istation's Indicators of Progress

Early Reading

Validity and Reliability Evidence

for Pre-Kindergarten

istation Research Report 2010-01

Rev A

March 2010

800 E Campbell Rd, Ste 224
Richardson, TX 75081
(866) 883-READ
info@istation.com

Summary

During the 2009-10 school year, a study using ISIP™, istation's Indicators of Progress, Early Reading assessment program was conducted in eleven Pre-Kindergarten classes from a large north Texas school district. Data were examined for concurrent validity with external measures, as well as internal consistency and test-retest reliability. Results show large to very large evidence of reliability and validity with regards to letter knowledge, vocabulary, phonemic awareness, and comprehensive reading ability for students in Pre-Kindergarten.

The principal investigator for the study was Patricia Mathes, PhD, Texas Instruments Foundation Chair in Reading Research and Director of the Institute for Reading Research at Southern Methodist University. Dawn Levy, MEd, was the study coordinator.

Correspondence concerning the study should be addressed to Dr Patricia Mathes, The Institute for Reading Research, Southern Methodist University, Post Office Box 750381, Dallas, Texas 75275-0381. E-mail: PMathes@smu.edu

Acknowledgments: The current study was conducted through the generous support of the Today Foundation.

Correspondence concerning this report should be addressed to Dr Kevin E Kalinowski, Director of Research, istation, 800 East Campbell Road, Suite 224, Richardson, Texas 75081. E-mail: KKalinowski@istation.com

istation's Indicators of Progress Early Reading
Validity and Reliability Evidence for Pre-Kindergarten

ISIP™, istation's Indicators of Progress, is a computer adaptive continuous progress monitoring assessment of critical reading skills. In addition to overall reading ability, ISIP Early Reading measures abilities in the key reading areas of phonemic awareness, alphabetic knowledge, fluency with text, vocabulary, and comprehension, as outlined by the National Reading Panel (National Institute of Child Health and Human Development, 2000). ISIP is Internet-based and can be administered individually or as a group. As an engaging computer animated program using a game-like interface, ISIP eliminates human error and subjectivity in measuring reading ability. Furthermore, ISIP provides immediate feedback for differentiated tiered instruction using the included teacher resources, istation Reading online reading intervention program, or a core reading system.

IRT-based CAT

During the 2007-08 school year, a two-parameter logistic item response theory (2PL-IRT) calibration study was conducted with early reading assessment items developed by Patricia Mathes and Joe Torgesen in the areas of phonemic awareness, letter knowledge, alphabetic decoding, spelling, vocabulary, and reading comprehension. The study resulted in a pool of 1,550 items with reliable discrimination and difficulty parameter estimates aligned on a common scale.

Subsequently, the items were encoded into a computerized adaptive testing (CAT) version of ISIP, called ISIP Early Reading. ISIP Early Reading dynamically presents the most informative item to students based on how well the item's difficulty matches the student's ability. When the standard error of the estimate falls below a preset threshold, the testing administration stops, and final estimates of ability are computed using a Bayesian estimator, one for each of the subtests, plus a comprehensive reading ability.

Current Study

Data from ISIP Early Reading have been shown to be valid and reliable (istation, 2009). Although the initial set of items was targeted for students in Kindergarten through Grade 3, the items were developed at a wide range of abilities, including older students performing below grade level, plus younger students such as those in Pre-Kindergarten (Pre-K). To establish validity evidence for the younger population, data were collected during the 2009-10 school year from eleven Pre-K classes at five elementary schools (A-E) from a large north Texas school district, which was different from the district used in the IRT calibration study or in the previous validity study. Demographics of the study participants are found in Table 1.

Table 1
Student Demographics

	Pre-K	
Students	179	
By School		
A	27	(15.1%)
B	33	(18.4%)
C	37	(20.7%)
D	28	(15.6%)
E	54	(30.2%)
By Gender		
Male	91	(50.8%)
Female	88	(49.2%)
By Race/Ethnicity		
African American	35	(19.6%)
Asian	26	(14.5%)
Hispanic	35	(19.6%)
Other	4	(2.2%)
Pacific Islander	1	(0.6%)
White	78	(43.6%)
Qualifying for Free/Reduced Lunch	140	(78.2%)
Receiving ESL Services	14	(7.8%)
In a Bilingual Classroom	2	(1.1%)
English Language Leaner (ELL)	17	(9.5%)
Having a disability	2	(1.1%)
Receiving Special Ed Services	2	(1.1%)

Note. Percentages may not add up to 100% for a given category due to rounding.

The schools included in the study used ISIP throughout the 2009-2010 school year. At the beginning of each month, ISIP assessments were automatically administered to students during regularly scheduled computer lab time. In some cases, school coordinators from Southern Methodist University (SMU) assisted teachers in proctoring ISIP. In addition to ISIP, SMU school coordinators administered external measures to participating students in each school over the course of a week during November. Prior to administering any external measure, the SMU school coordinators underwent training on each instrument to increase inter-rater reliability. A four group Latin squares design was utilized to reduce ordering effects. The external measures were selected based on the reading skills being measured, as well as their suitability for Pre-Kindergarten students, as indicated in Table 2.

Table 2
Assessments Administered by Skill

Assessment	*Letter Knowledge*	*Vocabulary*	*Phonemic Awareness*	*Comprehensive Ability*
ISIP Early Reading	Sep–Dec	Sep–Dec	Nov–Dec	Sep–Dec
ELSA	Nov		Nov	
Letter Names	Nov			
Letter Sounds	Nov			
PPVT-4		Nov		
TOPEL	Nov	Nov	Nov	Nov

The ISIP Early Reading assessment measures abilities in the domains of phonemic awareness, alphabetic knowledge, fluency with text, vocabulary, and comprehension. However, only the subtests Letter Knowledge (through alphabet letter recognition and letter-sound correspondence items), Vocabulary (through oral-picture correspondence items), and Phonemic Awareness (through initial sound and blending items) are appropriate for emergent readers enrolled in Pre-Kindergarten. At the end of each session, responses from all subtests are combined, and a comprehensive reading ability measure, called Overall Reading, is estimated using IRT.

Regarding the external measures used in the current study, the Early Literacy Skills Assessment (ELSA; DeBruin-Parecki, 2005) is unique in that the assessment is presented to students in the form of a children's storybook. ELSA measures Comprehension (through prediction, retelling, and connection to real life items), Phonological Awareness (through rhyming, segmentation, and phonemic awareness items), Alphabetic Principle (through sense of word, alphabet letter recognition, and letter-sound correspondence items), and Concepts about Print (through orientation, story beginning, direction of text, and book part items). ELSA is not norm-referenced. Instead, ELSA identifies children in one of three developmental levels for each subtest: Level 1, Early Emergent; Level 2, Emergent; and Level 3, Competent Emergent. Letter Names and Letter Sounds measure a student's ability to recognize each of the 26 letters, randomly presented, by name and by sound. The Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn & Dunn, 2007) was designed to measure the oral vocabulary of children and adults. The Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2007) was designed to identify students in Pre-Kindergarten who might be at risk for literacy problems that affect reading and writing. TOPEL consists of four subtests, Print Knowledge (through written language conventions and alphabetic knowledge items), Definitional Vocabulary (through oral vocabulary and word meaning items), Phonological Awareness (through elision and blending items), as well as a composite score known as the Early Literacy Index. Both PPVT-4 and TOPEL are norm-referenced tests.

Reliability Evidence

Cronbach's (1951) coefficient alpha is often used as an indicator of reliability across test items within a testing instance. However, alpha assumes all students in the testing instance respond to a common set of items. Due to its very nature, students taking a CAT-based assessment, such as ISIP Early

Reading, will receive a custom set of items based on their initial estimates of ability and response patterns. The IRT analogue to classical internal consistency is marginal reliability (Bock & Mislevy, 1982). In essence, marginal reliability is a method of combining the variability in estimating abilities at different points on the ability scale into a single index. Like Cronbach's alpha, marginal reliability is a unitless measure bounded by 0 and 1, and it can be used with Cronbach's alpha to directly compare the internal consistencies of classical test data to IRT-based test data. ISIP Early Reading has a stopping criteria based on minimizing the standard error of the ability estimate. As such, the lower limit of the marginal reliability of the data for any testing instance of ISIP will always be approximately 0.90.

To establish test-retest reliability evidence, Pearson product moment correlation coefficients between ISIP Early Reading administrations were computed. Results for ISIP Letter Knowledge, Vocabulary, and Overall Reading ability range from 0.532 to 0.735 across four months of testing sessions, September to December, as indicated in Tables 3 through 5. Students had to demonstrate minimal ability before being presented the ISIP Phonemic Awareness subtest; unlike the ISIP Letter Knowledge and Vocabulary subtests, where all students were given both every month. In November only four students met the criteria, and in December only 23 students met the criteria. Therefore, there was insufficient power to perform statistical analysis for Phonemic Awareness reliability.

Table 3
ISIP Early Reading Letter Knowledge Test-Retest Reliability[a] *between Testing Sessions*

	Sep	*Oct*	*Nov*	*Dec*
Sep	---			
Oct	0.632[**] (171)	---		
Nov	0.650[**] (165)	0.699[**] (172)	---	
Dec	0.538[**] (163)	0.532[**] (170)	0.735[**] (167)	---

[a]Pearson product moment correlations (*r*).
[**]Statistically significant (H_0: r=0) at p<.01.
Note. Sessions occurred at the start of the month indicated. *N* for each correlation is within parentheses.

Table 4
ISIP Early Reading Vocabulary Test-Retest Reliability[a] *between Testing Sessions*

	Sep	*Oct*	*Nov*	*Dec*
Sep	---			
Oct	0.683[**] (171)	---		
Nov	0.577[**] (168)	0.658[**] (175)	---	
Dec	0.571[**] (169)	0.691[**] (176)	0.644[**] (173)	---

[a]Pearson product moment correlations (*r*).
[**]Statistically significant (H_0: r=0) at p<.01.
Note. Sessions occurred at the start of the month indicated. *N* for each correlation is within parentheses.

Table 5

ISIP Early Reading Overall Reading Test-Retest Reliability[a] *between Testing Sessions*

	Sep	*Oct*	*Nov*	*Dec*
Sep	---			
Oct	0.687[**] (171)	---		
Nov	0.706[**] (168)	0.701[**] (175)	---	
Dec	0.669[**] (169)	0.652[**] (176)	0.707[**] (173)	---

[a]Pearson product moment correlations (*r*).

[**]Statistically significant (H_0: r=0) at p<.01.

Note. Sessions occurred at the start of the month indicated. *N* for each correlation is within parentheses.

Validity Evidence

Content validity was established through a series of steps to substantiate the test development process. First, early reading content experts, Patricia Mathes and Joe Torgesen, created ISIP Early Reading assessment items in key developmental areas, as suggested by the National Reading Panel (National Institute of Child Health and Human Development, 2000). Next, the items underwent review by a panel of reading specialists. Then, the items were operationally used in a previous version of ISIP and revised as necessary. For ISIP Early Reading, the items were calibrated under a 2PL-IRT model. Finally, item parameters were examined and those items with unacceptable fit statistics with regards to the subtest to which they measured were removed from the pool. Based on the combined processes used to establish content validity, the items in the operational pool grouped by subtest are believed to be accurate representations of the domain in which they intend to measure.

Concurrent validity evidence was established by computing Pearson product moment correlation coefficients between ISIP Early Reading subtests and appropriate external measures, as illustrated in Table 6. Because students had to demonstrate minimal ability before being presented the ISIP Phonemic Awareness subtest only four students met the criteria in November. Therefore, December ISIP Phonemic Awareness scores were used for validity analysis.

Table 6

Correlations[a] *between External Measures and ISIP Early Reading Scores*

ISIP Subtest / *External Measure*	*r (N)*
ISIP Letter Knowledge (November)	
ELSA Alphabetic Principle Level	0.747** (172)
ELSA Upper Case Subtest Score	0.726** (172)
ELSA Lower Case Subtest Score	0.692** (172)
ELSA Letter Sounds Subtest Score	0.636** (172)
Letter Name Score	0.727** (172)
Letter Sound Score	0.669** (172)
TOPEL Print Knowledge Std Score	0.735** (170)
ISIP Vocabulary (November)	
PPVT-4 Std Score	0.625** (173)
TOPEL Definitional Vocabulary Std Score	0.520** (173)
ISIP Phonemic Awareness (December)	
ELSA Phonological Awareness Total Score	0.549** (23)
ELSA Rhyming Subtest Score	0.485* (23)
ELSA Phonemic Awareness Subtest Score	0.620** (23)
TOPEL Phonological Awareness Std Score	0.242 (23)
ISIP Overall Reading (November)	
TOPEL Total Std Score	0.677** (173)
TOPEL Early Literacy Index	0.676** (173)

[a]Pearson product moment correlations (*r*).

*Statistically significant (H_0: r=0) at p<.05. **Statistically significant (H_0: r=0) at p<.01.

Note. Sessions occurred at the start of the month indicated. *N* for each correlation is within parentheses.

Discussion

Reliability and validity are two important qualities of measurement data. Reliability can be thought of as consistency, either consistency over items within a testing instance or over scores from multiple testing instances, whereas validity can be thought of as accuracy, either accuracy of the content of the items or of the constructs being measured. In this study, both qualities were examined using ISIP Early Reading data collected from Pre-Kindergarten students in north Texas elementary schools during the 2009-2010 school year.

Regarding measures of reliability in the current study, ISIP Early Reading produced stable scores over time, even between testing instances four months apart (see Tables 3–5). These test-retest reliability results could stem from a number of converging reasons. First, the exit criteria of the adaptive algorithm used in ISIP produces consistently strong levels of internal consistency, at approximately 0.90, both in the subtest ability scores as well in the overall reading ability scores. Second, the authors, reading experts Patricia Mathes and Joe Torgesen, took great care in constructing the ISIP Early Reading item pool, basing the item types and content on contemporary findings in early reading research. Furthermore, the ISIP Early Reading items have been operational for several years in previous versions of the program. Inconsistent items have been culled over time, resulting in a very stable item pool.

Finally, ISIP Early Reading is an engaging and adaptive computer-based assessment program. Items are presented to students at their ability and using high quality computer animation. Students feel they are "playing a game" rather than "taking another test," which probably results in less off task behavior during assessment, producing more consistent results.

Evidence of concurrent validity can be found in the numerous strong, positive relationships to external measures of reading constructs. Cohen (1988) suggested correlations around 0.3 could be considered moderate and those around 0.5 could be considered large. Hopkins (2010) expanded the upper end of Cohen's scale to include correlations around 0.7 as very large, and those around 0.9 as nearly perfect. Given those criteria, the data from the current study (see Table 6) show mostly large to very large criterion validity with scores from well known norm-referenced measures, such as TOPEL and PPVT-4, as well as the authentic assessment, ELSA.

Specifically for letter knowledge, scores from the ISIP Letter Knowledge (LK) subtest showed strong, positive correlations to scores from comparative ELSA subtests, such as the Upper Case (r = 0.726), Lower Case (r = 0.692), and Letter Sounds (r = 0.636) subtests. In addition, ISIP LK scores correlated very well with Letter Names (r = 0.727) and Letter Sounds (r = 0.669), as well as TOPEL Print Knowledge (r = 0.735). These results suggest that the ISIP Letter Knowledge subtest measures the same construct as other early reading assessments.

Regarding vocabulary, PPVT-4 is most similar to the item format used in ISIP Vocabulary for students with early-emergent reading abilities, namely oral-picture correspondence. Therefore, it is not surprising that the correlation between the two sets of scores was large (r = 0.625). TOPEL Definitional Vocabulary (DV) also uses the oral-picture correspondence item format, but adds a task where participants state the meaning of the target word. Appropriately, the correlation between ISIP Vocabulary and TOPEL DV (r = 0.520) scores was somewhat less than between ISIP and PPVT-4 scores, but it is still considered large.

Participants had to demonstrate repeated minimal ability in ISIP Early Reading to be offered the ISIP Phonemic Awareness (PA) subtest. Because students first took ISIP in September, the first opportunity to take ISIP PA as a Pre-Kindergarten student was in November, when 4 students met the criteria. With insufficient power to compute correlations to external measures, it was decided that ISIP PA scores from December (N = 23) would be used for validity analyses even though the collection of external measures data occurred in November. Both ELSA and TOPEL assess the broader concept of phonological awareness, including onset, rime, and segmentation, whereas ISIP PA assesses phonemic awareness concepts, such as initial sound and phoneme blending. The correlation between ISIP PA and ELSA Phonemic Awareness subtest scores (r = 0.620) was large. However, even the phonological concept of rhyming (as measured by ELSA Rhyming subtest) correlated well with ISIP PA scores (r = 0.485). The overall correlation between ELSA Phonological Awareness and ISIP Phonemic Awareness scores was large (r = 0.549). ISIP PA scores did not show any meaningful correlation to TOPEL Phonological Awareness standard scores (r = 0.242). However, the correlation between TOPEL Phonological Awareness standard scores and ELSA Phonological Awareness total scores was equally insignificant (r = 0.278), which suggests that ISIP Phonemic Awareness subtest and the ELSA phonological/phonemic

subtests were measuring the same construct, but this construct was very different from the construct measured by the TOPEL Phonological Awareness subtest.

Finally, ISIP Early Reading computes a comprehensive measure of reading ability, named Overall Reading, through IRT modeling that utilizes the response pattern from all subtests in a testing session. Scores from ISIP Overall Reading correlated highly with the total standard scores from the TOPEL (r = 0.677), as well as with the TOPEL Early Literacy Index (r = 0.676), which is a 7-level interpretation of performance ranging from Very Poor to Very Superior.

Taken together, the evidence supports the claim that ISIP Early Reading produces reliable and valid data for measuring key domains of emerging reading, such as letter knowledge, vocabulary, phonemic awareness, as well as comprehensive reading ability, for students in Pre-Kindergarten.

References

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

DeBruin-Parecki, A. (2005). *Early literacy skills assessment* (ELSA). Ypsilanti, MI: High/Scope Press.

Dunn, L. M., & Dunn, L. M. (2007). *Peabody picture vocabulary test, fourth edition* (PPVT-4). Minneapolis, MN: Pearson Assessments.

Hopkins, W. G. (2010, February). *A new view of statistics: A scale of magnitudes for effect statistics*. Retrieved from: http://www.sportsci.org/resource/stats/index.html

istation. (2009, August). *istation's indicators of progress early reading reliability and validity evidence* (istation Research Report No. RR2009-01, Rev C). Retrieved from: http://www2.istation.com/research/pdfs/isip_rr.pdf

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *Test of preschool early literacy* (TOPEL). Austin, TX: PRO-ED.

National Institute of Child Health and Human Development. (2000). Report of the National Reading Panel. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.