Guidance for Developing a Technical Manual for English Language Proficiency Assessments (ELPA)

evea

Washington
Idaho
Indiana
Montana
Oregon

edCount, LLC
Center for Assessment
UCLA
Synergy Enterprises, Inc
PIRE

Marianne Perie, Ph.D.

December, 2011

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

Table of Contents

List of Exhibits

Preface

For any assessment, the documentation of the technical quality is a key aspect of providing validity evidence for the score interpretation and use. When working with assessments for special populations, including English learners and students with disabilities, the documentation becomes even more important as the assessments are less well understood. Over the past several years, work has been done through various federally funded grants to develop a framework for documenting the quality of alternate assessments based on alternate achievement standards (AA-AAS; see for example, Marion & Pellegrino, 2006). That work is now applied to English Language Proficiency Assessments (ELPAs).

Typically, ELPAs tend to be more similar to general assessments than AA-AAS. They usually involve multiple-choice items with some constructed-response portions. And, with the exception of the speaking portion, they can be given to whole classes of students simultaneously and scored electronically. Thus, much of the technical documentation is similar to that of a traditional assessment. However, some key differences include identifying the population for the assessment, providing appropriate accommodations, assessing speaking at an individual level, combining scores across the four language domains (reading, writing, speaking, listening), and determining when the scores reflect readiness to exit an English language development program.

One of the strengths of the work done on the documentation of the AA-AAS is in the focus on the building of a validity argument into the technical documentation. This approach was most recently applied in a U.S. Department of Education funded Enhanced Assessment Grant *Evaluating the Validity of English Language Proficiency Assessments* (EVEA; CFDA 84.368). The project involved five states, Idaho, Indiana, Montana, Oregon, and Washington, that worked on collaborative and independent validity plans over a 24-month period between 2009 and 2011. This paper starts from the work done in the EVEA project and proposes a structure for documenting technical and validity evidence for an ELPA.

The approach uses a two-part report structure: 1) a traditional approach to technical information that provides information on various aspects of technical design, implementation, and data collection; and 2) a validity evaluation that organizes that information around a validity argument designed to clearly present the evidence concerning the claims made by the assessment regarding appropriate uses of the scores. The outline in Appendix A shows the suggested structure for the technical documentation. In addition to the two main sections, an overview of the report should be provided to orient the reader to the structure of the document and location of various pieces of technical evidence.

Technical Information

The first section of the technical manual should contain traditional information about test design and development, administration, scaling, scoring, and reporting. Some unique aspects of this structure can be found in the first two chapters, where it is recommended that the purpose of the testing program and the guiding philosophy behind the program be described. For instance, are the ELPA results intended to be used only for determination of whether the student has met the criteria to exit the program? Or are they also used for other purposes such as class placement or program evaluation? Is the philosophy of the program to move students through quickly? Or to keep them in the program as long as there is any need for further instruction in English? Does the program weight the four English modalities equally or value one (e.g., reading) over the others? Likewise, it is important to document the English language development standards and how they incorporate academic English. Additionally, it is important to document the criteria for selecting students to take the assessment and the characteristics of those who do. Because this assessment is not offered to everyone but is intended for a specific population, it is important to clearly describe the intended and actual population taking the ELPA (see the sidebar for the outline for the first two chapters).

Exhibit 1: Outline for First Two Chapters

I. Introduction
   - A. Description of English Language (EL) Program
   - B. Description of English Language Development/Proficiency (ELD/P) Standards
      - 1. How they were developed
      - 2. How they are disseminated and used
   - C. Context for the Assessment
      - 1. Rationale
      - 2. Purpose and use of results
   - D. General Description of the Assessment

II. Students
   - A. Description of intended student population for the ELPA
   - B. Process for entry into the EL program
      - 1. Home language survey
      - 2. Screener
   - C. Counts and demographics of students in current assessment year*

After the first two chapters, the remainder of the technical information is similar to traditional technical reports. The full outline for this section can be found in Appendix A. The asterisks on various portions of the outline indicate where information should be updated annually.

In many of the chapters, the only differences between an ELPA technical manual and that of a typical assessment come as a result of ELPAs including a speaking section. Speaking assessments are typically given in a one-on-one setting with the teacher both administering and scoring the assessment. This scenario raises additional reliability and validity concerns that must be addressed. As described in sections V and VI of the outline, it is important to document the training provided to teachers who will be administering and scoring the speaking section. In addition, some form of quality control should be implemented and documented. Were teachers monitored? Were two scorers used to calculate the rater reliability? For example, in section V of the outline, the following section is included:

C. Scoring of Speaking Section
    1. Teacher training
    2. Teacher monitoring

In Chapters IX and X, standard setting and reporting, the author can follow a traditional format for documenting these components of the assessment. Care should be taken, however, to address the mechanism for dealing with the four modalities. Was a compensatory approach used, or are students required to achieve a minimum score on each section? Are separate scores reported for each modality? This information should be linked to the purpose of the test to facilitate the aggregation of evidence in the next section on validity.

Finally, the reporting section should be specific to the ELPA. This specificity involves including data elements unique to this assessment. For example, Section XI shown in the text box to the right provides one suggestion for information to include in the reports of operational results.

Exhibit 2: Chapter on Reporting Results

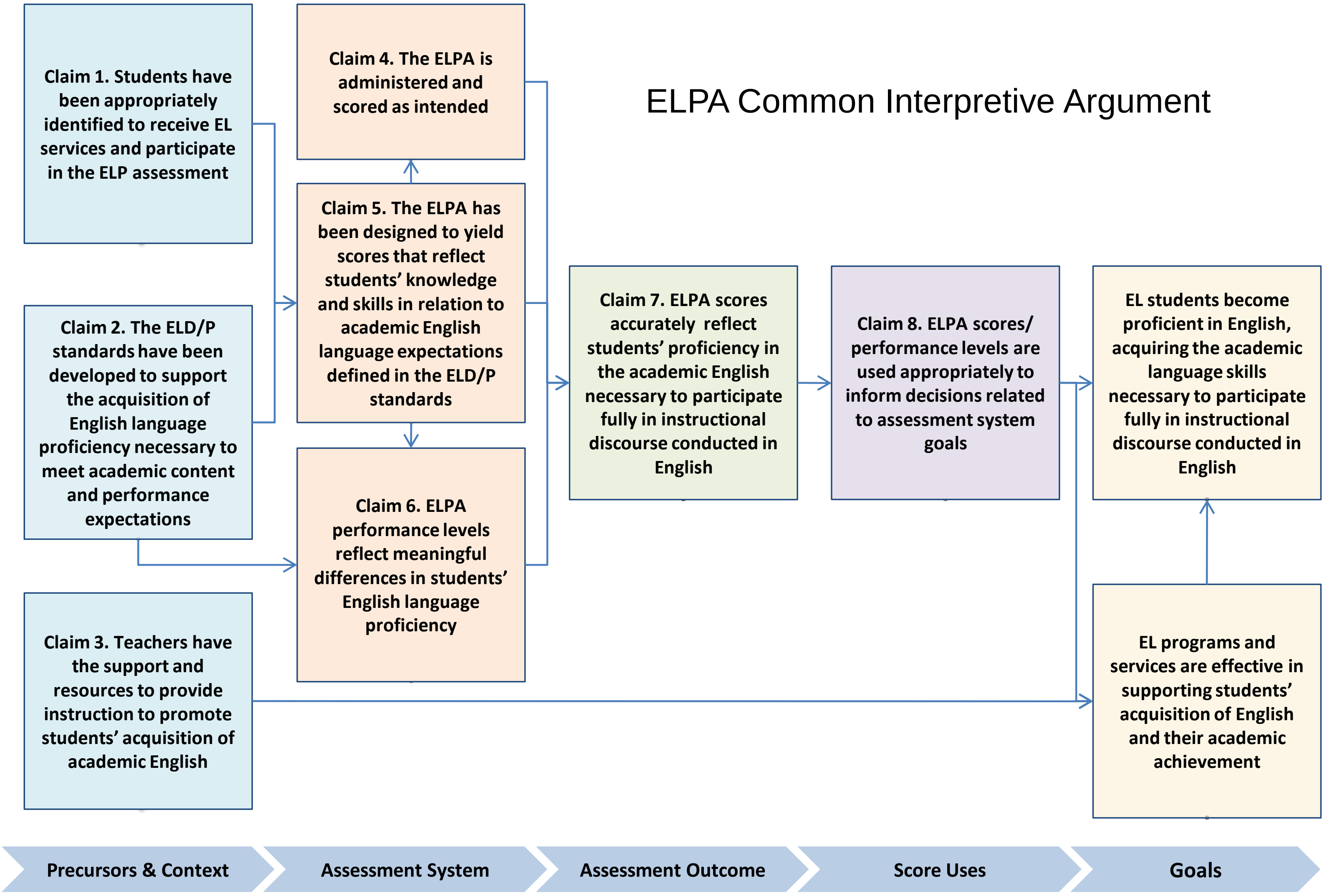
XI. Summary of Operational Results*
    A. Average Composite Score
        1. By Grade and Ethnicity
        2. By Grade and Home Language
    B. Average Score for each Modality
        1. By Grade and Ethnicity
        2. By Grade and Home Language
    C. Percentage Categorized in Each Performance Level
        1. By Grade and Ethnicity
        2. By Grade and Home Language

Validity Information

Once all of the technical information has been documented, it is crucial to weave that information together to “tell a story” as to how that information provides evidence of the validity of the results. When discussing validity, it is important to first articulate the intended uses and interpretations of the test scores and then gather evidence of the validity of using those scores for those purposes.

The previously mentioned work completed documenting validity evidence for the AA-AAS has used the model of a validity argument (see, for example, Kane (2006) or Messick (1989)). In this model, the first step is to document the interpretive argument and then gather and synthesize evidence of the various claims to provide the validity argument. The EVEA project developed a common interpretive argument to be used as a starting point for state or consortia-developed ELPAs. That interpretive argument is shown in Exhibit 3.

Exhibit 3. EVEA Common Interpretive Argument



ELPA Common Interpretive Argument

Claim 1. Students have been appropriately identified to receive EL services and participate in the ELP assessment

Claim 2. The ELD/P standards have been developed to support the acquisition of English language proficiency necessary to meet academic content and performance expectations

Claim 3. Teachers have the support and resources to provide instruction to promote students' acquisition of academic English

Claim 4. The ELPA is administered and scored as intended

Claim 5. The ELPA has been designed to yield scores that reflect students' knowledge and skills in relation to academic English language expectations defined in the ELD/P standards

Claim 6. ELPA performance levels reflect meaningful differences in students' English language proficiency

Claim 7. ELPA scores accurately reflect students' proficiency in the academic English necessary to participate fully in instructional discourse conducted in English

Claim 8. ELPA scores/ performance levels are used appropriately to inform decisions related to assessment system goals

EL students become proficient in English, acquiring the academic language skills necessary to participate fully in instructional discourse conducted in English

EL programs and services are effective in supporting students' acquisition of English and their academic achievement

In the paper written by Marion and Pellegrino (2006), they recommend organizing the validity documentation around the type of evidence collected. That is, after the chapters on the students, content, and interpretive argument, there would be chapters on content-related evidence, evidence based on internal structure, evidence based on response processes, evidence based on relationship to other variables, and consequential evidence. That is one approach, and it certainly has a lot of appeal from a traditional validity point of view. Another approach, which will be described here more fully, is to organize the technical manual around the claims, or groups of claims, in the interpretive argument. The outline in the appendix follows the claims approach, using the claims articulated in the EVEA common interpretive argument shown in Exhibit 3.

If the validity section is incorporated within the full technical manual, then there is no need to repeat the information on the purpose of the test, the students served, and the content covered. Instead, the manual can start with an explanation of the concept of validity and specifically the approach of using a validity argument. The next chapter of the validity section can then describe the interpretive argument specific to that ELPA. Each claim should be outlined and detailed with its underlying assumptions. Then, the remaining chapters can focus on the separate claims. For the EVEA common interpretive argument, the chapters follow the claims as shown in Exhibit 4.

Exhibit 4: Claims from the Interpretive Argument Used to Specify Each Chapter

1. Student identification
2. ELP/D standards
3. Teacher skill/knowledge/orientation
4. Administration and scoring
5. Test design
6. Performance level distinctions
7. Score accuracy
8. Score use at the individual level
9. Score use at the program level

For each claim, evidence can be presented from the first section on technical quality or produced as a result of a special study done specifically to gather validity evidence. For example, the third claim about the ELP/D standards could include evidence from the process of developing the standards, citations that support the process as following best practices, and results from an alignment study between the ELP/D standards and another source, such as grade-level standards or other published standards.

Probably the most complex and scrutinized claim is the claim regarding test design. It includes multiple assumptions about the content and construct. In this chapter, the author could consider organizing the evidence around types of evidence, as outlined in Marion & Pellegrino (2006), or on each assumption. For instance, think-aloud protocols are examples of evidence about response processes. They can be used to test the assumption that the assessment items elicit responses from the student related to the construct being assessed while minimizing construct-irrelevant variance. External alignment studies provide evidence of content validity and address the assumption related to the relationship between the content of the test and the ELP/D standards.

The next set of claims deal with scores and performance levels. The first focus is on their accuracy and reliability, followed by a study of their usage. For ELPAs, there are two main types of score uses: individual level, to determine when a student is ready to exit the EL program, and program level, to evaluate the instructional effectiveness of the EL program. The scores and performance level chapters need to address both uses. In the sample outline, these issues are addressed in separate chapters, focusing first on the reliability and accuracy of the scores and student classifications, and then on their interpretation and use. Other structures are certainly defensible, but only one is provided as an example. See Appendix A for examples of evidence that can be used to support each claim.

The final chapter should be on synthesizing the evidence to draw conclusions about the validity evidence collected. For each assumption, the validity evidence should be weighed, and a judgment should be made about the degree of support the evidence provides for the assumption. Then, by examining the degree to which each assumption is supported, a statement can be made about the degree to which the claims are supported. Claims that appear fully supported become part of the validity argument. Other claims may need to be revised, or additional evidence may be needed. In other cases, the evidence may provide direction on changes needed to the ELPA itself.

The validity evaluation is an ongoing process. Not all validity evidence can be gathered during the design of the assessment. Some must be gathered from operational data, and those specifically related to consequences will need to be collected over time. As with the technical quality section of the documentation, the validity evaluation will need to be updated annually with new evidence and data.

Roles and Responsibilities for Documentation

Because this approach to documenting an assessment includes several components, more than one author will be required. First, the policymaker or operating authority (e.g., a state assessment director or superintendent) should clearly specify the goals and guiding philosophy of the assessment program. Once that has been documented, the vendor responsible for producing the assessment should document the remaining components of Section I. The validity section incorporates an evaluation of the assessment and should be done by someone besides the vendor. Ideally, the policymaker will hire a validity consultant to gather and/or review this evidence and document Section II. However, the policymaker should maintain strong oversight of this section as it is the backbone of the program. If the vendor that designs the assessment writes it, the policymaker will need to conduct a thorough overview acknowledging the vendor bias of showing as much supporting evidence as possible and potentially downplaying any refuting evidence. Because the validity evidence is so important for demonstrating the appropriateness of various interpretations and uses of the scores, it is vital that the policymaker stay involved in these analyses and documentation. Moreover, this document should be available to the public so that they themselves can evaluate the evidence of the validity of the score interpretation and uses.

References

Kane, M. (2006). Validity. In R. L. Linn (Ed.), *Educational Measurement* (4th ed., pp. 13–103). New York: American Council on Education, Macmillan Publishing.

Marion, S. F. & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25* (4),pp. 47-57.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: American Council on Education, Macmillan Publishing.

Appendix A: Table of Contents for a ELPA Technical Manual

This outline provides a suggested table of contents for the technical documentation of an ELPA.

0. Overview of the Report
   - A. Purpose and use of manual
   - B. Intended audience
   - C. Organization of the report

Section 1: Traditional Technical Information

I. Introduction
   - A. Description of English Language (EL) Program
   - B. Description of English Language Development/Proficiency (ELD/P) Standards
     - 1. How they were developed
     - 2. How they are disseminated and used
   - C. Context for the Assessment
     - 1. Rationale
     - 2. Purpose and use of results
   - D. General Description of the Assessment

II. Students
   - A. Description of intended student population for the ELPA
   - B. Process for entry into the EL program
     - 1. Home language survey
     - 2. Screener
   - C. Counts and demographics of students in current assessment year*

III. Test Design and Development
   - A. Overview
   - B. Test Specifications
     - 1. Number of items and points by modality and grade span
   - C. Item Mapping to ELD Standards by Grade Span
     - 1. Alignment study
   - D. Item Development
   - E. Item Review
     - 1. Content Review
     - 2. Bias/Sensitivity Review
   - F. Item Piloting
     - 1. Data Review
     - 2. DIF
   - G. Form Construction
   - H. Determination of allowable accommodations

IV. Administration
   - A. Schedule*
   - B. Process

    1. Whole class and individual portions
    2. Including monitoring of accommodations
  C. Training
    1. General for the assessment as a whole
    2. Specific to the speaking section

V. Scoring
  A. Scoring of Open-ended Items
    1. Description of Rubrics
    2. Scorer Qualifications[*]
    3. Range-finding and Anchor Sets
    4. Training
      a. Process
      b. Materials
    5. Resolution Scoring Rules
    6. Results[*]
      a. Intra-rater agreement
      b. Inter-rater agreement
  B. Scoring of Speaking Section
    1. Teacher training
    2. Teacher monitoring

VI. Reliability[*]
  A. Internal Consistency
  B. Standard Errors of Measurement
    1. Classical
    2. IRT Conditional SEM
  C. Inter-rater Reliability
  D. Reliability of each of the Four Modalities/Dimensions
  E. Decision Consistency and Accuracy

VII. Classical Statistics
  A. Item-Level Statistics
    1. Difficulty (p-value)
    2. Discrimination (bi-serial correlation)
  B. Composite Descriptive Statistics
    1. By Grade and Ethnicity
    2. By Grade and Home Language
  C. Modality-Level Descriptive Statistics
    1. By Grade and Ethnicity
    2. By Grade and Home Language

VIII. Calibration, Scaling and Equating
  A. Description of IRT Model
  B. Calibration
  C. Scale Score Development

D. Equating
E. Raw-to-Scale Score Conversion Tables*
IX. Standard Setting
A. Performance Level Descriptors
B. Standard-Setting Process
C. Scale Score Cut Scores
X. Summary of Operational Results*
A. Average Composite Score
1. By Grade and Ethnicity
2. By Grade and Home Language
B. Average Score for each Modality
1. By Grade and Ethnicity
2. By Grade and Home Language
C. Percentage Categorized in Each Performance Level
1. By Grade and Ethnicity
2. By Grade and Home Language

Section 2: Validity Information

I. Section Introduction
A. Definition of Validity
B. Kane's Argument Approach
II. Define Interpretive Argument for ELPA
A. Claims
B. Assumptions
III. Claim about Student Identification
A. Refer back to Section 1, Chapter II
B. Describe any further studies done/planned to further validate claim*
IV. Claim About ELP/D Standards
A. Describe process for development (refer back to Section 1, Chapter IB1)
B. Reference literature to place process in professional context
C. Describe any alignment studies between the ELP/D standards and subject standards or ELP/D standards from another source*
V. Claim About Teacher Skill/Knowledge/Orientation
A. Results from any studies/surveys/focus groups addressing this claim*
VI. Claim About Administration
A. Refer back to Section 1, Chapter IVB & C to describe training
B. Document percentage of teachers administering the ELPA who received training*
C. Document number of years since training and/or number of training sessions attended*
D. Describe any surveys/evaluations of teachers after training*
1. Immediately after training
2. After initial administration
E. Document details on number of students and number of days of administration.

F. Monitor fidelity of implementation of assessment as intended

  1. Including implementation of approved accommodations

G. Document any studies involving observation of administration[*]

VII. Claim About Test Design

A. Assumption on Content

  1. Test development procedural evidence (refer back to Section 1, Chapter III)
  2. Alignment study

B. Assumption on Construct

  1. Internal structure of ELPA
  2. Evidence of unidimensionality within each modality
  3. Item analyses from Section 1, chapter VII
  4. Description of any other studies done in areas such as think-alouds[*]

VIII. Claim about Performance Level Distinctions

A. Refer back to Section I, Chapter IX for process evidence

B. Describe any further studies done/planned to determine[*]

  1. Classification accuracy
  2. Classification consistency

IX. Claim About Score Accuracy

A. Scorer training (refer to Section 1, Chapters VA4 and VB1)

B. Scorer reliability[*] (refer to Section 1, Chapters VA6 and VIC)

C. Describe any validity studies examining the accuracy of teacher scoring[*]

X. Claim About Score Use

A. Describe studies conducted on[*]

  1. Score interpretation
  2. Use to determine student readiness to participate in English-based instruction
  3. Use to evaluate program
  4. Other uses

XI. Claim about Students Progressing Through EL Program

A. Exit rates & years between entry and exit[*]

  1. By Grade
  2. By Home Language
  3. By Program

B. Describe any further studies on program effectiveness[*]

XII. Synthesize Evidence[*]

A. Describe claims that are supported

B. Describe claims that need further research

C. Present validity argument that is backed by evidence

[*] Section should be updated every year or when new data are available.