



Assessment and Accountability Comprehensive Center (AACC)
Evaluation of the Technical Adequacy of Evidence of Assessments of
English Language Proficiency: Body of Evidence Summary

Assessment: IDEA Proficiency Test-Reading & Writing (IPT-R/W) 2004.

This body of evidence summary reports the results of the evaluation of technical evidence in support of the IPT-R/W 2004, as analyzed against a validated list of technical adequacy criteria. The table below outlines the types of validity, reliability, and bias and sensitivity evidence associated with various phases of test development in the order they are discussed in this summary. The detailed text following this table explains which quality elements met or exceeded quality expectations. It also provides recommendations for additional types of evidence that would provide support for the technical quality of the assessment. Elements of evidence are divided into Tier 1 and Tier 2. Tier 1 elements ought to be parts of a test’s body of evidence, considering both phase and type of development. Tier 2 elements are important, but may include elements that are specific to a particular test. For a complete list of the Tier 1 and Tier 2 elements, see document: *Assessments for English Language Learners: Technical Adequacy Criteria—Tiers*. For operational definitions of the criteria, see document: *ELL Assessment Technical Criteria—Operational Definitions*. For information regarding the evaluation of the assessment’s technical evidence and the technical criteria used, refer to the AACC/WestEd report titled *Evaluation of the Technical Evidence of Assessments for Special Student Populations* at <http://www.aacompcenter.org> (see Special Populations page).

Type	Phase (Number of possible elements)
Construct validity	Test design and development (10)
Content validity	Test design and development (13)
	Scoring (4)
	Field testing (3)
Consequential validity	Test design and development (1)
	Security (1)
	Reporting (4)
Criterion validity	Test design and development (2)
Reliability	Test design and development (11)
	Scoring (2)
Bias and sensitivity	Test design and development (13)



Body of Evidence

The following documents comprise the body of evidence analyzed for this assessment:

- Technical Manual, IPT 1 & 2, Reading & Writing, Grades 2–6, English, Forms 1A & 2A. (2004) Ballard & Tighe.
- Technical Manual, IPT 1 & 2, Reading & Writing, Grades 2–6, English, Forms 1B & 2B. (2004) Ballard & Tighe.
- Technical Manual, IPT 3, Reading & Writing, Grades 7–12, English, Forms 3A & 3B. (2004) Ballard & Tighe.
- IPT NCLB Compliance Handbook (2004) Ballard & Tighe, including IPT 1 Reading and Writing Examiner’s Manual Addendum; IPT 2 Reading and Writing Examiner’s Manual Addendum; IPT 1 & 2 Reading and Writing Technical Manual Addendum; IPT 3 Reading and Writing Technical and Examiner’s Manual Addendum
- Examiner’s Manual, IPT Early Literacy, Grades K–1, English, Second Edition. (2006) Ballard & Tighe.
- Technical Manual, IPT Early Literacy, Grades K–1, English, Second Edition. (2006) Ballard & Tighe.
- Examiner’s Manual, IPT 1, Grades 2–3, English, Form 1C. (2005) Ballard & Tighe.
- Examiner’s Manual, IPT 1, Grades 2–3, English, Form 1D. (2005) Ballard & Tighe.
- Technical Manual, IPT 1, Technical Manual, Grades 2–3, English, Form 1C & 1D. (2005) Ballard & Tighe.
- Examiner’s Manual, IPT 2, Grades 4–6, English, Form 2C. (2005) Ballard & Tighe.
- Examiner’s Manual, IPT 2, Grades 4–6, English, Form 2D. (2005) Ballard & Tighe.
- Technical Manual, IPT 2, Grades 4–6, English, Form 2C & 2D. (2005) Ballard & Tighe.
- Examiner’s Manual, IPT 3, Examiner’s Manual, Grades 7–12, English, Form 3C. (2005) Ballard & Tighe.
- Examiner’s Manual, IPT 3, Grades 7–12, English, Form 3D. (2005) Ballard & Tighe.
- Technical Manual, IPT 3, Technical Manual, Grades 7–12, English, Form 3C & 3D. (2005) Ballard & Tighe.

Across all types and phases, the technical evidence associated with the IPT-RW received a rating of meeting or exceeding technical quality expectations in 14 of the 64 evidence/method elements. Further description of specific evidence/method elements of technical adequacy follows.

Construct Validity

Test Design and Development

Of the ten evidence/method elements of construct validity in the test design and development phase, five received a rating of meeting or exceeding technical quality expectations. The five elements that met or exceeded expectations were

- *test purpose,*
- *population/classification,*
- *theoretical foundation/framework,*



- *multi-trait/multi-method/subtest inter-correlation*, and
- *fidelity*.

Three of these elements are Tier 1 elements. Evidence of *standardization* that meets or exceeds expectations is also desired to determine construct validity.

Content Validity

Test Design and Development

Of the 13 evidence/method elements of content validity in the test design and development phase, two received a rating of meeting or exceeding technical quality expectations.

The elements that met or exceeded expectations were

- *p-values/point biserials* and
- *linking/equating*.

P-values/point biserials is a Tier 1 element, but evidence of *alignment (items-to-standards)*, *expert judgment*, *test blueprint*, *alignment (test form-to-blueprint)*, *IRT/test fit*, and that meets or exceeds expectations is desired to determine content validity.

Scoring

Of the four evidence/method elements of content validity in the scoring phase, three received a rating of meeting or exceeding technical quality expectations. The elements that met or exceeded expectations were

- *scale*,
- *rubric*, and
- *training of scorers/scoring protocol*.

Scale is a Tier 1 element, but evidence of *standard setting* that meets or exceeds expectations is desired to determine content validity.

Field Testing

Of the three evidence/method elements of content validity in the field testing phase, *blueprint*, *sampling*, and *norming*, none received a rating of meeting or exceeding technical quality expectations.

Consequential Validity

Test Design and Development

The one evidence/method element of consequential validity in the test design and development phase, *use of results*, received a rating of meeting or exceeding technical quality expectations.

Reporting

Of the four evidence/method elements of consequential validity in the reporting phase, two received a rating of meeting or exceeding technical quality expectations. The elements that met or exceeded expectations were

- *N* and
- *central tendency/variation*.



These are Tier 1 elements, but evidence of *reporting category* that meets or exceeds expectations is desired to determine consequential validity.

Security

The one evidence/method element of consequential validity in the security phase, *protocols*, did not receive a rating of meeting or exceeding technical quality expectations.

Criterion Validity

Test Design and Development

Of the two evidence/method elements of criterion validity in the test design and development phase, *cross tabulations* and *Pearson correlation*, neither received a rating of meeting or exceeding technical quality expectations.

Reliability

Test Design and Development

Of the 11 evidence/method elements of reliability (stability and consistency, internal consistency, generalizability, and classification consistency) in the test design and development phase, four received a rating of meeting or exceeding technical quality expectations. The elements that met or exceeded expectations were

- *coefficient alpha*,
- *standard error of measurement/confidence intervals*,
- *alternate form*, and
- *test-retest*.

These four elements are Tier 1 elements.

Scoring

Of the two evidence/method elements of reliability (inter-rater) in the scoring phase, both received a rating of meeting or exceeding technical quality expectations. The elements that met or exceeded expectations were

- *correlation (kappa)* and
- *percent correspondence*.

These two elements are Tier 1 elements.

Bias and Sensitivity

Test Design and Development

Of the 13 evidence/method elements of *expert review* and *DIF analysis* across seven types of bias and sensitivity (linguistic, ethnicity/race, cultural/religious, geographic, SES, disability, and gender), none received a rating of meeting or exceeding technical quality expectations.

Preliminary AACC Comments:

In response to notification of the IPT 2004 evidence evaluation and request for additional documentation, Ballard & Tighe responded that the IPT 2004 series of tests are “currently out of



date and print.” Additionally, the tests are “specifically used for placement only at the district and school levels and are not relevant for NCLB Title I and Title III purposes.” The publisher requested that the AACC instead review the evidence associated with their current, NCLB-compliant test materials, and provided the AACC with materials for review. Analysts will review the new IPT materials. However, given that the IPT 2004 is currently approved for use in at least two states¹ and is in use at the local level for placement, a review of the evidence associated with the IPT 2004 was conducted. The evidence presented in this summary represents that review.

Test Publisher Comments:

The publisher provided a detailed response to the AACC evidence summary (below), including sources of evidence that the publisher believed addressed the technical criteria. (IPT Reading & Writing 1A/2A Technical manual and the IPT Reading & Writing 1A Examiner’s manual are noted below, providing specific page and section references. Reference to other IPT Technical and Examiner’s manuals follow the same organizational format and similar information will be found in them as well. Also, unless otherwise stated, section and page references refer to the Technical manual.)

IPT 2004 Reading and Writing Tests

Additional Evidence:

Construct Validity—Test Design and Development

- *multi-trait/multi-method/subtest inter-correlation*: Subtest inter-correlations are found in Section 4.4.2 on page 24.
- *standardization*: This is covered in a couple of places in the *Examiner’s Manual*. See specifically Section 2.3 (Procedures for Testing) on pages 5 and 6, and Section 8 (Rules for Testing) on page 52.

Content Validity—Test Design and Development

- *alignment (items-to-standards)*: This is and has been available on our web site at http://www.ballard-tighe.com/assess/alignment_assess.html
- *expert judgment*: This is discussed in Section 3.0 (item development) on page 6.
- *test blueprint and alignment (test form-to-blueprint)*: This information is presented in Tables R and S (content validity matrix) on pages 20–23.

Content Validity—Scoring

- *standard setting (cut score and proficiency levels)*: Section 4.5 (Establishment of Cutoff Scores for Non-, Limited, and Competent English Reading Designations), pages 29–40, and section 5.4 (IPT Designation Standards for Writing), pages 51–53, both deal with this topic.

¹ United States Government Accountability Office. (2006). *No Child Left Behind Act: Assistance from Education Could Help States Better Measure Progress of Students with Limited English Proficiency* (GAO-06-815). Washington, DC: U.S. Government Printing Office.



- *rubric and training of scorers/scoring protocol*: This is extensively detailed in the *Examiner's Manual* in section 6.0 (Rating the Writing Test Samples), pages 25–48. Additional information can be found in Appendix E (Rating Sheet for Writing Samples), Appendix F (Sample Student Writing Tests for Scoring Practice), and Appendix G (Ratings Key for Sample Student Writing Tests in Appendix F), pages 70–117. In fact, approximately 60% of the 121 pages in the *Examiner's Manual* addresses these two topics.

Content Validity—Field Testing

- *sampling*: This is presented in section 4.1 (demographics), pages 7–12, and in section 5.1 (demographics), pages 41–43.
- *norming*: This is discussed predominantly in section 7.0, pages 55–62.

Criterion Validity—Test Design and Development

- *cross tabulations*: These are in Tables Z and AA on pages 28 and 29, respectively.
- *Pearson correlation*: These are on pages 26–27, in Table W (which shows correlations between the IPT and CTBS tests), and in Tables X and Y (which contain correlations between IPT Reading and Writing scores and teacher's opinions of students' Reading and Writing ability).

Reliability—Test Design and Development

- *test-retest*: This is presented in Tables P and Q on page 18.
- *correlation coefficient, percent correspondence, and classification error*: These are presented in section 5.2, in Tables WW and XX, on page 44.

Bias and Sensitivity—Test Design and Development

- *expert review (ethnicity/race), expert review (cultural/religious), and expert review (gender)*: These are discussed in section 3.0, on page 6.

Final AACC Comments and Recommendations:

AACC analysts considered the additional evidence provided by the publisher (noted above) vis-à-vis the technical review criteria. When the additional evidence was found to meet or exceed expectations, ratings were revised. This summary reflects the analysts' final ratings.

This summary is intended to inform consumers and test publishers of the breadth and depth of evidence relevant to supporting an assessment's validity, reliability, and freedom from bias.

We appreciate the thoughtful comments from Ballard-Tighe in reference to the summary evaluation of the IPT-Reading and Writing 2004 materials.

Examples of our consideration of the additional evidence provided by the publisher is presented below:



Construct Validity—Test Design and Development

- *multi-trait/multi-method*: Inter-correlations were noted but rated as “below expectations.”

Comment provided was as follows:

The intercorrelations presented in this document appear to address the construct of reading comprehension over time, not the development of language proficiency as demonstrated by change in reading ability.

Content Validity—Test Design and Development

- *alignment*: Regarding the link to the alignment documentation, no explicit reference to such studies was found in the Technical Manuals. Additionally, no detailed description of the nature of the studies (e.g., independent versus internal; qualifications of judges) associated with these alignment findings was provided. The criteria for this type of evidence are as follows:

In-process alignment and/or ex post facto alignment studies done (independent); appropriate unit(s) of analysis and model/appropriate dimensions evaluated; explanation of process or results (including limitations). In-process alignment may be done by writers, editors, or other developers and expert reviewers during the item development process. Ex post facto alignment should be done by independent experts in assessment, standards, and relevant content areas. Alignment procedures and studies should look for appropriateness of item content and cognitive level as described in individual standards, and coverage (breadth and depth) as reflected by the set of standards.

Finally, specific information about item-to-standard alignments was provided only in relation to the TESOL standards and the California, Colorado, and Virginia proficiency standards. Alignment information was very general (text comments, checkmarks, bullet points) for all other cited studies (Council of Great City Schools/National Clearinghouse Assessment Standards and the proficiency standards for Arkansas, Hawaii, Idaho, Maryland, Missouri, and Oklahoma).

The rating for this additional evidence changed from a “0” to a “1” and is “below expectations.”

- *expert judgment*: Expert judgment was noted but rated as “below expectations.”

Comment provided was as follows:

A general reference to expert judgment is provided with no accompanying explanation of group composition (e.g., levels of expertise, demographics) or of protocol/methodology used.

Content Validity—Scoring

- *standard setting*: Standard setting was noted but rated as “below expectations.” Comment provided was as follows:

This document provides only limited information about standard setting; little detail is presented about the method used (classification analysis), the expert judges, the standard errors associated with scores at cut points, or proficiency level definitions (NER, LER, CER). Explanation for how proficiency levels have evolved is more detailed.



Content Validity—Field Testing

- *norming*: The pages cited for norming (55-62) do not provide descriptive statistics about the population or information about the processes used for establishing the normative basis for this test. Hence, the score remains a “0” (not present).

Criterion Validity—Test Design and Development

- *cross tabulations*: Cross tabulations were noted but rated as “below expectations.” Comment provided was as follows:

Some technical quality standards for methodology are addressed, but a number of concerns emerged about the information provided: the tables are very difficult to interpret; unclear whether the writing test discriminates well for those students at lower levels (e.g., only 3 fourth grade students designated NES by teachers); and the narrative asserts the appropriateness of the test for proficiency testing, but results do not support this conclusion.

Reliability—Test Design and Development

- *correlation coefficient* and *percent correspondence*: Information about correlations and percent correspondence was noted but rated as “below expectations.” Comment provided for both was as follows:

Information about reliability is presented in this document (e.g., Cramer’s V), but accompanying explanations (e.g., proficiency levels; how teachers integrated state levels with IPT) are not provided.

Bias and Sensitivity—Test Design and Development

- *expert review (ethnicity/race, cultural/religious, gender)*: While expert review is mentioned, few details about the nature of the judges’ expertise nor the judging process were provided; hence the quality ratings for all remain a “1” (below expectations).

The contents of this evaluation summary were developed under a grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.