 0 | 1 | 1 |
| 11 | 3 | 9 | 0.04 | 0 | 1 | 0 | 0 |
| 12 | 3 | 9 | 0.06 | 0 | 1 | 0 | 1 |
| 13 | 4 | 6 | 0.08 | 5 | 0 | 1 | 1 |
| 14 | 4 | 8 | 0.09 | 2 | 0 | 0 | 1 |
| 15 | 4 | 10 | 0.10 | 0 | 1 | 0 | 0 |
| 16 | 4 | 10 | 0.07 | 0 | 1 | 0 | 0 |
| 17 | 4 | 8 | 0.10 | 3 | 0 | 1 | 0 |
| 18 | 4 | 8 | 0.10 | 3 | 0 | 0 | 1 |
| 19 | 4 | 10 | 0.09 | 0 | 1 | 0 | 0 |
| 20 | 4 | 9 | 0.03 | 0 | 1 | 1 | 1 |
| 21 | 4 | 7 | 0.08 | 3 | 0 | 1 | 1 |
| 22 | 4 | 9 | 0.11 | 1 | 1 | 0 | 1 |
| 23 | 4 | 10 | 0.07 | 0 | 1 | 0 | 1 |
| 24 | 4 | 9 | 0.05 | 0 | 1 | 0 | 1 |
| 25 | 5 | 10 | 0.10 | 1 | 0 | 0 | 1 |
| 26 | 5 | 9 | 0.10 | 1 | 0 | 1 | 1 |
| 27 | 5 | 10 | 0.13 | 0 | 0 | 0 | 1 |
| 28 | 5 | 10 | 0.11 | 0 | 0 | 0 | 1 |
| 29 | 5 | 11 | 0.17 | 1 | 0 | 0 | 1 |
| 30 | 5 | 9 | 0.04 | 0 | 0 | 1 | 0 |
| 31 | 5 | 10 | 0.14 | 1 | 0 | 1 | 0 |
| 32 | 5 | 11 | 0.10 | 0 | 1 | 0 | 1 |
| 33 | 5 | 10 | 0.12 | 1 | 0 | 0 | 1 |
| 34 | 5 | 11 | 0.12 | 1 | 1 | 0 | 1 |
| 35 | 5 | 11 | 0.16 | 0 | 0 | 0 | 0 |
| 36 | 5 | 10 | 0.10 | 0 | 0 | 0 | 1 |

APPENDIX E

HUMAN SUBJECTS DOCUMENTATION

To: Maria Restrepo
COOR
CC: Gareth Morgan
From: Mark Roosa, Chair 
Soc Beh IRB 
Date: 12/10/2010
Committee Action: Renewal
Renewal Date: 12/10/2010
Review Type: Expedited F7
IRB Protocol #: 0701001450
Study Title: Spanish Screener for Language Impairment in Children (SSLIC)
Expiration Date: 12/09/2011

The above-referenced protocol was given renewed approval following Expedited Review by the Institutional Review Board.

It is the Principal Investigator's responsibility to obtain review and continued approval of ongoing research before the expiration noted above. Please allow sufficient time for reapproval. Research activity of any sort may not continue beyond the expiration date without committee approval. Failure to receive approval for continuation before the expiration date will result in the automatic suspension of the approval of this protocol on the expiration date. Information collected following suspension is unapproved research and cannot be reported or published as research data. If you do not wish continued approval, please notify the Committee of the study termination.

This approval by the Soc Beh IRB does not replace or supersede any departmental or oversight committee review that may be required by institutional policy.

Adverse Reactions: If any untoward incidents or severe reactions should develop as a result of this study, you are required to notify the Soc Beh IRB immediately. If necessary a member of the IRB will be assigned to look into the matter. If the problem is serious, approval may be withdrawn pending IRB review.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, or the investigators, please communicate your requested changes to the Soc Beh IRB. The new procedure is not to be initiated until the IRB approval has been given.