The International Journal of Educational and Psychological Assessment
August 2011, Vol. 8(1)

Using the Rasch Model to Determine Form Equivalence in the Trilingual Lollipop Readiness Test

W. Steve Lang
University of South Florida St. Petersburg, U.S.A.

Alex L. Chew
Georgia Southern University, U.S.A.

Carol Crownover
Seminole (FL) County Public Schools, U.S.A.

Judy R. Wilkerson
Florida Gulf Coast University, U.S.A.

Abstract

Determining the cross-cultural equivalence of multilingual tests is a challenge that is more complex than simple horizontal equating of test forms. This study examines the functioning of a trilingual test of preschool readiness to determine the equivalence. Different forms of the test have previously been examined using classical statistical techniques on bilingual forms (Lang, Chew, & Schomber, 1992). A family of psychometric models that maximize information from the data while independent of incidental circumstances is the Rasch model (Wright & Mok, 2004). The Rasch model met the requirements of analysis appropriate with different aggregations of items, handling of missing data, and sensitivity to misfit from the construct. Graphic and parametric statistical analyses resulted from the Rasch analysis that identified test items that functioned differently by language. Even though conclusions based on this sample size and the limitation of analysis with no common people or items would require further investigation after some test corrections, this exercise was revealing to the authors and convincing that the French version of the *Lollipop Test* was not performing as equivalent compared to the Spanish and English versions. This was completely unknown prior to this study.

Keywords: Rasch Model, preschool readiness, multilingual testing

Introduction

Multilingualism has been a problem in psychoeducational assessment for almost a century (Arsenian, 1937). A number of studies have described the difficulties of using standardized instruments developed for English language instruments with limited-English-proficiency (LEP) or bilingual children (Figueroa, 1983, Valencia & Rankin, 1985). The Lollipop Test: A Diagnostic Screening Test of School Readiness-III (Chew, 2007) is a criterion-referenced test with four developmental subtests designed for diagnostic assessment of pre-school and kindergarten children. Translations of the Lollipop Test of School Readiness into Spanish (Chew, 1989), Portuguese (C. Martins, personal communication, September 28, 2010), and French (Venet, Normandeau, Letarte, & Bigras 2003) from the original English (Chew, 1981) may be linguistically sound, but the

question of cultural construct equivalence is also important to valid, reliable, and fair assessment. Mardell-Czudnowski (1987) recommends that American norm tests should be interpreted with extreme caution when used with multilingual children.

The need for multilingual editions of a culturally-fair school readiness test was the motivation for the development of Lollipop Test in Spanish and French translations. In at least one study with the Lollipop test, equal total scores did not necessarily translate into construct equivalence for bilingual students (Vargas & Lang, 1994). With four or five year old children who sometimes mix two languages in a multilingual cultural, a developmental readiness test needs to work at the item level and be as culturally-fair as possible for valid readiness measurement.

Background of the Readiness Construct and Lollipop Test

The concept of readiness can be related to achievement in various school subjects and at various levels of the curriculum. "Traditionally, however, readiness refers to the intellectual and social characteristics of kindergarten and first grade children. Readiness may be defined as a general responsiveness to instruction or as specific intellectual abilities that are predictive of the development of reading or of arithmetic skills" (Leton & Rutter, 1973, p. 293). Readiness as defined by Ausubel (1959, p. 247) refers to "the adequacy of existing capacity in relation to the demands of a given learning task." Generally, readiness refers to the capacity for meeting successfully certain expectancies or for achieving particular levels of performance (Brandt, 1971). Even if we think of readiness only as capacity, Brandt suggests that at least three different, though interrelated, kinds of readiness stand out as especially important. One is physical maturity, which has long been recognized as a precursor to jumping, skipping, bicycling and other gross motor activities of early childhood. Physical maturity is also related to various intellectual activities and school accomplishments.

The second kind of readiness is socio-economical, having to do with developing personality qualities and the strides already taken toward independence. Kindergarteners who are relatively outgoing and self-assured are apt to find the school world exciting and enjoyable. The dependent, withdrawn child, on the other hand, is at a different stage of development and initially needs the warm assurance and helping hand which good schools offer before other horizons can be fully developed. The third kind of readiness discussed by Brandt might be labeled intellectual-educational, referring to the fact that during the first years of life children have already had many experiences that shape their thinking and fill their minds. By age 6 their vocabulary averages well over 2,500 words, although culturally deprived children may know only half of this number. In the *Encyclopedia of Educational Research (4th),* Tyler (1969) offers an in depth, concise review of the literature on the concept of readiness. In a similar earlier paper, Tyler (1964) presents a discussion of issues related to readiness to learn. The idea of readiness has been and still is inherent in our thinking about education, although Tyler's review suggests that there is considerable diversity of opinion about both practice and theory. Different theorists have different notions about the nature of

child development. Some like Gesel consider that it is a matter of maturation forces, experiences, formal and informal teaching, and equilibration i.e., the organization of knowledge occasioned by the appearance of contradictions in the child's system of beliefs. Still others, e.g., learning psychologists, regard it as a matter of accumulating new behavior through learning from experiences. More recent discussions still debate the exact components and theories of readiness, but there remains a general consensus that readiness assessment is common and useful for preschool and readiness programs (Kagan, 1990; Gredler, 1997).

Purpose of this Study

Although multiple studies utilizing classical methods have been conducted on the *Lollipop Test* and each individual language translation of it, no studies prior to this one have equated the test results for children who are native speakers of the three different languages tested; English, French, and Spanish. The purpose of this study was to determine the equivalence of the three versions of the test by exploring item functioning employing the Rasch model (1980). This study was the first Differential Item Functioning (DIF) analysis of the *Lollipop Test* and the first trilingual scaling of the items.

Method

Participants

The initial data analysis included 597 children tested, including 191 English speaking, 234 French speaking, and 112 Spanish speaking children. Of the 597 children, 529 children were tested twice (pre-post). Sixty-eight children were available for one, but not both testing opportunities. Many programs test twice to provide evidence of effectiveness for various funding requirements. Some of the sample, particularly the migrant population, was only available for one testing session. This reduced the Spanish speaking group most often. Of the 112 Spanish children, 70 were post tested while 42 left during the year. A small number of subjects that varied with each process were occasionally dropped from exploratory analyses as extreme or misfitting persons.

The children in the sample were racially and ethnically diverse with 110 black, 115 Hispanic (representing South American, European, and island nations), 21 Eastern (representing Asian and Pacific Island nations), and 350 white (representing North American, African, European, and South American nations) children identified by ethic origin. The sample was composed of 310 female and 284 male subjects.

Instrument

Extensive details of the validity and reliability of *The Lollipop Test* can be found in *The Developmental and Interpretive Manual for the Lollipop Test* (Chew, 2007) that summarizes twenty-eight research studies. The *Lollipop Test* is

an individually administered test of school readiness that requires approximately 15 minutes for administration and scoring. Reliability has been reported as high as KR20 = .93 (Chew, 1981). Predictive and concurrent validity studies have been conducted (Chew & Morris, 1984; Chew & Morris, 1989; Chew & Lang, 1990; Eno & Woehike, 1995).

The *Lollipop Test* consists of three components and 52 items: (1) a set of seven stimulus cards used for presentation of items throughout the test, (2) the administration and scoring booklet, and (3) the *Developmental and Interpretive Manual*. The test contains four sections or subtests: Test 1: Identification of colors and shapes, and copying shapes; Test 2: Picture description, position, and spatial recognition; Test 3: Identification of numbers, and counting; and Test 4: Identification of letters, and writing. Through several revisions, the *Lollipop Test* has been popular and demonstrated validity and reliability in the English edition with classical item analysis, longitudinal predictions, and comparative correlations with other instruments.

Procedure

The sample was collected in North America from multiple locations and all subjects were enrolled in pre-kindergarten or kindergarten programs. Many children represented cultures new to the United States or Canada and therefore the spoken language at home was French or Spanish. All schools used English as the language of instruction. Children were typically tested at program entrance and exit. The children resided in Florida, Georgia, Michigan, and Quebec (Canada). The ages of the sample ranged from 48 to 72 months (4 to 6 years old). Specific locations and schools are not identified here, but a variety of urban to rural populations, levels of SES, and preschool programming models was included. The sample generalizability would not be randomly drawn from or specific to a defined population, but the functioning of different test forms within this heterogeneous sample was considered the focus of the investigation.

Since preschool programming varied widely as to curriculum, teacher quality, required age for entrance, and contact hours; generalizable sample gains were considered globally as a normal consequence of the kindergarten experience, but without specific consideration of differential program impacts. Many children were selected as qualified for targeted remedial programs such as Head Start, but enrollment was controlled by the local school requirements.

Analyses

Modern test development methods now include vertical and horizontal equating. The specific methodology for equating that incorporates the Rasch model is well established (Wolfe, 2004), but only provides partial application to this effort since there are no common people to calibrate and items translation equivalence is the issue. The Rasch model was deemed appropriate for this effort because the construct of the instrument was considered established, but the need for sample independent analysis was evident (Wright & Mok, 2004). The Rasch

model also met the requirements of analysis appropriate with different aggregations of items, handling of missing data, and sensitivity to misfit from the construct.

This study used an exploratory analysis that started with overall model statistics and followed up as patterns were revealed, because a strict equating model utilizing common items or persons is impossible. Perhaps a better description of the problem would be to compare different forms of a test with no common persons and parallel, but not common items. The Rasch model was expected to be useful for overall consideration of dimensionality, reliability, separation and fit. Applying the Rasch model started with calibration of items, and examined the overall estimates of the model parameters (Smith, 2001). At least one demonstration of item calibration on a school readiness test has been previously demonstrated (Gumpel, 1999). Additionally, a recent analysis a cross-cultural comparison using the Rasch model has been effectively demonstrated (Choi, Mericle, & Harachi, 2006). Those efforts guided the study reported here, even though the planned analysis was admittedly investigative.

A statistic produced by the Rasch model is Differential Item Functioning (DIF), but DIF analysis requires both an appropriately modeled test and detailed knowledge of the construct (Bond & Fox, 2001). DIF was examined for each language form of the *Lollipop Test* separately and as a pooled set of items. DIF analyses by the four subtests were performed. Analyses by pretest scores, posttest scores, and combined pretest-posttest were performed. Particular attention was focused on examining patterns, item functioning, and contrasts that could be interpreted as culturally loaded or indicating possible language bias. Analysis results were interpreted within the framework of language expertise and construct knowledge. The choice of separate analysis of pretest, posttest, or both sets of items is dependent on the frame of reference that is meaningful to the user's assertions (Wright, 1996). The comparative language of testing was the primary variable of interest while the most common use of the test is for pretest-posttest gains utilizing the four subtests, therefore that was the focus of the researchers.

Results

Basic Results

The first observation is that the Lollipop Test meets Rasch dimensional requirements for continued analysis and calibration. All analyses (separate and pooled) were examined with both Winsteps (Linacre, 2004) Diagnosis/Separation tables and RUMM2020 (Andrich, 2004a) Summary Statistics tables. Since RUMM2020 and Winsteps use slightly different estimation algorithms, the results in side-by-side comparisons can be confusing. For consistency, Winsteps output is reported except in the cases where RUMM2020 reports or graphics are indicated in this report. A sample summary result is included in Table 1 below, produced with Winsteps software. The real person separation of .91 indicates that the scale discriminates between the persons well. The real item separation of 1.0 indicates that the items create a well defined variable. An outfit mean of .1 (expected value = 0) and S. D. of 1.3 (expected value = 1.0) suggest that the data fit the model. Our

separation table provides evidence that a Rasch analysis is reasonable (Smith, 1999). In all cases, runs met the guidelines suggested for Winsteps (Smith, 1999) or were rated Excellent on RUMM2020's Power of Test-of-Fit. Extreme items and persons (misfitting) were typically dropped from some later sub analyses.

Table 1
Lollipop Test Rasch Analysis of Three Languages

INPUT: 597 persons, 104 items MEASURED: 597 persons, 104 items, 84 CATS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	67.8	89.0	5.00	.45	1.30	.9	1.21	.1
S.D.	29.3	23.5	1.98	.16	.48	1.4	1.01	1.3
MAX.	135.0	104.0	13.14	1.85	3.54	5.2	9.90	5.5
MIN.	11.0	50.0	-1.63	.34	.47	-2.5	.04	-2.3

REAL RMSE .58 ADJ.SD 1.89 SEPARATION 3.24 person RELIABILITY .91
MODEL RMSE .48 ADJ.SD 1.92 SEPARATION 4.01 person RELIABILITY .94
S.E. OF person MEAN = .08

VALID RESPONSES: 85.6% person RAW SCORE-TO-MEASURE CORRELATION = .68
CRONBACH ALPHA (KR-20) person RAW SCORE RELIABILITY = .96

SUMMARY OF 104 MEASURED items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	389.1	511.0	12.92	.23	.98	-1.0	1.22	-1.0
S.D.	259.5	16.7	5.27	.12	.30	4.6	1.14	4.1
MAX.	1657.0	529.0	20.62	.85	2.16	9.9	6.58	9.9
MIN.	78.0	492.0	1.79	.07	.51	-9.0	.21	-6.8

REAL RMSE .26 ADJ.SD 5.26 SEPARATION 20.43 item RELIABILITY 1.00
MODEL RMSE .25 ADJ.SD 5.26 SEPARATION 20.64 item RELIABILITY 1.00
S.E. OF item MEAN = .52

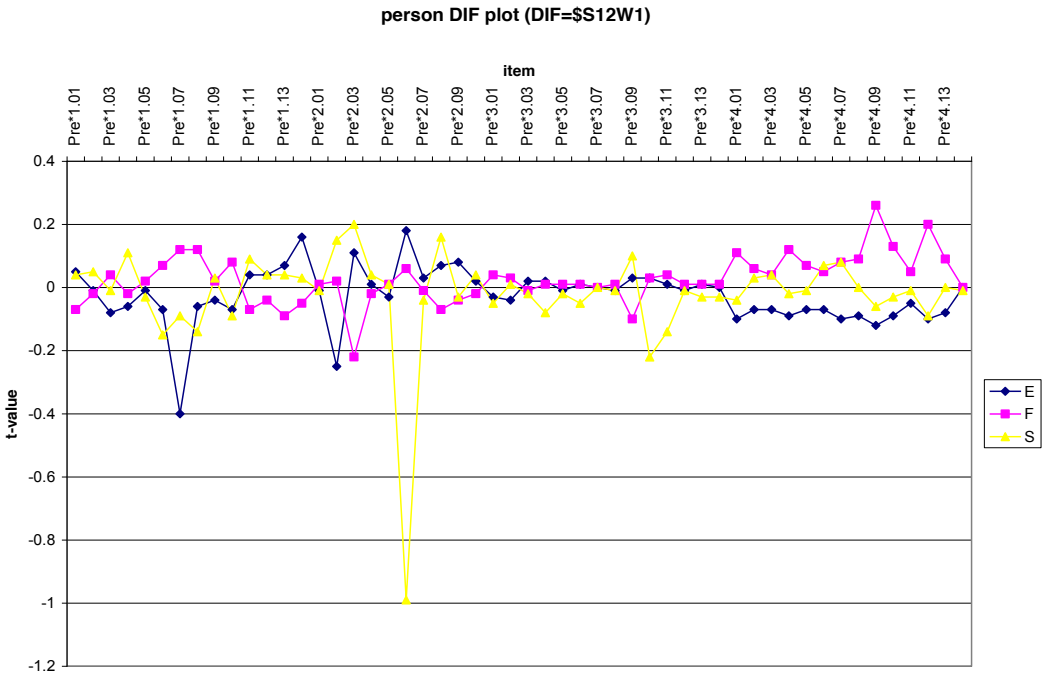
UMEAN=5.000 USCALE=2.000 item RAW SCORE-TO-MEASURE CORRELATION = -.81

Differential Item Analyses

Smith (2004) described the detection of item bias with the t-statistic. In a simulation, Smith suggests parameters bias detection using reference groups and focal groups of different sizes and varying the proportion of biased items. Smith concludes, "The Rasch item bias statistics lack the power to detect bias of less than .5 logits unless there are 500 people in each subpopulation....The use of preset item bias cutoffs like +2 and -2 to detect items that favor one subpopulation over another in pairwise comparisons can be misleading." (p. 415-416). Given the complexity of a three-way comparison plus the differing degrees of freedom in each comparison, a graphic representation is provided here in Figure 1 (pre-test items) and Figure 2 (post-test data) for exploratory analysis by Winsteps / Excel.

On first viewing the pre-test calibration, item 1.07 appears to be unusually easy for English language students. Item 1.07 item is a performance item, "Show

me the circle." Item 2.06 appears as unusually easy for Spanish language students. That item refers to a picture of a cat and kittens and asks, "Show me which is the biggest." On the post-test calibration, a number of items asking for color recognition and counting appeared to be easier for Spanish speaking students. On both the pre-test and post-test, Identification of Letters and Writing (subtest four), appeared more difficult for French language students. Keeping in mind the power and Type I error rate issues raised by Smith (2004), a table showing the visually suspect items narrowed to those with statistically significant differences ($p < .05$) and logit contrasts greater than 1.0 were identified.

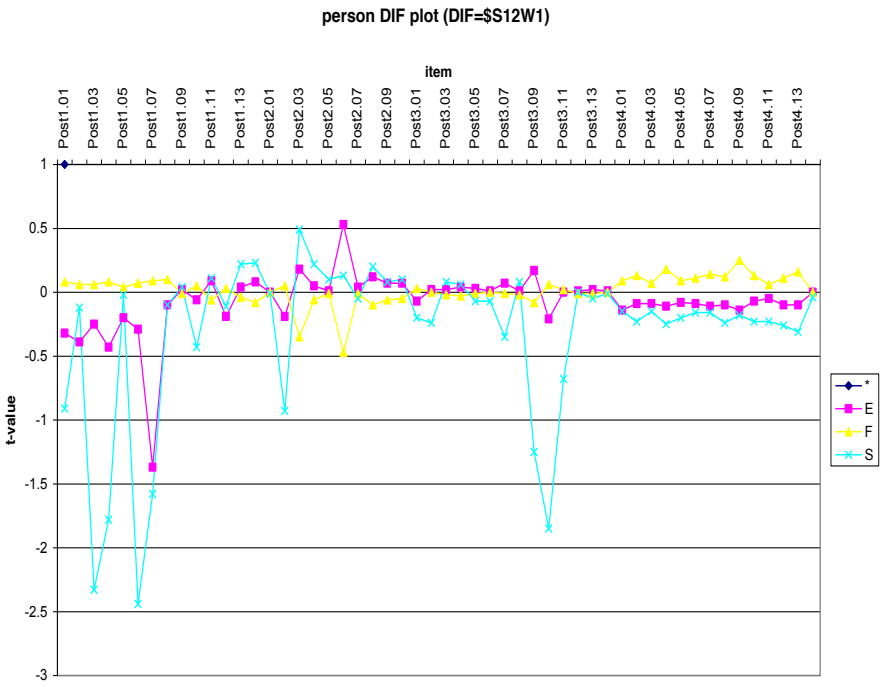


E=English, F=French, S=Spanish
Subtest and Item Number label items. For example Pre-test 3.04= Subtest 3, Item #4. Only the odd number items are labeled on the graph.

Figure 1. Plot of Pretest Items Comparing English, French, and Spanish

Language Version Analysis

In the plots the DIF measure is the difficulty of the item for each group with all else held constant plotted on overlapping lines where the t-value gives the significance of the unit normal deviate (Linacre, 2003). Graphically, the French version appears to differ on subtest four (to the right) on the plot.



E=English, F=French, S=Spanish Items are labeled by Subtest and Item Number. For example Pre-test 3.04 = Subtest 3, Item #4. Only the odd number items are labeled on the graph.

Figure 2. Plot of Posttest Items Comparing English, French, & Spanish

Figure 2 demonstrates that some items appear easy on subtest one. That is somewhat expected as a posttest after instruction and a year's growth. Again, the French test appears to separate on subtest one and subtest four on the plot.

One item that graphically appears to be extremely different (pre-test 2.06) actually is a case of a very easy item that is even easier for the Spanish subgroup. With the worst point biserial correlation on the test (r_{bis}=.14) and a difficulty of -3.45, the graph is misleading as the contrast is large, but not significant. Some items in subtests one and two have some variability without a clear pattern explaining the variance other than error. A few post-test items seem to have become very easy for Spanish speaking students, most likely as a result of targeted instruction. On the pre-test, the English students appear to perform best across subtest four, while on the post-test the Spanish students are favored. Possible interpretations of the results are addressed in the discussion section of this article, but the plots indicate enough deviation by language contrast to suggest additional inspection of contrasts is warranted.

The International Journal of Educational and Psychological Assessment
August 2011, Vol. 8(1)

Table 2

Items with significant DIF (p < .05) & a contrast > 1 logit

CLASS	Person MEASURE	DIF S.E.	DIF CLASS	Person MEASURE	DIF S.E.	DIF CONTRAST	DIF S.E.	JOINT t	d.f.	Prob.	item Number
E	1.71	.17	F	2.77	.17	-1.05	.24	-4.37	414	.0000	15 Pre*1.08
F	2.77	.17	S	1.39	.21	1.37	.27	5.02	339	.0000	15 Pre*1.08
E	-4.22	.10	F	-5.78	.11	1.56	.14	10.81	414	.0000	25 Pre*1.13
F	-5.78	.11	S	-4.62	.12	-1.16	.16	-7.17	339	.0000	25 Pre*1.13
E	-2.90	.15	F	-4.50	.08	1.60	.17	9.43	414	.0000	27 Pre*1.14
F	-.57	.18	S	1.57	.21	-2.15	.28	-7.60	339	.0000	33 Pre*2.03
F	1.42	.14	S	2.57	.25	-1.15	.29	-3.98	339	.0001	43 Pre*2.08
E	1.39	.16	F	2.65	.17	-1.26	.24	-5.36	412	.0000	77 Pre*4.01
E	2.37	.18	F	3.46	.21	-1.09	.28	-3.88	412	.0001	83 Pre*4.04
E	2.65	.19	F	4.21	.28	-1.56	.34	-4.63	412	.0000	93 Pre*4.09
F	4.21	.28	S	3.06	.28	1.15	.40	2.91	339	.0039	93 Pre*4.09
E	2.39	.18	F	3.50	.22	-1.12	.28	-3.96	413	.0001	95 Pre*4.10
E	2.42	.18	F	3.81	.24	-1.39	.30	-4.62	413	.0000	99 Pre*4.12
F	3.81	.24	S	2.63	.25	1.18	.35	3.38	339	.0008	99 Pre*4.12
E	2.08	.17	F	3.25	.16	-1.17	.24	-4.96	423	.0000	16 Post1.08
F	2.00	.14	S	.63	.42	1.37	.44	3.08	301	.0023	20 Post1.10
E	-5.50	.20	F	-4.20	.08	-1.30	.21	-6.15	423	.0000	24 Post1.12
F	-5.70	.13	S	-4.26	.21	-1.43	.24	-5.94	301	.0000	26 Post1.13
E	-3.29	.10	F	-4.99	.09	1.70	.14	12.55	421	.0000	28 Post1.14
F	-4.99	.09	S	-2.64	.16	-2.35	.18	-12.8	300	.0000	28 Post1.14
E	2.31	.17	F	-.38	.22	2.69	.28	9.77	423	.0000	34 Post2.03
F	-.38	.22	S	3.05	.27	-3.43	.35	-9.91	301	.0000	34 Post2.03
E	-2.87	.10	F	-4.09	.08	1.22	.13	9.57	423	.0000	36 Post2.04
F	-4.09	.08	S	-2.11	.17	-1.97	.19	-10.4	302	.0000	36 Post2.04
E	2.25	.17	F	.86	.16	1.39	.23	6.03	423	.0000	44 Post2.08
F	.86	.16	S	2.21	.29	-1.35	.33	-4.12	301	.0001	44 Post2.08
E	-3.77	.10	F	-5.04	.09	1.27	.14	9.11	423	.0000	48 Post2.10
F	-5.04	.09	S	-3.97	.19	-1.07	.21	-5.15	301	.0000	48 Post2.10
E	.61	.23	F	-.48	.23	1.09	.33	3.35	422	.0009	66 Post3.09
E	.61	.23	S	-1.44	.93	2.05	.96	2.14	257	.0336	66 Post3.09
E	1.40	.19	F	2.72	.15	-1.32	.24	-5.46	422	.0000	78 Post4.01
F	2.72	.15	S	1.68	.32	1.04	.35	2.96	300	.0033	78 Post4.01
E	2.28	.17	F	3.58	.17	-1.30	.24	-5.39	423	.0000	80 Post4.02
F	3.58	.17	S	2.06	.30	1.52	.34	4.42	300	.0000	80 Post4.02
E	2.27	.17	F	3.87	.19	-1.60	.25	-6.36	422	.0000	84 Post4.04
F	3.87	.19	S	2.06	.30	1.81	.35	5.16	300	.0000	84 Post4.04
E	2.17	.17	F	3.22	.16	-1.05	.23	-4.52	423	.0000	86 Post4.05
F	3.22	.16	S	1.97	.30	1.25	.34	3.66	300	.0003	86 Post4.05
E	2.24	.17	F	3.46	.17	-1.22	.24	-5.11	422	.0000	88 Post4.06
F	3.46	.17	S	2.23	.29	1.23	.34	3.67	300	.0003	88 Post4.06
E	2.16	.17	F	3.58	.17	-1.43	.24	-5.86	422	.0000	90 Post4.07
F	3.58	.17	S	2.23	.29	1.35	.34	4.00	300	.0001	90 Post4.07
E	2.10	.17	F	3.35	.17	-1.25	.24	-5.27	422	.0000	92 Post4.08
F	3.35	.17	S	1.88	.31	1.47	.35	4.23	300	.0000	92 Post4.08
E	2.18	.17	F	4.23	.21	-2.04	.27	-7.59	422	.0000	94 Post4.09
F	4.23	.21	S	2.40	.28	1.83	.35	5.20	300	.0000	94 Post4.09
E	2.60	.16	F	3.74	.18	-1.14	.24	-4.65	422	.0000	96 Post4.10
F	3.74	.18	S	2.23	.29	1.51	.34	4.41	300	.0000	96 Post4.10
F	2.66	.15	S	1.58	.33	1.08	.36	3.01	300	.0028	98 Post4.11
E	1.93	.17	F	3.17	.16	-1.24	.24	-5.25	423	.0000	100 Post4.12
F	3.17	.16	S	1.68	.32	1.48	.36	4.17	300	.0000	100 Post4.12
E	2.19	.17	F	3.67	.18	-1.48	.25	-6.03	423	.0000	102 Post4.13
F	3.67	.18	S	1.79	.31	1.88	.36	5.22	299	.0000	102 Post4.13

E=English, F=French, S=Spanish. Contrasting items (p < .05) with a magnitude greater than 1 logit.

Grey indicates the item contrasted on both the pretest and the posttest. Yellow is the only non-French contrast identified.

Table 2 contains the DIF analysis by Winsteps for the Pretest and Posttest items identified as significant ($p < .05$) contrasts by language with a magnitude greater than one logit. The grey bars indicate items which show the parallel contrast if the item appeared on both the Pretest and Posttest. Only one item appears to show an English-Spanish contrast. With 103 contrasts, it is possible that some appear significant due to type I error. Table 3 illustrates the Winsteps DIF analysis by subtests for both pretest and posttest analyses where the contrast was significant and the magnitude of difference was greater than .25 logits. Again, all the subtest contrasts that met these admittedly arbitrary criteria contained the French language test.

Table 3
Winsteps output of subtest scores (1-4) with significant DIF ($p < .05$); a contrast > .25 logits

596 persons, 104 items MEASURED: 529 persons, 52 items, 45 CATS
CLASS-LEVEL BIAS/INTERACTIONS FOR PRETEST

Person CLASS	DIF MEASURE	DIF S.E.	Person CLASS	DIF MEASURE	DIF S.E.	DIF CONTRAST	JOINT S.E.	t	d.f.	Item Prob.	
E	.21	.04	F	-.21	.04	.41	.05	-7.60	INF	.0000	1
F	-.21	.04	S	.09	.05	-.30	.06	4.67	INF	.0000	1
F	-.12	.03	S	.16	.05	-.28	.06	4.54	INF	.0000	2
F	.09	.03	S	-.18	.05	.27	.06	-4.50	INF	.0000	3
E	-.27	.04	F	.26	.04	-.52	.05	9.54	INF	.0000	4
F	.26	.04	S	-.07	.06	.32	.07	-4.75	INF	.0000	4

CLASS-LEVEL BIAS/INTERACTIONS FOR POSTTEST

Person CLASS	DIF MEASURE	DIF S.E.	Person CLASS	DIF MEASURE	DIF S.E.	DIF CONTRAST	JOINT S.E.	t	d.f.	Item Prob.	
E	.14	.03	F	-.18	.03	.33	.04	-8.13	INF	.0000	1
F	-.18	.03	S	.20	.04	-.38	.05	7.72	INF	.0000	1
E	.18	.03	F	-.22	.02	.40	.04	-11.10	INF	.0000	2
F	-.22	.02	S	.23	.04	-.45	.05	10.00	INF	.0000	2
E	-.31	.03	F	.40	.03	-.71	.04	19.63	INF	.0000	4
F	.40	.03	S	-.28	.04	.68	.05	-14.90	INF	.0000	4

E=English, F=French, S=Spanish. Grey indicates the item contrasted on both the pretest and the posttest.

Sample Individual Item Analyses

Finally, an investigation of a subset of items in order to observe item and subtest functioning graphically with expected Item Characteristic Curves (ICC's) was performed. RUMM2020 was chosen for this analysis. The results here include examples that illustrate the process as space does not allow comprehensive output. In Figures 3 and 4 are the plots of the observed ICC confidence interval means from Posttest item 2.07 that appears to model an expected Rasch ICC reasonably well. This is followed by a contrast plot of the language functioning where there was no significant difference detected by the contrasts.

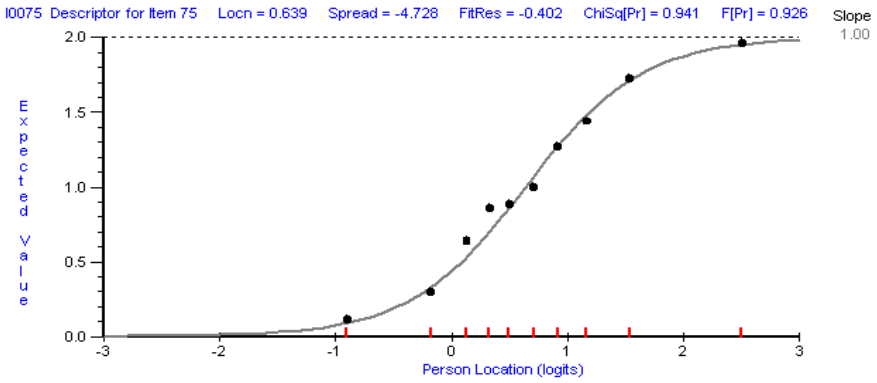


Figure 3. Function plot of item 14 from subtest 3 as a pretest

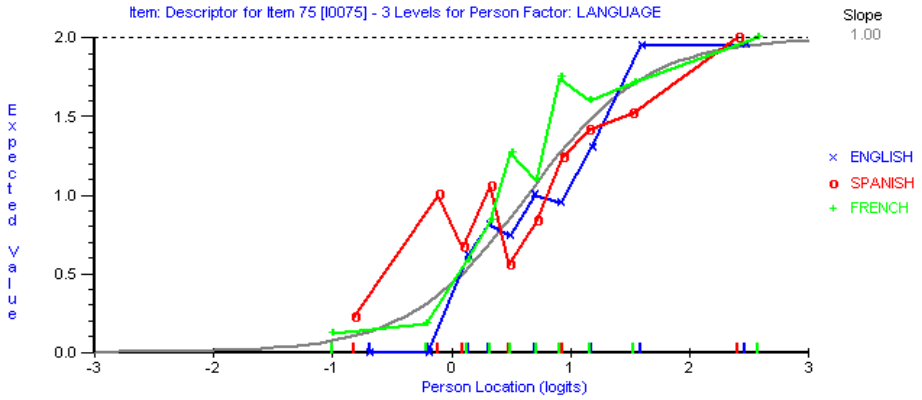


Figure 4. Language contrast plot of item 14 from subtest 3 as a pretest

Now we can observe an item that does not appear to behave well as an expected Rasch ICC. In Figure 5 is a plot of item 3 from subtest 2 (2.03 in Figure 4) whose expected CI means do not appear to have appropriate discrimination. In Figure 6, the same item is plotted showing the language DIF contrasts. This pretest item was originally identified statistically and is now confirmed graphically.

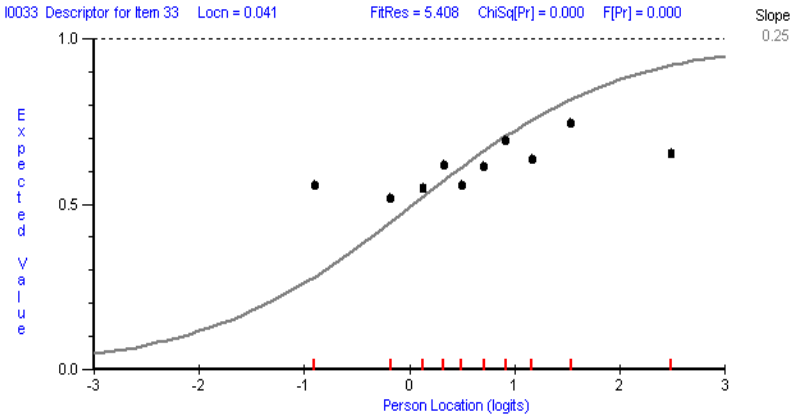


Figure 5. Observed plot of item 3 from subtest 2 as a pretest

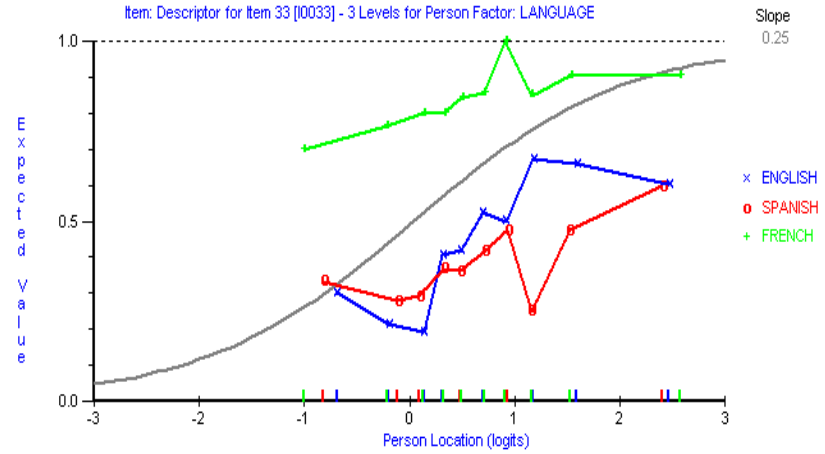


Figure 6. Language contrast plot of item 3 from subtest 2 as a pretest

It was also confirmed that the time of testing may be important to the item functioning on this preschool readiness test. Originally, this was a logical assertion of the authors, but in Figures 7 and 8 are plots of the pretest and posttest functions for item (1.12). The DIF is still present, but now the difficulty of the bias has reversed. This item met our statistical criteria on the posttest analysis in but did not have a large enough contrast to meet the criteria for the pretest analysis. Graphically, this is clearly an item of interest.

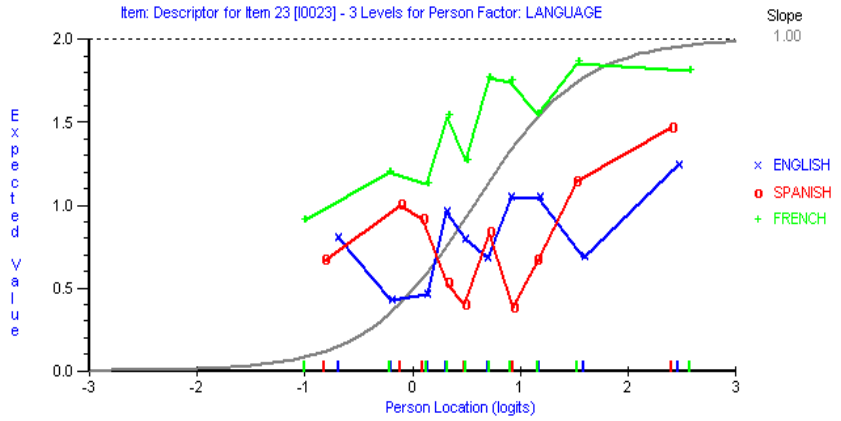


Figure 7. Language contrast plot of item 12 from subtest 1 as a pretest

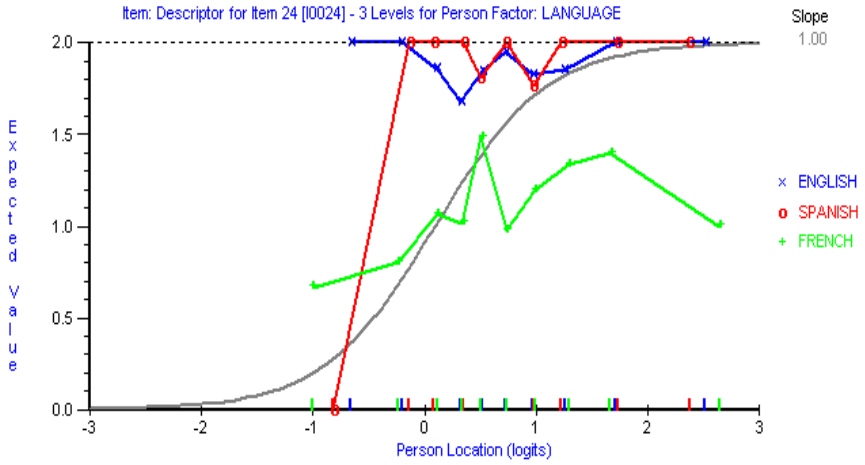


Figure 8. Language contrast plot of item 12 from subtest 1 as a posttest

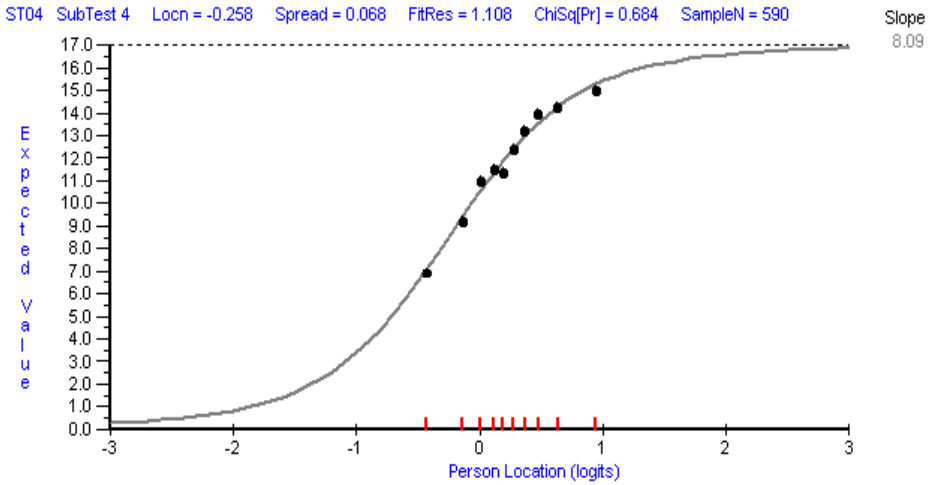
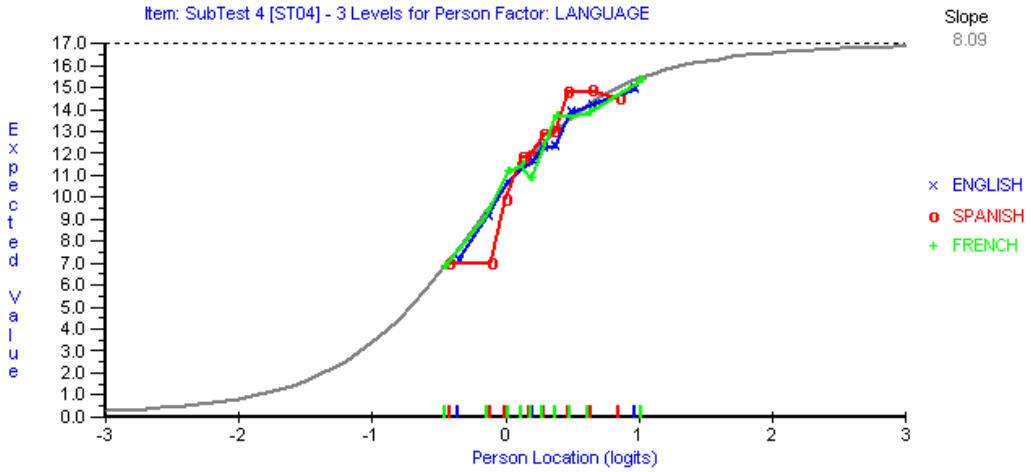


Figure 9. ICC and language contrast plots of subtest 2 administered as a posttest



Finally, analysis of the subtest scores was performed. It is common to report the *Lollipop Test* by subtest scores as the design of the test separates items into subtests that reflect item types and increasing difficulty as part of its original design. Similar to the individual item graphs, Figures 9 and 10 demonstrate the observed ICC of subtest two given as a posttest, and subtest four given as a posttest. These reveal performance differences contrasting by language. This is consistent with our original Figures 2 and 3. It is also consistent with the larger number of individual items expressing DIF in subtest four as opposed to subtest two.

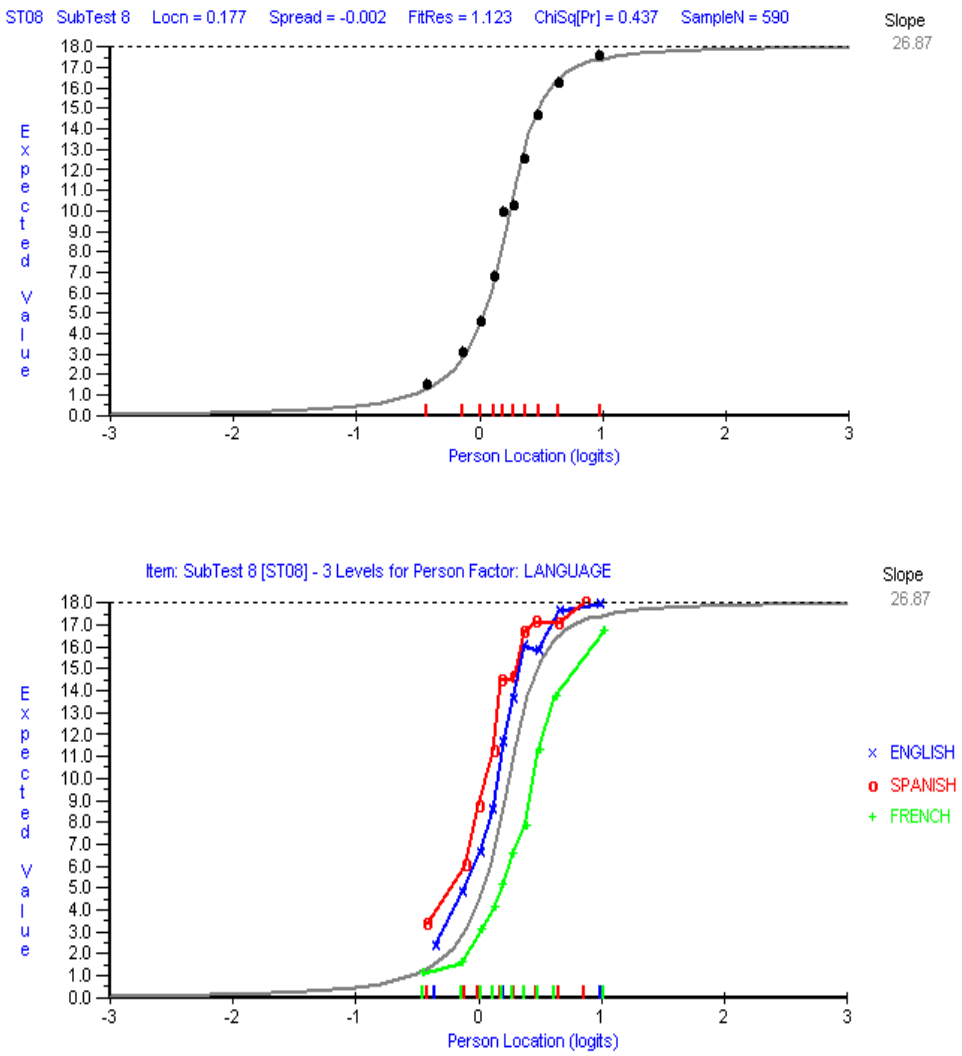


Figure 10. ICC and language contrast plots of subtest 4 as a posttest

Discussion

For purposes of summary of the graphic and statistical analysis, some items that appeared to demonstrate the most DIF across language form and testing times (pre-post) were examined in the results section in detail. The items identified as functioning differently across languages on the *Lollipop Test* are reproduced in Table 4 below.

Table 4
Items on the Trilingual Lollipop that demonstrated Differential Item Functioning

Subtest	*Item Number*	*Item*	*Language Contrast*	
			Pre-test	*Post-test*
1: Colors & Shapes	7	"Show me the circle."	E-F, E-S, F-S	E-F, F-S
1: Colors & Shapes	8	"Show me the rectangle."	E-F, F-S	E-F
1: Colors & Shapes	13	"Draw a cross just like this one."	E-F, F-S	F-S
1: Colors & Shapes	14	"Draw a square just like this one."	E-F	E-F, F-S
2: Position & Spatial	3	"Show me (point to) the kitty that is on top."	E-F, F-S	E-F, F-S
2: Position & Spatial	8	"Which one is first?"	F-S	E-F, F-S
4: Letters & Writing	1	"Show me the letter B."	E-F	E-F, F-S
4: Letters & Writing	4	"Show me the letter P."	E-F	E-F, F-S
4: Letters & Writing	9	"What letter is this? (D)"	F-S	E-F, F-S
4: Letters & Writing	10	"What letter is this? (H)"	E-F	E-F, F-S
4: Letters & Writing	12	"Write the letter B."	E-F, F-S	E-F, F-S

E=English, S=Spanish, F=French

Our purpose in this discussion is not to exhaust all items and possible explanations, but to describe how construct knowledge combined with statistical evidence might be used to explain DIF and provide reasonable targets for instrument improvement. We noted that even though it was not the question of interest for this study, an initial DPF (Differential Person Fit) analysis indicated no contrasts larger than .5 logits between ethnic subpopulations or gender. No contrasts by gender or origin were significant for total scores. As such, item differences did not appear to have an interaction effect with person demographics.

First, in both the statistical and graphical analyses, there was a clear trend for the French language version to function as more difficult in subtest four on the posttest which targeted letter identification. The stimulus cards were the same for all languages, and the authors' explanation for the contrast is that European / French handwriting and letter formation is different from American / English. There is a possible cultural difference in handwriting instruction that may confound the subtest four items. For example, the General Teaching Council for England notes that handwriting is taught differently in French than in English (2006). If French posttest children have been taught cursive as opposed to print or D'Nealian letter formation, and may not have the same exposure to uppercase letters, they may simply fail to recognize the capital block letters on the stimulus cards as readily as their Spanish and English speaking counterparts.

The next most interesting DIF concerns both the pretest and posttest items calling for recognition of shapes or drawing of simple shapes (circle, cross, square). Again, the French speaking children appeared to differ from Spanish and English test takers. No apparent curricular or cultural differences were obvious despite

consideration of religion (crosses), playtime activities (games that used shapes such as hop scotch), or television (Sesame Street) that may have provided cultural loading for the items. The best explanation seems to be that the reproduction of the French test contained slightly smaller figures in the vertical direction (estimated at 5 to 10 %) and drawing space. Even though the area is not known to be a consideration and was unnoticed until this investigation of DIF, it may have caused an inadvertent effect on item functioning and this hypotheses needs to be investigated in the future.

Another DIF finding for the French language items was a subset of a stimulus card with a picture of a "mother cat" and "kittens". Questions of spatial relations are asked, "Show me the kitty that is on top." Even though there is no clear reason for the DIF, there is some speculation among the authors that different words for cat and kitty in French (petits chats, chats, chatons) differ in difficulty from English (cat, kitty) or Spanish (gatitos, gatos) simply because the blends and diphthongs in French for "cat" is more complex for young language learners. There is also a theory that the French children are hesitant to point to the picture (pointing is rude). For the purposes of these discussions, these are only educated guesses at this point, but seem like viable construct explanations that warrant inquiry.

The discussion here is a sample of a larger set of interpretations that demonstrate the use of Rasch DIF analysis and construct understanding that would lead to test improvements in the language translations of the *Lollipop Test.* As more multilingual / multicultural tests are used in schools with growing international populations, form equivalence becomes important, especially for placement or high-stakes decisions. Analyses that are sensitive to item differences are difficult within populations where the total score is confounded by many demographic and curricular factors. Simply translating from one language to another may not produce construct measures that are similar enough to be acceptable. Test translation from English to French and English to Spanish had demonstrated concurrent validity and predictive validity in longitudinal studies. (Lang, W. S., Chew, A.L. & Schomber, J., 1992; Venet, M., et. al., 2003), but had not been examined for DIF in comparative settings. One reason that such studies would be unpopular is simply the sample size necessary to detect item bias, and even in this study the number of subjects is adequate, but not as large as desirable. In this case, a conclusion is that the English and Spanish editions of the *Lollipop Test* appear reasonably equivalent, but the French form appears to have bias on items that is significant and warrants item and translation review. Rasch model DIF exposed the contrasts where other research had missed it.

Another conclusion of the authors is that DIF analysis using the Rasch model revealed a number of possible item discrepancies that were hidden in previous classical studies incorporating factor analysis, regression line comparisons, discriminant analysis, and canonical correlations. The Rasch analysis provides a powerful tool for detection of item response to DIF in different language translations of tests. Even though the burden of item writing and construct consideration is always part of test authorship, an analysis that pinpoints potential issues and provides clues to the questions at hand is welcome indeed.

A possible advantage in the use of Rasch software such as RUMM2020 is the split item function so that a bias item can be scored differently for different groups of test takers. Whether a given measurement philosophy or application believes that method to be appropriate, it certainly is another useful tool to be able to equate total scores or subtest scores by simply creating a separate item for the biased group and scoring it differently (Andrich, 2004b, p. 31-39). The Amend Sample Size function in RUMM2020 that allowed a "prophecy formula" for Rasch analysis when large sample sizes are not easily obtainable is also an interesting and useful tool. A recommendation would be to explore if these software functions were able to provide adjusted equating of a limited amount of item or subtest bias and remain true to the construct.

Even though conclusions based on this sample size and the limitation of analysis with no common people or items would require further investigation after some test corrections, this exercise was revealing to the authors and convincing that the French version of the *Lollipop Test* was not performing as equivalent compared to the Spanish and English versions. This was completely unknown prior to this study and gives the authors another challenge for the future. For now we recommend that interpretation of the French version of the Lollipop be considered cautiously, especially with regard to subtest 4. Additionally, this type of analysis demonstrates empirically that tests cannot be simply translated from language to language while assuming that item difficulty and construct validity are the same for all forms. In fact, it takes substantial study to determine if test items function in a similar manner when used in multilingual applications.

References

Andrich, D. (2004a). RUMM2020 Rasch unidimensional measurement models: Calibration software [Computer program]. Perth, WA, Australia: RUMM Laboratory Pty Ltd.

Andrich, D. (2004b). *Interpreting RUMM2020: Part I Dichotomous Data*. Perth, WA, Australia: RUMM Laboratory Pty Ltd.

Arsenian, S.(1937). *Bilingualism and mental development.* New York: Columbia University.

Ausubel, D. P. (1959). Viewpoints from related disciplines: Human growth and development. *Teachers College Record, 60,* 245-254.

Brandt, R. M. (1971). The readiness issue today. In D. Hold & H. Kicklighter (Eds.), *Psychological services in the schools: Reading in preparation, organization and practice.* Dubuque, Iowa: Wm. C. Brown Company, Publishers.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: LEA.

Chew, A. L. (1977). *The design, development, and validation of an individually administered school readiness test.* Unpublished doctoral dissertation, The University of Mississippi.

Chew, A. L. (1981). *The Lollipop Test: A Diagnostic Screening Test of School Readiness.* Atlanta, GA: Humanics Limited.

Chew, A. L. (1983, November). *Validation of an individually administered school readiness test.* Paper presented at the Georgia Educational Research Association Meeting, Athens, GA.

Chew, A. L. (1985, November). *Investigation of the Lollipop Test as a pre-kindergarten screening instrument.* Paper presented at the Georgia Educational Research Association Meeting, Atlanta, GA.

Chew, A. L. (1989a). La Prueba Lollipop: Una prueba diagnostica y seleccionadora de la preparacion para la escuela enmendada [*The Lollipop Test: A diagnostic screening test of school readiness-revised*]. Atlanta, GA: Humanics Limited.

Chew, A. L. (1989b). *The Lollipop Test: A diagnostic screening test of school readiness-revised.* Atlanta, GA: Humanics Limited.

Chew, A. L. (2007). *The Lollipop Test: A diagnostic screening test of school readiness-III.* Ft. Worth, FL: Brumby Holdings

Chew, A. L. (2007). *The developmental and interpretive manual for the Lollipop Test.* Lake Worth, FL: Brumby Holdings

Chew, A. L., Kesler, E. B., & Sudduth, D. H. (1984). A practical example of how to establish local norms. *The Reading Teacher, 38,* 160-163.

Chew, A. L., & Lang, W. S. (1990). Predicting academic achievement in kindergarten and first grade from pre-kindergarten scores on the *Lollipop Test* and DIAL. *Educational and Psychological Measurement, 50,* 431-437.

Chew, A. L., & Lang, W. S. (1992). *Validation of the Spanish edition of the Lollipop Test (La Pruebe Lollipop).* Paper presented at the Eastern Education Research Association Conference, Hilton Head, SC.

Chew, A. L. & Lang, W. S. (1993). Concurrent validation and regression line comparison of the Spanish edition of the *Lollipop Test* (La Prueba Lollipop) on a bilingual population. *Educational and Psychological Measurement, 51,* 173-181.

Chew, A. L., & Morris, J. D. (1984). Validation of the *Lollipop Test*: A Diagnostic Screening Test of School Readiness. *Educational and Psychological Measurement, 44,* 987-991.

Chew, A. L., & Morris, J. D. (1987). Investigation of the *Lollipop Test* as a pre-kindergarten screening instrument. *Educational and Psychological Measurement, 47,* 467-471.

Chew, A. L., & Morris, J. D. (1989). Predicting later academic achievement from kindergarten scores on the Metropolitan Readiness Tests and Lollipop Test. *Educational and Psychological Measurement, 49,* 461-465.

Eno, L., & Woehlke, P. (1995). Use of the *Lollipop Test* as a predictor of California achievement test scores in kindergarten and transitional first-grade status. *Psychological Reports, 76,* 145-146.

Figueroa, R. A. (1983). Test bias and Hispanic children. *The Journal of Special Education, 17*, 431-440.

General Teaching Council for England (2006, July). *Adopting a French approach to teaching handwriting.* Retrieved from http://www.gtce.org.uk/tla/rft/eng_fr0708/eng_fr0708cs/casestudy4/

Gredler, G. R. (1997). Issues in early childhood screening and assessment. *Psychology in the Schools,* 34, 99-113.

Gumpel, T. P. (1999). Use of item response theory to develop a measure of first-grade readiness. *Psychology in the Schools,* 36(4), 285-293.

Kagan, S. L. (1990). Readiness 2000: Rethinking rhetoric and responsibility. *Phi Delta Kappan, 72*, 272-279.

Lang, W. S. (1994). Bilingual testing. *Rasch Measurement; Transactions of the Rasch Measurement SIG of the American Education Research Association, 8*(1), 343.

Lang, W. S., & Chew, A. L. (1988). *The predictive validity of the Lollipop Test.* Paper presented at the Georgia Education Research Association, Atlanta, GA.

Lang, W. S., & Chew, A. L. (1989a). *The Lollipop Test norms-technical manual for Head Start children in Georgia.* Unpublished manuscript, Georgia Southern University.

Lang, W. S., & Chew, A. L. (1989b). *Developing norms for the Head Start children of Georgia.* Paper presented at the annual meeting of the Georgia Education Research Association, Atlanta, GA.

Lang, W. S., & Chew, A. L. (1990). *Using the Lollipop norms to answer your educational questions.* Paper presented at the meeting of the Georgia Head Start Coordinators, Macon, GA.

Lang, W. A., & Chew, A. L. (1997). *Improving assessment to improve retention benefits in kindergarten.* Paper presented at the annual meeting of the Eastern Education Research Association, Hilton Head, SC.

Lang, W. S., Chew, A. L., & Gill, C. C. (1995, March). *Predicting academic success of promoted and retained kindergarten students from multiple variables.* Paper presentation to the Eastern Education Research Association.

Lang, W. S., Chew, A. L., & Schomber, J. (1990). *An initial study of the Spanish edition of the Lollipop Test.* Paper presented at the annual conference of the Georgia Education Research Association, Atlanta.

Lang, W. S., Chew, A. L., & Schomber, J. (1992). The comparative use of the Spanish and English editions of the *Lollipop Test* : a cross cultural study. *Journal of Research in Education, 2*(1), 23-30 .

Lang, W. S., Chew, A. L., & Schomber, J. (1992). *The comparative use of the Spanish and English editions of the Lollipop Test : a cross cultural study.* Paper presented at the Eastern Education Research Association Conference. Hilton Head, SC.

Lang, W. S., Chew, A. L., & Vargas, C. (1993). *Rasch model applications to determine the equivalence of a readiness test in two languages.* Paper presented at the Seventh International Objective Measurement Workshop. Atlanta, GA.

Lang, W. S., Chew, A. L., & Vargas, C. (1995) Rasch model applications to determine the equivalence of a readiness test in two languages. *Resources in Education*, (ERIC Document Reproduction Service No. ED 379 293).

Linacre, J. M. (2004). WINSTEPS [Computer program]. Chicago: MESA Press.

Linacre, J. M. (2003). A user's guide to Winsteps: Rasch-model computer programs, Chicago: Winsteps.com.

Mardell-Czudnowski, C. D., (1987). Screening kindergartners at an international school. *School Psychology International, 8*, 127-132.

Vargas, C. & Lang, W.S. (1994). Rasch Model detection of cultural bias in preschool testing of South American and Hispanic American children. *Proceedings of the Eight National Conference on Undergraduate Research, 1*, 283-286.

Valencia, R. R. & Rankin, R. J. (1985). Evidence of content bias on the McCarthy Scales with Mexican American children: Implications for test translation and nonbiased assessment. *Journal of Educational Psychology, 77*(2), 197-207.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281-311.

Tyler, F. T. (1969). Readiness. In R. L. Edel (Ed.), *Encyclopedia of educational research (4th ed.).* Toronto, Canada: The MacMillian Company.

Venet, M., Normandeau, S., Letarte, M. J., & Bigras, M. (2003). Les propriétés psychométriques du Lollipop. *Revue de psychoéducation, 32,* 165-176.

Wolfe, E. W. (2004). Equating and item banking with the Rasch Model. In Smith, E. & Smith, R. (Eds.) *Introduction to Rasch Measurement* (pp. 360-390), Maple Grove, MN: JAM Press.

Wright, B. D. (1996). Comparisons require stability. *Rasch Measurement Transactions.* 10 (2), 506.

Wright, B. D., & Mok, M. M. C (2004). An overview of the family of Rasch measurement models. In Smith, E. & Smith, R. (Eds.) *Introduction to Rasch Measurement* (pp. 1-24), Maple Grove, MN: JAM Press.

About the Authors

W. Steve Lang, Ph.D. is a Professor of Educational Measurement and Research at the University of South Florida St. Petersburg. He specializes in the Rasch model of Item Response Theory. He has taught in the public school and at the university level. E-mail: lang@mail.usf.edu

Alex L. Chew, Ed. D. is Professor Emeritus of Educational Psychology and Counseling at Georgia Sothern University. He has worked as an elementary and secondary teacher, school counselor, school psychologist, and educational administrator. He is a member of the Licensed Professional Counselors Association. alexchew@frontiernet.net

Carol Crownover, Ed. D. is the lead resource teacher for the Seminole County pre-school kindergarten program and an adjunct professor in Early Childhood Education at Nova Southeastern University. She was an elementary school classroom teacher for ten years. Carol_Crownover@scps.k12.fl.us

Judy R. Wilkerson, Ph.D., is an Associate Professor of Research and Assessment at Florida Gulf Coast University. She and Dr. Lang have co-authored two books on teacher assessment in the cognitive and affective domains and present regularly at international and U.S. professional meetings in these areas, on assessment of student learning outcomes, and accreditation and accountability.
E-mail: jwilkers@fgcu.edu