Istation's Indicators of Progress Early Reading Reliability and Validity Evidence

Istation Research Report 2009-01

Rev C

8150 North Central Expressway
Suite 2000
Dallas, TX 75206

1-866-883-READ (7323)

Summary

During the 2008-09 school year, a validity and reliability study using Istation's ISIP-ER computer-adaptive reading assessment program was conducted in five elementary schools from a north Texas school district. Data were examined for internal consistency, test-retest reliability, concurrent validity with external measures (including DIBELS, TPRI, AND ITBS), and predictive validity with TAKS. Results show moderate to strong evidence of reliability and validity with regards to phonemic awareness, alphabetic knowledge, vocabulary, and reading comprehension.

Conducting the study was Dr. Patricia Mathes, Texas Instruments Foundation Chair in Reading Research and Director of the Institute for Reading Research at Southern Methodist University.

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. Victoria N. Locke, Research Director, Istation, 8150 N. Central Expressway, Suite 2000, Dallas, Texas 76205. E-mail: vlocke@istation.com

ISIP-ER: Istation Indicators of Progress for Early Reading
Reliability and Validity Evidence

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 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, ISIP eliminates human error and subjectivity. Furthermore, ISIP provides immediate feedback for differentiated tiered instruction.

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 Drs. Patricia Mathes and Joe Torgesen in the areas of Phonemic Awareness (PA), Letter Knowledge (LK), Alphabetic Decoding (AD), Spelling (SPL), Vocabulary (VOC), and Reading Comprehension (CMP). The study resulted in a pool of 1,550 Kindergarten through Grade 3 items with reliable discrimination and difficulty parameter estimates aligned on a common scale ranging from 140 to 320.

Subsequently, the items were encoded into a computerized adaptive testing (CAT) version of ISIP, called ISIP-ER (*Istation's Indicators of Progress for Early Reading*). The CAT-based ISIP-ER 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, one for each of the six reading ability subtests, plus an overall reading ability.

Current Study

To establish reliability and validity evidence, data were collected during the 2008-09 school year at five elementary schools (A-E) from a large north Texas independent school district, different from the previous IRT calibration study. Demographics of the study participants are found in Table 1.

Table 1
Student Demographics

					Grade Level	
	K	*1*	*2*	*3*	*K-3*	
Students	122	103	95	96	416	
By School						
A	20	16	15	19	70	(16.8%)
B	21	15	18	18	72	(17.3%)
C	43	37	36	16	132	(31.7%)
D	17	15	11	12	55	(13.2%)
E	21	20	15	31	87	(20.9%)
By Gender						
Male	68	55	52	40	215	(51.7%)
Female	54	48	43	56	201	(48.3%)
By Ethnicity						
African American	21	28	17	10	76	(18.3%)
Caucasian	48	31	32	18	129	(31.0%)
Hispanic	40	38	40	65	183	(44.0%)
Asian	13	6	4	3	26	(6.3%)
Other	0	0	2	0	2	(0.5%)
Qualifying for Free/Reduced Lunch	63	52	44	73	232	(55.8%)
Qualifying for ESL Services	20	15	13	27	75	(18.0%)
Receiving ESL Services	17	15	10	25	67	(16.1%)
In a Bilingual Classroom	0	0	0	32	32	(7.7%)
Receiving Special Ed Services	1	5	6	7	19	(4.6%)

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

Seven testing sessions occurred every two weeks between October and February. For each session, students were escorted by trained data collectors from Southern Methodist University (SMU) in convenience groupings to the school's computer lab for sessions on the CAT-based ISIP-ER program. On average, six items were needed per subtest to establish an ability estimate with a standard error below the threshold, resulting in 13- to 18-minute ISIP-ER testing sessions, depending on the number of skills assessed. The remaining time in each session was spent administering external measures. A seven group Latin squares design was utilized to reduce ordering effect. Students were given assessments for reading skills appropriate for their age as indicated in Table 2.

Table 2

Assessments Administered by Grade

	ISIP-ER							*DIBELS*					
Grade Level	*PA*	*LK*	*AD*	*SPL*	*TF*	*VOC*	*CMP*	*PSF*	*NWF*	*ORF*	*TPRI*[a]	*ITSB*[a]	*TAKS*[a]
K	X	X	X			X		X	X		X		
1	X		X	X	X	X	X	X	X	X		X	
2			X	X	X	X	X		X	X		X	
3				X	X	X	X			X			X

[a]Tests administered by the district.

In addition to ISIP-ER, SMU data collectors administered *Dynamic Indicators of Basic Early Literacy* (DIBELS; Kaminski & Good, 1998) Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF) assessments. Furthermore, one or more additional external measures were administered during each session. These additional assessments include well known instruments in Phonemic Awareness: *Comprehensive Test of Phonological Processes* (CTOPP; Wagner, Torgesen, & Rashotte, 1999); Letter Knowledge: *Woodcock Language Proficiency Battery-Revised* (WLPB-R; Woodcock, 1991); Alphabetic Decoding: *Test of Word Reading Efficiency* (TOWRE; Torgesen, Wagner, & Rashotte, 1999), WLPB-R, and *Wechsler Individual Achievement Test* (WIAT-II; Wechsler, 2005); Spelling: *Woodcock-Johnson III Tests of Achievement* (WJ-III ACH; Woodcock, McGrew, & Mather, 2001) and WIAT-II; Vocabulary: *Peabody Picture Vocabulary Test* (PPVT-III; Dunn & Dunn, 1997) and WLPB-R; and Comprehension: *Gray Oral Reading Tests* (GORT-4; Wiedeholt & Bryant, 2001), WLPB-R, and WIAT-II.

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-ER, 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. It can be used with Cronbach's alpha to directly compare the internal consistencies of classical test data to IRT-based test data. ISIP-ER 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-ER will be approximately 0.90.

To establish test-retest reliability evidence, Pearson product-moment correlation coefficients between ISIP-ER administrations were computed. Results for overall reading ability range from 0.927 to 0.970 (*N* = 416) across all seven sessions spanning from October to February. Table 3 shows the individual test-retest reliability results for overall reading ability.

Table 3
ISIP-ER Overall Reading Test-Retest Reliability[a] *between Testing Sessions*

	Oct 20	Nov 3	Nov 17	Dec 8	Jan 12	Jan 26	Feb 9
Oct 20	---						
Nov 3	0.970	---					
Nov 17	0.962	0.975	---				
Dec 8	0.947	0.962	0.969	---			
Jan 12	0.946	0.963	0.964	0.960	---		
Jan 26	0.936	0.956	0.962	0.960	0.963	---	
Feb 9	0.927	0.945	0.951	0.949	0.958	0.961	---

[a] Pearson product-moment correlations (*r*).
Note. Sessions were two weeks in length and started on the date indicated.

Validity Evidence

Content validity was established through a series of steps to substantiate the test development process. First, early reading content experts, Drs. Patricia Mathes and Joe Torgesen, created ISIP-ER assessment items in key developmental areas. 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-ER, 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-ER subtests and appropriate external measures. Table 4 shows results by grade level. During each of the seven testing sessions, both ISIP-ER and DIBELS were administered to the students in the study. Pearson correlations between DIBELS and ISIP-ER are shown in Table 5. Prior to testing, the SMU testers were trained on administering DIBELS. Inter-rater reliability was ensured during training so that no more than a two point difference in scoring occurred between testers.

The *Texas Primary Reading Inventory* (TPRI; Texas Education Agency, 1998) was administered to all Kindergarten students by the district three times during the school year, beginning of the year (BOY), middle of the year (MOY), and end of the year (EOY). Data for students in the current study were provided by the district at the end of the school year. Pearson correlations between TPRI subtests and ISIP-ER subtests are found in Table 6. It is unknown when these testing administrations occurred, so data from the most appropriate ISIP-ER testing sessions were used in the comparisons. The training and inter-rater reliability of the district testers is also unknown.

The *Iowa Tests of Basic Skills* (ITBS; Hoover, Dunbar, & Frisbie, 2007) was administered by the district in October to all students in Grades 1 and 2. Data for students in the current study were provided by the district at the end of the school year. Pearson correlations between ITBS Reading and ISIP-ER Overall Reading are shown in Table 7.

To establish predictive validity evidence, Pearson correlations between ISIP-ER Overall Reading ability and the *Texas Assessment of Knowledge and Skills* (TAKS; Texas Education Agency, 2003) were computed for Grade 3. Results are found in Table 8. TAKS was administered by the district in March.

Table 4

Correlations[a] *between External Measures and ISIP-ER Subtest Scores for Grades K-3*

ISIP-ER Subtest			*Grade Level*				
	External Measure		*K*	*1*	*2*	*3*	*K-3*
Phonemic Awareness (PA)							
	CTOPP Blending Words	*r*	**.688**	.431			**.702**
		N	120	100			220
	CTOPP Blending Non Words	*r*	**.676**	.336			**.650**
		N	120	100			220
	CTOPP Segmenting Words	*r*	**.644**	.344			**.620**
		N	122	101			223
	CTOPP Sound Matching	*r*	**.624**	.474			**.662**
		N	122	101			223
Letter Knowledge (LK)							
	Letter Names	*r*	**.593**				**.593**
		N	121				121
	Letter Sounds	*r*	**.693**				**.693**
		N	121				121
	WLPB-R Letter Word Identification	*r*	**.711**				**.711**
		N	120				120
Alphabetic Decoding (AD)							
	TOWRE Phonemic Decoding	*r*	**.582**	**.679**	**.539**		**.838**
		N	122	103	93		313
	TOWRE Sight Word Efficiency	*r*	**.583**	**.626**	**.586**		**.811**
		N	120	100	93		313
	WLPB-R Word Attack	*r*	**.535**	**.701**	**.702**		**.830**
		N	122	102	94		316
	WIAT-II Target Words	*r*		**.624**	**.507**		**.589**
		N		101	92		193
Spelling (SPL)							
	WJ-III ACH Spelling	*r*		**.800**	**.823**	**.798**	**.890**
		N		103	94	96	293
	WIAT-II Spelling	*r*		**.726**	**.774**	**.788**	**.875**
		N		101	91	96	288
Fluency with Text (TF)							
	DIBELS ORF[b]	*r*		**.741**	**.667**	**.627**	**.766**
		N		103	92	94	289
Vocabulary (VOC)							
	PPVT-III	*r*	**.687**	**.696**	**.582**	**.785**	**.814**
		N	121	101	94	95	411
	WLPB-R Vocabulary	*r*	.368	**.656**	**.702**	**.716**	**.836**
		N	121	103	94	96	414
Comprehension (CMP)							
	GORT-4 Comprehension	*r*		.456	.354	.473	**.621**
		N		102	95	94	291
	WLPB-R Comprehension	*r*		**.707**	**.597**	**.569**	**.794**
		N		102	92	93	287
	WIAT-II Reading Comprehension	*r*		**.630**	**.554**	**.596**	**.682**
		N		101	91	96	288

[a] Pearson product-moment correlations (*r*). [b] Feb 9 session data used for correlations.

Note. Empty cells indicate no students were administered that instrument for that grade level. Correlations above 0.5 are highlighted.

Table 5

Correlations[a] between DIBELS Scores and ISIP-ER Subtest Scores for Grades K-3

		PSF[b]					*NWF[c]*					*ORF[d]*				
Testing		*Grade Level*					*Grade Level*					*Grade Level*				
Session		*K*	*1*	*2*	*3*	*K-3*	*K*	*1*	*2*	*3*	*K-3*	*K*	*1*	*2*	*3*	*K-3*
Oct 20	*r*	**.645**	.476			**.707**	.454	.434	.375		**.724**		**.657**	**.699**	**.811**	**.826**
	N	98	92			190	96	94	84		274		87	81	73	241
Nov 3	*r*	**.612**	.388			**.678**	.432	**.519**	.497		**.794**		**.593**	**.711**	**.709**	**.794**
	N	121	103			224	121	103	93		317		100	93	91	284
Nov 17	*r*	**.712**	.365			**.711**	**.578**	**.571**	**.524**		**.807**		**.656**	**.735**	**.733**	**.827**
	N	121	102			223	121	102	93		316		102	93	96	291
Dec 8	*r*	**.649**	.406			**.649**	**.574**	**.636**	**.607**		**.820**		**.640**	**.682**	**.619**	**.752**
	N	121	102			223	121	102	92		315		101	93	94	288
Jan 12	*r*	**.624**	.238			**.558**	**.605**	.490	.649		**.802**		**.590**	**.707**	**.601**	**.748**
	N	120	102			222	120	102	86		308		102	91	95	288
Jan 26	*r*	**.532**	.171			.478	**.547**	**.593**	**.514**		**.780**		**.661**	**.708**	**.647**	**.777**
	N	121	102			223	121	102	91		314		102	91	94	287
Feb 9	*r*	.496	.253			**.517**	**.597**	**.539**	.438		**.764**		**.741**	**.667**	**.627**	**.766**
	N	122	102			224	122	103	92		317		103	92	94	289

[a]Pearson product-moment correlations (*r*). [b]ISIP-ER PA subtest scores used for correlations. [c]ISIP-ER AD subtest scores used for correlations. [d]ISIP-ER TRM subtest scores used for correlations.

Note. Empty cells indicate no students were administered that instrument for that grade level. Correlations above 0.5 are highlighted.

Table 6

Correlations[a] between TPRI Subtest Scores and ISIP-ER Subtest Scores for Kindergarten

		Phonemic Awareness[b]					*Graphophonemic Knowledge[c]*	
		Rhy[d]	*BWP[e]*	*BP[f]*	*DIS[g]*	*DFS[h]*	*LNI[i]*	*LtSL[j]*
BOY[k]	*r*	.475	**.557**	**.555**	.483	.404	**.728**	**.561**
	N	109	97	91	88	88	109	97
MOY[l]	*r*	.334	**.598**	**.602**	**.575**	**.558**	**.629**	**.551**
	N	109	101	98	97	88	109	106
EOY[m]	*r*	.267	.426	.431	.471	.440	.387	.408
	N	110	110	108	106	97	110	109

[a]Pearson product-moment correlations (*r*). [b]ISIP-ER PA subtest scores used for correlations. [c]ISIP-ER LK subtest scores used for correlations. [d]Rhyming. [e]Blending Word Parts. [f]Blending Phonemes. [g]Deleting Initial Sounds. [h]Deleting Final Sounds. [i]Letter Name Identification. [j]Letter to Sound Linking. [k]ISIP-ER Nov 17 session data used for correlations. [l]ISIP-ER Jan 12 session data used for correlations. [m]ISIP-ER Feb 9 session data used for correlations.

Note. TPRI administered by the district. It is unknown when in the school year TPRI was administered or by whom. Correlations above 0.5 are highlighted.

Table 7

Correlations[a] between ITBS Reading Scale Scores and ISIP-ER Overall Reading Scores for Grades 1 and 2

Testing		*Grade Level*		
Session		*1*	*2*	*1-2*
Oct 20	*r*	**.807**	**.845**	**.895**
	N	62	75	137
Nov 3	*r*	**.808**	**.821**	**.884**
	N	65	78	143
Nov 17	*r*	**.793**	**.839**	**.888**
	N	65	78	143
Dec 8	*r*	**.806**	**.741**	**.845**
	N	65	78	143
Jan 12	*r*	**.748**	**.837**	**.874**
	N	64	78	142
Jan 26	*r*	**.725**	**.806**	**.854**
	N	65	78	143
Feb 9	*r*	**.699**	**.768**	**.829**
	N	65	77	142

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

Note. ITBS administered by the district in October. Correlations above 0.5 are highlighted.

Table 8

Correlations[a] between TAKS Reading Scale Scores and ISIP-ER Overall Reading Scores Plus DIBELS ORF Scores for Grade 3

Testing		*ISIP-ER*	*DIBELS*
Session		*Overall Reading*	*ORF*
Oct 20	*r*	**.740**	**.630**
	N	64	60
Nov 3	*r*	**.741**	**.551**
	N	74	75
Nov 17	*r*	**.698**	**.598**
	N	77	77
Dec 8	*r*	**.695**	.450
	N	77	76
Jan 12	*r*	**.698**	**.582**
	N	76	77
Jan 26	*r*	**.741**	**.555**
	N	74	75
Feb 9	*r*	**.710**	**.533**
	N	77	76

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

Note. TAKS administered by the district in March. Correlations above 0.5 are highlighted.

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-ER data collected from Kindergarten through Grade 3 students in north Texas elementary schools during the 2008-09 school year.

Regarding measures of reliability, the data from the current study suggest very high levels of internal consistency, both in the subtest ability scores as well in the overall reading ability scores. In addition, ISIP-ER produced extremely stable scores over time, even between testing instances five months apart. These outstanding results could stem from a number of converging reasons. First, the authors, reading experts Drs. Patricia Mathes and Joe Torgesen, took great care in constructing the ISIP-ER item pool. They utilized the most up to date findings in early reading research as a basis for the item types and content they produced for Istation. Furthermore, the ISIP-ER 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-ER 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 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 (2009) 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 show mostly large to very large criterion validity with scores from well-known external measures, such as CTOPP, GORT-4, PPVT-III, TOWRE, WJ-III ACH, WLPB-R, and WIAT-II, as well as with TPRI and ITBS. In addition, validity results show that ISIP-ER Overall Reading is a stronger predictor than DIBELS ORF for TAKS Reading, using scores from one to six months prior to TAKS administration.

Taken together, the evidence supports the claim that ISIP-ER produces reliable and valid data for measuring key areas of reading development, such as phonemic awareness, alphabetic knowledge, vocabulary, and reading comprehension, as well as overall reading ability.

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.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test-Third edition* (PPVT-III). Circle Pines, MN: American Guidance Service.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (DIBELS). Eugene, OR: Institute for the Development of Education Achievement.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A. (2007). *Iowa tests of basic skills* (ITBS). Rolling Meadows, IL: Riverside Publishing.

Hopkins, W. G. (2009). *A new view of statistics: A scale of magnitudes for effect statistics*. Retrieved July 16, 2009, from http://www.sportsci.org/resource/stats/index.html

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.

Texas Education Agency. (2003). *Texas assessment of knowledge and skills* (TAKS). Austin: Author.

Texas Education Agency. (1998). *Texas primary reading inventory* (TPRI). Austin: Author.

Torgesen, J. K., Wagner, R., & Rashotte, C. (1999). *Test of word reading efficiency* (TOWRE). Austin, TX: Pro-Ed.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive test of phonological processes* (CTOPP). Austin, TX: Pro-Ed.

Wechsler, D. (2005). *Wechsler individual achievement test-Second edition* (WIAT-II). San Antonio, TX: Harcourt Assessment.

Wiederholt, J. L., & Bryant, B. R. (2001). *Gray oral reading tests-Fourth edition* (GORT-4). Austin, TX: Pro-Ed.

Woodcock, R. W. (1991). *Woodcock language proficiency battery-Revised* (WLPB-R). Rolling Meadows, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement* (WJ-III ACH). Rolling Meadows, IL: Riverside Publishing.