# *Evaluating the Validity of a High-Stakes ESL Test: Why Teachers' Perceptions Matter*

**PAULA WINKE**
*Michigan State University*
*East Lansing, Michigan, United States*

The English Language Proficiency Assessment (ELPA) is used in the state of Michigan in the United States to fulfill government-mandated No Child Left Behind (NCLB) requirements. The test is used to promote and monitor achievement in English language learning in schools that receive federal NCLB funding. The goal of the project discussed here was to evaluate the perceived effectiveness of the ELPA and to see if those perceptions could meaningfully contribute to a broad concept of the test's validity. This was done by asking teachers and test administrators their views on the ELPA directly after its administration. Two hundred and sixty-seven administrators took a survey with closed and open-ended questions that aimed to tap into the consequential dimensions of test validity. An exploratory factor analysis identified five factors relating to the participants' perceptions of the ELPA. Analysis of variance results revealed that educators at schools with lower concentrations of English language learners reported significantly more problems in administering the ELPA. Three themes (the test's appropriateness, impacts, and administration) emerged from an analysis of qualitative data. This article discusses these results not only as a means of better understanding the ELPA, but also to contribute to larger-scale discussions about consequential validity and standardized tests of English language proficiency. It recommends that broadly defined validity data be used to improve large-scale assessment programs such as those mandated by NCLB.
*doi: 10.5054/tq.2011.268063*

The study discussed in this article explored the concept of test validity (the overall quality and acceptability of a test; Chapelle, 1999) and how teachers' expert judgments and opinions can be viewed as part of a test's validity argument. It did so in the context of a statewide battery of tests administered in Michigan, a large Midwestern state in the United States, to English language learners (ELLs) from kindergarten through 12th grade. Data were gathered in the weeks after the testing by

surveying educators who had been involved in administering the tests. Using these data, in this article I examine the validity of the tests by analyzing what the teachers believed the tests measure and what they believed the tests' impacts are. In essence, this paper investigates the social consequences of a large-scale testing program and scrutinizes "not only the intended outcome but also the unintended side effects" of the English language tests (Messick, 1989, p. 16). Because the testing in Michigan was required by the federal government's No Child Left Behind (NCLB) Act, I begin by summarizing that law.

## THE NO CHILD LEFT BEHIND ACT AND LANGUAGE POLICY TESTS

Although the United States has no official national language policy (Crawford, 2000), NCLB is sometimes viewed as an ad hoc federal language policy that promotes an English-only approach to education (Evans & Hornberger, 2005; Menken, 2008; Wiley & Wright, 2004). NCLB has created stringent education requirements for schools and states. Title I of NCLB,[1] which provides federal funding to schools with low-income students, requires those schools to meet state-established Annual Yearly Progress (AYP) goals and to achieve 100% proficiency relative to those goals by 2014. (For details about how states define proficiency, see Choi, Seltzer, Herman, & Yamashiro, 2007; Lane, 2004; Porter, Linn, & Trimble, 2005.) If a school fails to meet the AYP goals, it suffers increasingly serious consequences, eventually including state takeover.

The law has special requirements for seven identified subgroups: African Americans, Latinos, Asian/Pacific Islanders, American Indians, students with low socioeconomic status, special education students, and ELLs. Before NCLB, achievement gaps between subgroups and the overall student population were often overlooked (Lazarín, 2006; Stansfield & Rivera, 2001). The law addresses this problem in three ways. First, there are special testing requirements for the subgroups. Now 95% of students within each subgroup—including ELLs who have been in the United States less than 1 year—must be tested for a school or district to meet its AYP (U.S. Department of Education, 2004b). Second, since each subgroup must achieve the same school and statewide AYP goals that apply to the general population, ELLs must also meet English language proficiency benchmarks through additional tests. Finally, schools must report the scores of these subgroups separately, enabling stakeholders such as funders, test developers, teachers, test takers, and parents (McNamara, 2000) to hold the educational system accountable for discrepancies in scores.

[1] The word *Title* refers to a major portion of a law. NCLB has 10 Titles. See the NCLB Table of Contents, which lists the Titles and their subparts, and has the text of the statute (U.S. Department of Education, 2002).

Title III of NCLB allocates federal education funding to states based on the state's share of what the law calls Limited English Proficient and recent immigrant students. This article refers to those students as ELLs, because the statutory term Limited English Proficient is controversial (e.g., Wiley & Wright, 2004, p. 154). Public schools in the United States have more than 5.5 million ELLs; 80% speak Spanish as their first language, with more than 400 different languages represented overall (U.S. Department of Education, 2004a). To continue to receive Title III funding, states must demonstrate that they are achieving two Annual Measureable Achievement Objectives: an increase in the number or percentage of ELLs making progress in learning English, as measured by state-issued tests of English language proficiency, and an increase in the number or percentage of ELLs obtaining state-defined levels of English proficiency as demonstrated, normally, on the same state tests. Therefore, the failure of ELLs to meet Title I AYP requirements (achieved through 95% of ELLs being tested in reading, math, and science and, if the ELLs have been in the U.S. for more than a year, acceptable performance on the tests) will jeopardize schools' Title I (low-income) funding. Additionally, failure of ELLs to demonstrate gains in English on the English language proficiency tests will jeopardize the school's Title I and state's Title III (ELL program) funding. Thus, the consequences attached to this testing are severe. These overlapping NCLB policies regarding ELLs are diagramed in Figure 1.

NCLB thus requires that ELLs take several standardized tests every year, regardless of their preparation for or ability to do well on the tests. The tests have been criticized as constituting a de facto language policy, because they stress the importance of English over all other languages (Menken, 2008) and have been used to justify bypassing transitional bilingual education and focusing solely on English language instruction (Shohamy, 2001; Wiley & Wright, 2004).

Whether they support the tests, NCLB policy makers, second language acquisition researchers, and test designers agree on at least one principle: The tests created for NCLB purposes must be reliable and valid. Those two terms are used throughout the text of the NCLB Act, but are never defined. It is up to states to establish reliable and valid tests, and, ultimately, to design accountability systems adhering to the law that are both reliable and valid (Hill & DePascale, 2003). The next few paragraphs address those key terms.

## PERSPECTIVES ON TEST VALIDITY AND RELIABILITY

Validity encompasses several related concepts (see Chapelle, 1999). To be valid, a test needs *reliability* (Bachman, 1990; Chapelle, 1999,

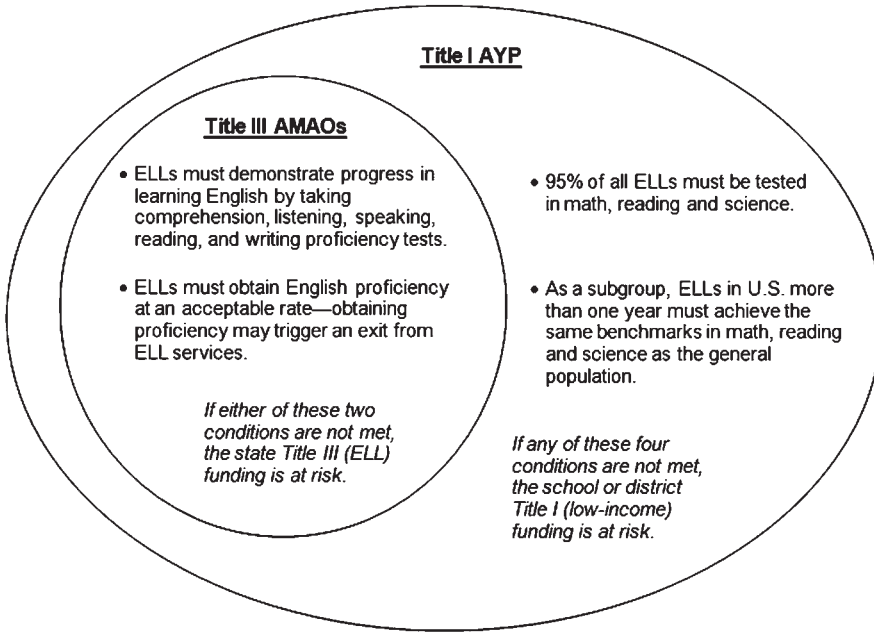


**FIGURE 1. Annual Measureable Achievement Objectives (AMAOs) under Title I and Title III of No Child Left Behind that affect English language learners (ELLs). AYP = Annual Yearly Progress.**

p. 258; Lado, 1961). In other words, "reliability is a requirement for validity, and . . . the investigation of reliability and validity can be viewed as complementary aspects of identifying, estimating, and interpreting different sources of variance in test scores" (Bachman, 1990, p. 239). In simple terms, a test's reliability estimate tells users of a language test how typical (generalizable) the students' test scores are, whereas an evaluation of a test's validity will tell them whether it is appropriate to use the scores from the test to make particular decisions (see Figure 2.7 in McNamara & Roever, 2006). Reliability and validity can be viewed as on a continuum (Bachman, 1990), or, as I like to think of them, as both within the sphere of a broad concept of validity with reliability at the core and consequential validity at the outer layer (see Figure 2).

Reliability is the core component of validity, and no test can be valid if it is not reliable. However, a test can be reliable and not valid. This is because these aspects of the test validation process are different and they are measured differently. A test is reliable if it will consistently yield the same scores for any one test taker regardless of the test examiner and of the time of testing (Bachman, 1990; Chapelle, 1999; Messick, 1989). Reliability can be assessed by sampling the content of the test, testing individual students more than once and measuring differences, and by

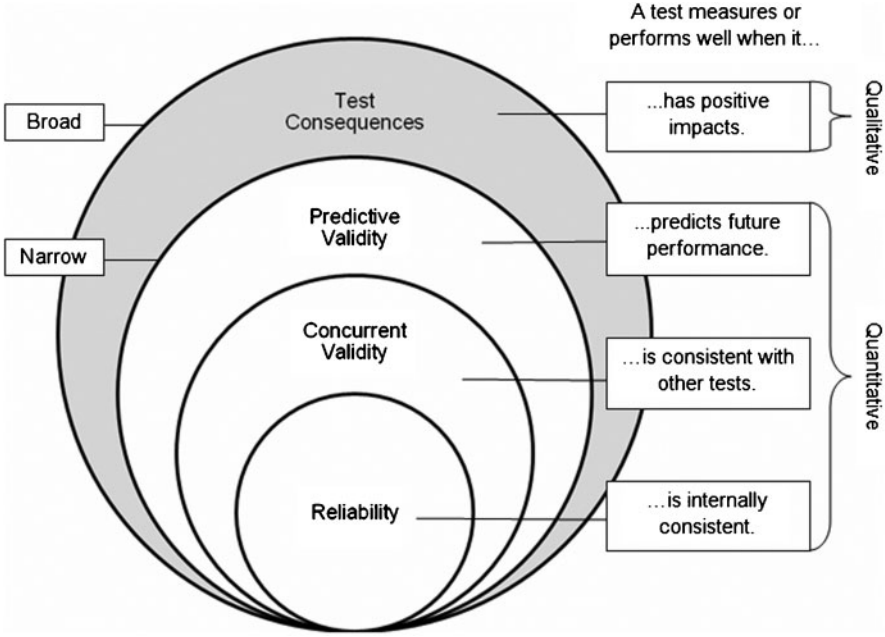


**FIGURE 2. Levels of validity evidence.**

comparing scores assigned by different raters (Brown, 2005). Estimating a test's reliability is part of validating a test—in fact it is often considered a prerequisite for validation (Bachman, 1990). However, a reliable test is not necessarily valid. For example, a speaking test comprised of multiple-choice, discourse-completion tasks in which test takers select correct or appropriate responses to recorded questions may have high reliability, but may not be a valid test of speaking ability because scores from it may not accurately represent the test takers' true, real-world speaking abilities.

In addition to reliability, a valid test requires *concurrent validity*, meaning that the test is consistent with other tests that measure the same skills or knowledge (Chapelle, 1999). Another important trait of a valid test is *predictive validity*, meaning that a test predicts later performance or skill development (Chapelle, 1999). Reliability, concurrent validity, and predictive validity can all be measured quantitatively. However, these purely statistical conceptions of validity are rather narrow (see Norris, 2008, p. 39, for a description of the "narrow-vein" of statistical validity evidence). Tests should be more than just statistically valid (Messick, 1980, 1989, 1994, 1996). They should be fair, meaningful, and cost efficient (Linn, Baker, & Dunbar, 1991). They should be developmentally appropriate (Messick, 1994). They must be able to be administered

successfully (Hughes, 2003). More broadly construed, then, validity includes test consequences (Bachman, 1990; Messick, 1989). Tests affect students, the curriculum, and the educational system as a whole (Crooks, 1988; Moss, 1998; Shohamy, 2000). They can be "engines of reform and accountability in education" (Kane, 2002, p. 33). *Consequential validity* thus encompasses ethical, social, and practical considerations (Canale, 1987; Hughes, 2003; McNamara & Roever, 2006). This article uses the term *broad validity* to refer collectively to reliability, concurrent validity, predictive validity, and consequential validity. An extensive or broad validation process will not only provide evidence that a test's score interpretations and uses of the scores derived from the test are good, but it will also investigate the ethics of the test and the consequential basis of the test's use.

Stakeholders' judgments about a test are an important tool for determining its consequential validity (Chapelle, 1999; Crocker, 2002; Haertel, 2002; Kane, 2002; Shohamy, 2000, 2001, 2006). Teachers and school administrators are certainly stakeholders, especially when assessment programs are designed primarily to improve the educational system (Lane & Stone, 2002). Students' performance on tests can affect teachers' and administrators' reputations, funding, and careers. Teachers and school administrators, moreover, have unique insight into the collateral effects of tests. They administer tests, know their students and can see how the testing affects them, and they recognize—sometimes even decide—how the tests affect what is taught. Because they have personal knowledge of the students and come to directly understand how testing affects them in their day-to-day lives, teachers are well positioned to recognize discrepancies between classroom and test practices. They have a unique vantage point from which to gauge the effects of testing on students (Norris, 2008). The teachers' perspectives are therefore valuable pieces of information concerning whether tests affect the curriculum as intended. In this way, the teachers can shed light on the validity of the tests, that is, whether the tests measure what they are supposed to and are justified in terms of their outcomes, uses, and consequences (Bachman, 1990; Hughes, 2003; Messick, 1989).

It is surprising, then, that most statewide assessment programs in the United States do not, as part of the annual review of the validity of their tests, anonymously survey teachers about test content or administration procedures. The teachers and administrators are, after all, expert judges who can inform the content-related validity of a test. According to Chapelle (1999), "accepted practices of test validation are critical to decisions about what constitutes a good language test for a particular situation" (p. 254). Researchers have found that teachers and school administrators normally do not contribute meaningfully to the test validation process unless the test managers have a plan for collecting

and using information from them (Crocker, 2002), even though it is clear that including the perspectives of teachers and school administrators in the assessment validation process can improve the validity of high-stakes assessments (Ryan, 2002). Although panels of teachers and technical experts are often employed during policy development, standards drafting, and test creation, in practice their opportunity to express their full opinions after a test becomes fully operational does not exist as part of the validation process (Haertel, 2002).

Investigations into the validity of the Title I (reading, math, science) tests for ELLs have been the subject of considerable debate (Evans & Hornberger, 2005; Menken, 2008; Stansfield, Bowles, & Rivera, 2007; Wallis & Steptoe, 2007; Zehr, 2006, 2007). However, little or no research has been conducted on the validity of the English language proficiency tests mandated by Title I and Title III. These tests may be less noticed because they are less controversial and only indirectly impact funding received by the larger population under Title I. Nonetheless, these tests are being administered to over 5.5 million students (U.S. Department of Education, 2004a) and can impact funding for ELL and Title I programs. Therefore, an investigation into their broad validity is warranted.

## CONTEXT OF THE STUDY: ENGLISH LANGUAGE PROFICIENCY TESTING IN MICHIGAN

This study investigated the views of teachers and school administrators (collectively called *educators* in this article) on the administration of English language proficiency tests in Michigan, United States. Michigan's English Language Proficiency Assessment (ELPA) has been administered to students in kindergarten through 12th grade annually since the spring of 2006, as a part of Michigan's fulfillment of NCLB Title I and Title III requirements.

The ELPA is used in Michigan to monitor the English language learning progress of all kindergarten through 12th grade students eligible for English language instruction, regardless of whether they are currently receiving it (Roberts & Manley, 2007). The main goal of the ELPA, according to the Michigan Department of Education, is to "determine—on an annual basis—the progress that students who are eligible for English Language Learner (ELL) services are making in the acquisition of English language skills" (Roberts & Manley, 2007, p. 8). A secondary goal of the Michigan ELPA is to forge an overall improvement in test scores over time for individuals, school districts, and/or subgroups and to spur English language education to more closely align with state proficiency standards. Additionally, the ELPA is viewed

by the state as a diagnostic tool for measuring proficiency, revealing which schools or districts need more attention.

As required by NCLB, the test is based on English language proficiency standards adopted by the state and includes subtests of listening, reading, writing, speaking, and comprehension.[2] Each subtest is scored based on three federal levels of performance: basic, intermediate, and proficient. In the spring of 2007, there were five levels (or forms) of the test:

- Level I for kindergarteners
- Level II for grades 1 and 2
- Level III for grades 3 through 5
- Level IV for grades 6 through 8
- Level V for grades 9 through 12

The 2006 Michigan ELPA (MI-ELPA) Technical Manual reported on the validity of the ELPA (Harcourt Assessment, 2006, pp. 31–32). In the manual, Harcourt asserted that the ELPA was valid because (a) the item writers were trained, (b) the items and test blueprints were reviewed by content experts, (c) item discrimination indices were calculated, and (d) item response theory was used to measure item fit and correlations among items and test sections. However, the validity argument the manual presented did not consider the test's consequences, fairness, meaningfulness, or cost and efficiency, all of which are part of a test's validation criteria (Linn et al., 1991). In 2007 teachers or administrators could submit their opinions about the ELPA during its annual review, but they had to provide their names and contact information. As explained by Dörnyei (2003), surveys and questionnaires that are anonymous are more likely to elicit answers that are less self-protective and more accurate. Respondents who believe that they can be identified may be hesitant to respond truthfully (Kearney, Hopkins, Mauss, & Weisheit, 1984). Therefore, it is not certain if the teachers' opportunities to submit opinions on the ELPA could be considered a reliable way of obtaining information on the broad validity of the ELPA, if the opinions are not as truthful as they could have been out of fear of monitoring or censuring on the part of the state.

The aim of the study described here was to understand how educators can shed light on a test's consequential validity. More locally, the goal was to fill the gap in the analysis of the ELPA's validity. Two research questions were therefore formulated:

[2] NCLB requires subtests of listening, reading, writing, speaking, and comprehension, but in practice most states only present students with subtests of listening, speaking, reading, and writing, but report a separate *comprehension* score to be compliant with the law. In Michigan, select items from the listening and reading sections are culled to construct a separate comprehension score.

(1) What are educators' opinions about the ELPA?
(2) Do educators' opinions vary according to the demographic or teaching environment in which the ELPA was administered?

The null hypothesis related to the second question was that educators' opinions would not vary according to any demographics or teaching environments in which the ELPA was administered.

## METHOD

### Participants

Two hundred and sixty-seven teachers, principals, and school administrators (henceforth, educators) participated in this study. Of these, 159 classified themselves as English as a second language (ESL) or language arts, that is, mainstream English teachers (many stated that they taught more than one subject or identified themselves as both ESL and language arts teachers). Five of these reported that they primarily taught English or other subjects besides ESL or language arts. Sixty-nine identified themselves as school administrators ($n$ = 27), school principals ($n$ = 21), or a specific type of school administrator ($n$ = 21), such as a school curriculum director, curriculum consultant, testing coordinator, or test administrator. Thirty-nine explained that they were ESL or ELL tutors, special education teachers, Title I teachers, ESL teachers on leave or in retirement who came in to assist with testing, or literacy assistants or coaches.

### Materials

The data for the present study were collected using a three-part, online survey with items that aimed to investigate the social, ethical, and consequential dimensions of the ELPA's validity. The survey was piloted on a sample of 12 in-service teachers and two external testing experts, after which the survey was fine-tuned by changing the wording of several items and deleting or collapsing some items. The final survey included six discrete-point items that collected demographic information (Appendix A) and 40 belief statements (which can be obtained by emailing the author; see Appendix B for a subset of the statements) that asked the educators their opinions about the following aspects of the ELPA's validity: (a) the effectiveness of the administration of the ELPA (items 1–7); (b) the impacts the ELPA has on the curriculum and stakeholders (items 8–19); (c) the appropriateness of the ELPA subsections of listening, reading, writing, and speaking (items 20–35);

and (d) the overall validity of the instrument (items 36–40). These questions were asked because a test's broad validity is related to the test's successful administration (Hughes, 2003), impacts (Bachman, 1990; Messick, 1989; Shohamy, 2000), and appropriateness (Messick, 1989). For each belief statement, the educators were asked to mark on a continuous, 10-point scale how much they agreed or disagreed with the statement. Each statement was followed by a text box in which educators could type comments. Five open-ended questions were presented at the end of the survey (Appendix C).

## Procedure

The survey was conducted during and after the spring 2007 ELPA testing window, which was March 19 to April 27, 2007. Educators were first contacted through an email sent through the Michigan Teachers of English to Speakers of Other Languages (MITESOL) listserv on March 29, 2007. The email explained the purpose of the survey and asked the educators to take the survey as soon as possible after administering the ELPA. The Michigan Department of Education's Office of Educational Assessment and Accountability declined a request to distribute a similar email through the official state listserv for the administrators of the ELPA. Therefore, additional names and email addresses were culled from online databases and lists of Michigan school teachers, principals, and administrators. The completed list contained 2,508 educators, who were emailed on April 8, 2007. Five hundred and eighty-five emails bounced back. One hundred and fifty-six educators responded that they either were not involved in the ELPA or that they would forward the message on to others who they believed were. On May 14, reminder emails were sent through the MITESOL listserv and to the appropriately truncated email list. Two hundred and sixty-seven educators completed the online survey between March 29 and May 20, 2007. The survey took an average of 16 minutes and 40 seconds to complete.

## Analysis

The data for this study consisted of two types, the quantitative data from the Likert-scale items (the belief statements) and the qualitative data from the comment boxes attached to the belief statements. Because the research involved understanding how educators can shed light on a test's consequential validity, the goal was to analyze the quantitative data for general response patterns to the Likert-scale items, and then to illustrate, through the qualitative comments, the educators' opinions about the ELPA.

The quantitative, Likert-scale data were coded on the 10-point scale from −4.5 (strongly disagree) to +4.5 (strongly agree) with 0.5 increments in between; neutral responses were coded as zeros. All data were entered into SPSS 18.0. Negatively worded items (12, 16, 19, 25, and 36) were recoded positively before analysis. An exploratory factor analysis was conducted because there has been to date no empirical research identifying exactly what, and how many, factors underlie the broad concept of test validity. The factor analysis on the questionnaire data was meant to filter out items in the survey that were unrelated to the construct of consequential test validity (Field, 2009). In other words, the factor analysis was used to "reduce the number of variables to a few values that still contain most of the information found in the original variables" (Dornyei, 2003, p. 108). One-way analyses of variance (ANOVA) were used to detect differences in the factors among the educator subgroups.

The qualitative data were analyzed through an inductive approach in which themes and patterns emerged from the data. After all data were entered into NVivo 8, I read the data segments (a segment is a single response by an educator to one question on the survey) and first coded them as (a) being positive or negative in tone, (b) referring to a specific grade-level (kindergarten, first through second grade, etc.), or (c) referring to a specific skill area (listening, speaking, reading, or writing). I then reread the entire corpus and compiled a list of general themes, which are presented in their final form in Table 1. Once I had identified the initial themes, I and another researcher then read through approximately 10% of the corpus's data segments and coded them. (Ten percent was chosen because previous research with large qualitative data sets has established inter-rater reliability on 10% of the data—see, for example, Chandler, 2003.) The agreement level was 82%. Differences in opinion were resolved through discussion. The list of general themes was refined throughout this initial coding process by grouping related themes and then by renaming combined categories. As a final step, I completed the coding of the rest of the data segments with consultation from the second researcher.

## RESULTS

### Quantitative (Likert-Scale) Results

The 267 educators who responded to the survey were allowed to skip any item that they did not want to answer or that did not pertain to them. Accordingly, there are missing data. Cronbach's alpha for the 40 Likert-scale items was 0.94 when the estimate included a listwise deletion of all educators who did not answer any one item (134 educators

**TABLE 1**
**Coding Categories and Themes That Emerged From the Data**

| Major category | Subtheme |
|---|---|
| 1. Tone | a. Positive |
| | b. Negative |
| 2. Grade level | a. Kindergarten |
| | b. 1st–2nd grade |
| | c. 3rd–5th grade |
| | d. 6th–8th grade |
| | e. 9th–12th grade |
| 3. Skill tested | a. Listening |
| | b. Speaking |
| | c. Reading |
| | d. Writing |
| 4. Appropriateness | a. Cognitive |
| | b. Content relativeness |
| | c. Difficulty level |
| | d. Length |
| | e. For specific student populations |
| 5. Impact | a. Available resources |
| | b. Instruction |
| | c. Students' psyche |
| 6. Logistics | a. Amount of time |
| | b. Conflict with other tests |
| | c. Educator training |
| | d. Personnel |
| | e. Physical space |
| | f. Scoring process |
| | g. Test materials |

included, 133 excluded). When all missing values were replaced with the series mean, which allowed for the alpha coefficient to be based on all obtained data, Cronbach's alpha was 0.95. Either way, the high reliability estimate indicated that the data from the instrument were suitable for a factor analysis (Field, 2009).

The exploratory factor analysis[3] of the data from 40 Likert-scale items resulted in a clear five-factor solution. The five factors explain 72% of the variance found in the analysis. Table 2 reports the Eigenvalues and the total variance explained by each factor. Factor 1 items were related to the validity of the reading and writing portions of the test. Factor 2 items concern the effective administration of the test. Factor 3 items concern the test's impacts on the curriculum and students. Factor 4

[3] A maximum likelihood extraction method was applied. Because the factors were assumed to be intercorrelated, a subsequent oblique (Promax) rotation was used. After eliminating all items with communalities less than 0.4, the number of factors to be extracted was determined by the Kaiser criterion; only factors having an Eigenvalue greater than 1 were retained.

[4] Note that in Appendix B, because the factors are correlated (oblique), the factor loadings are regression coefficients, not correlations, and therefore can be larger than 1 in magnitude, as one factor score in the table is. See Jöreskog (1999).

**TABLE 2**
**Eigenvalues and Total Variance Explained by Factor**

| Factor | Initial Eigenvalues | | |
|---|---|---|---|
| | Total | % Variance | Cumulative % |
| 1. Reading and writing tests | 11.43 | 45.70 | 45.70 |
| 2. Effective administration | 2.43 | 9.73 | 55.43 |
| 3. Impacts on curriculum and learning | 1.52 | 6.08 | 61.51 |
| 4. Speaking test | 1.39 | 5.57 | 67.08 |
| 5. Listening test | 1.18 | 4.73 | 71.81 |

items concern the speaking portion of the test, whereas Factor 5 items concern the listening portion of the test. Appendix B presents the pattern matrix of the five-factor solution with loadings less than 0.5 suppressed.[4]

The average response rates for each factor are listed in Figure 3. On average, educators disagreed more with the statements from Factor 2 (effective administration) than any other factor. The average response to the four questions that make up Factor 2 was −1.811, where 4.5 is strongly agree and −4.5 is strongly disagree (zero is a neutral response). Also receiving negative averages were the statements from Factor 1 (reading and writing sections), Factor 4 (the speaking section), and

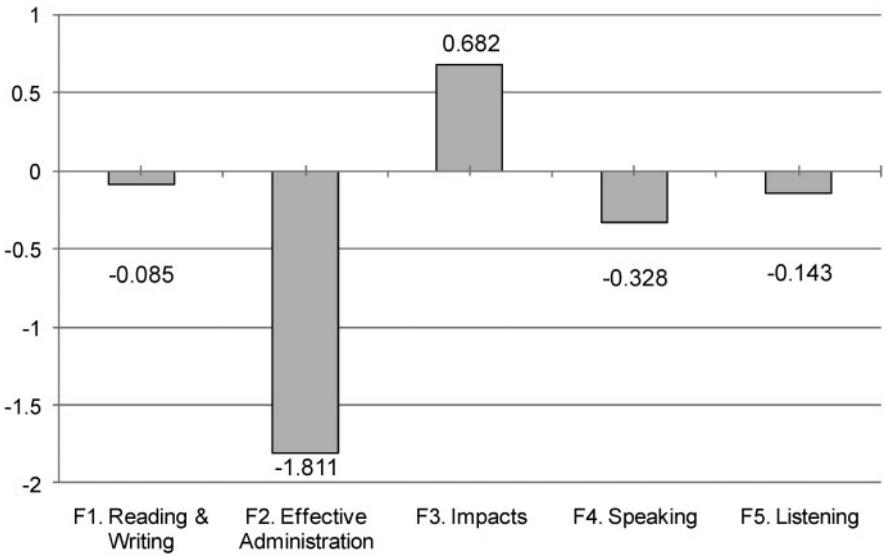


**FIGURE 3. Descriptive statistics of the average responses on the five factors (F).**

Factor 5 (the listening section). Receiving an overall positive score was Factor 3 (impacts on the curriculum and learning). To sum, the survey consisted of questions that clustered around five major issues, and the educators' overall opinions regarding these five issues varied. As a group, the educators were apprehensive about how effective the exam's administration was. They were, as a whole, slightly troubled about aspects of the different sections of the exam itself. But, generally, they were pleased with how the exam impacted certain aspects of the curriculum and the students' English language learning. These results are discussed in detail later in the discussion section.

One-way ANOVA was used to compare the mean factor scores (the means of the raw scores that loaded on the individual factors) by (a) the level of testing the educator administered (kindergarten through 2nd grade, grades 3 through 5, or grades 6 through 12) or (b) the school's concentration of ELLs (less than 5%, between 5 and 25%, and more than 25%). No significant differences in mean scores were found among the levels of testing. Significant differences in mean scores were found among the three subgroups of ELL concentration on Factor 2 (effective administration), $F$ (2,253) = 5.739, $p$ = 0.004, and Factor 4 (the speaking test), $F$ (2,253) = 3.319, $p$ = 0.038, but not on any of the other factors. What this means is that although, as a whole, the educators expressed (a) disfavorable opinions about the effectiveness of the exam's administration and (b) a slight unease concerning issues with the speaking test (see Figure 3), when the educators are divided into three subgroups according to the percentage of ELLs at their schools, there are significant differences in their opinions concerning these two issues. Table 3 provides the descriptive statistics for the three levels of ELL concentration on Factors 2 (effective administration) and 4 (the speaking test).

Post hoc Tukey analyses were examined to see which pairs of means were significantly different. In other words, post hoc analyses were conducted to reveal which of the three educator subgroups (grouped by the percentage of ELLs at their schools) differed in their opinions on

**TABLE 3**
**Descriptives for One-way ANOVAs Regarding ELL Concentration and Factors 2 and 4**

| ELL Concentration | Factor 2 (Effective administration) | | | Factor 4 (Speaking test) | | |
|---|---|---|---|---|---|---|
| | *N* | *M* (*SD*) | 95% CI | *N* | *M* (*SD*) | 95% CI |
| <5% | 80 | −2.14 (2.3) | [−2.65, −1.63] | 80 | −0.71 (2.18) | [−1.20, −0.23] |
| 5–25% | 103 | −1.99 (2.3) | [−2.44, −1.54] | 103 | −0.27 (2.24) | [−0.70, 0.17] |
| >25% | 73 | −0.97 (2.43) | [−1.53, −0.4] | 73 | 0.19 (2.05) | [−0.29, 0.67] |

*Note.* ANOVA = analysis of variance; ELL = English language learners; *N* = number; *M* = mean; *SD* = standard deviation or error; CI = confidence interval.

**TABLE 4**
**Post Hoc Tukey Results for ELL Concentration and Factors 2 and 4**

| Dependent Variable | ELL Concentration A | ELL Concentration B | Mean difference (A − B) | *P* | 95% CI |
|---|---|---|---|---|---|
| Factor 2 | <5% | 5–25% | −0.15 | 0.900 | [−0.97, 0.67] |
| (Effective | <5% | >25% | −1.17 | 0.006* | [−2.07, −0.28] |
| administration) | 5–25% | >25% | −1.02 | 0.013* | [−1.86, −0.18] |
| Factor 4 | <5% | 5–25% | −0.44 | 0.356 | [−1.21, 0.32] |
| (Speaking | <5% | >25% | −0.90 | 0.028* | [−1.73, −0.08] |
| test) | 5–25% | >25% | −0.46 | 0.349 | [−1.24, 0.32] |

*Note.* *The mean difference is significant at the 0.05 level.

these two issues (the effectiveness of the ELPA's administration and the speaking test). The standard errors and the confidence intervals for the Tukey post hoc analyses are displayed in Table 4.

Tukey post hoc comparisons of the three ELL concentration subgroups revealed no significant difference in opinion concerning the ELPA's administration (Factor 2) among schools with ELL populations less than 5% ($M = -2.14$) and between 5 and 25% ($M = -1.99$); however, both of these subgroups' means on the effective administration of the ELPA (Factor 2) were significantly more negative than the mean from educators at schools with more than 25% ELLs ($M = -0.97$), $p = 0.006$ and $p = 0.013$, respectively. In more general terms, what this means is that educators in schools with a lower concentration of ELLs tended to have a more negative view of test administration than educators in schools with a higher concentration. These differences can be seen in Figure 4.

Regarding Factor 4, opinions related to the speaking section, Tukey post hoc comparisons demonstrated that educators at schools with very low concentrations of ELLs (less than 5%; $M = -0.71$) had significantly more negative opinions concerning the speaking section of the ELPA (Factor 4) than did those at schools with high concentrations of ELLs (more than 25%; $M = 0.19$), $p = 0.028$. Comparisons between the mid-ELL-concentration subgroup and the other two subgroups were not statistically significant at $p < 0.05$. In other words, educators in schools with lower concentrations of ELLs were apt to have a more negative view of the speaking test than those in schools with higher ELL concentrations. These differences are illustrated in Figure 5.

In sum, the quantitative data relevant to Research Question 1 (What are educators' opinions about the ELPA?) show that, on average, educators were critical about the administration of the test, positive about the impacts of the test on the curriculum and learning, and slightly negative about the subtests. The quantitative data relevant to Research Question 2

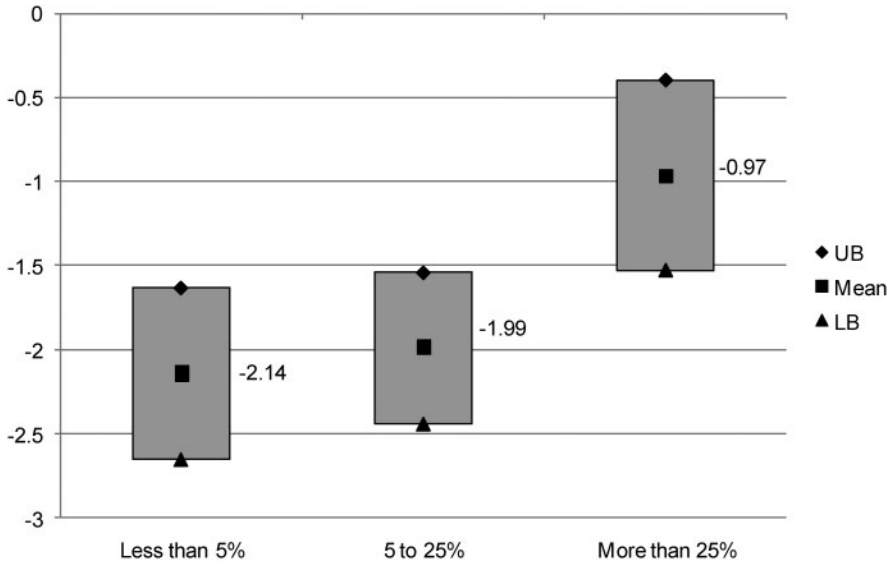


**FIGURE 4. Analysis of variance results of Factor 2 (effective administration) by English language learner concentration, with upper bound (UB) and lower bound (LB) confidence intervals.**

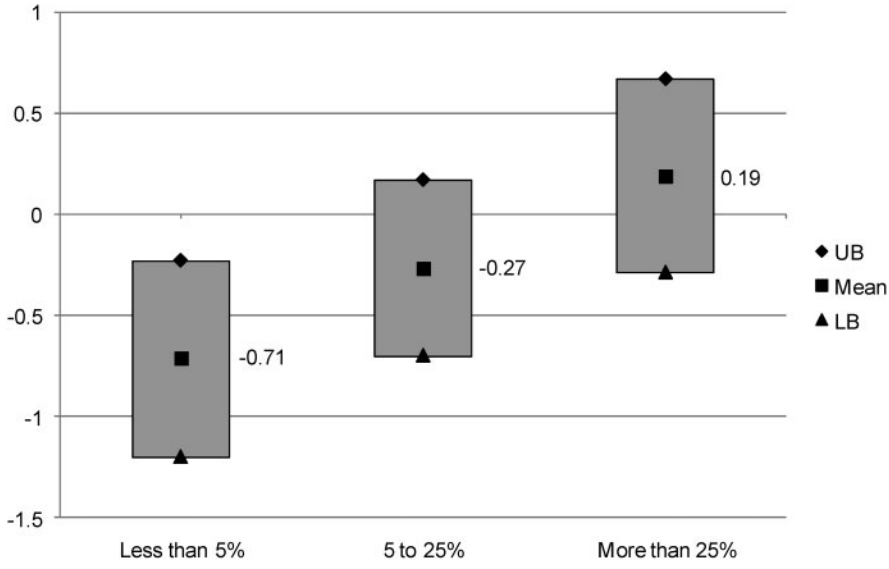


**FIGURE 5. Analysis of variance results of Factor 4 (speaking test) by English language learner concentration, with upper bound (UB) and lower bound (LB) confidence intervals.**

(Do educators' opinions vary according to the demographic or teaching environment in which the ELPA was administered?) reveal that the opinions do vary, rejecting the null hypothesis. Educators from schools with lower concentrations of ELLs had significantly more negative opinions about the administration of the exam than educators at schools with higher concentrations of ELLs. Those from schools with low ELL concentrations also expressed significantly lower opinions about the speaking portion of the exam. However, regarding the other issues (opinions about the reading, writing, and listening sections of the exam, and opinions concerning the impacts of the exam), opinions did not vary by demographic or teaching environment subgroups.

## Qualitative (Free-Response) Results

The educators' responses to the open-ended items provide valuable insight into the ELPA's broad validity. The educators' impressions centered on three themes: (1) the appropriateness of the exam, (2) the exam's impacts, and (3) the logistics of administering the exam. Table 5 indicates the number of educators who commented positively or negatively on these themes.

### 1. The Appropriateness of the Exam

Educators wrote mostly negative comments about the exam's appropriateness, opining that it was too difficult and too long. Their concerns were most pronounced for three groups of ELLs: young language learners (kindergarten and first graders), special education ELLs, and very recent arrivals. There were 11 positive comments about the appropriateness of the exam, but 10 of these were coupled with a negative comment. For example, one teacher wrote: "Overall, it is good for keeping districts accountable but the test needs some rewriting, needs to be shorter and teachers need help with administration" (Educator 182, an ESL teacher who administered Level II and III tests to 1st through 5th graders).

#### *The difficulty level of the exam*

Some educators explained that the exam was difficult because the content was unrelated to the underlying skills being assessed. Many comments emphasized the difficulty for young ELLs. They pointed out that young children are not used to reading silently, a test requirement. Educators commented that young ELLs were also asked to use transition words and to discuss speakers' opinions, and to perform other tasks that were new and even incomprehensible to them. Educators reported that

**TABLE 5**
**Summary Data of Qualitative Responses**

| | | No. of educators | | No. of comments | |
|---|---|---|---|---|---|
| 1. Tone | | Pos. | Neg. | Pos. | Neg. |
| 2. Grade level | a. Kindergarten | 7 | 72 | 8 | 176 |
| | b. 1–2 grade | 6 | 49 | 6 | 98 |
| | c. 3–5 grade | 2 | 13 | 2 | 18 |
| | d. 6–8 grade | 0 | 3 | 0 | 3 |
| | e. 9–12 grade | 0 | 8 | 0 | 8 |
| | | **8** | **89** | **9** | **212** |
| 3. Skill tested | a. Listening | 8 | 69 | 11 | 108 |
| | b. Speaking | 5 | 71 | 5 | 121 |
| | c. Reading | 3 | 75 | 3 | 113 |
| | d. Writing | 8 | 52 | 10 | 81 |
| | | **20** | **135** | **26** | **407** |
| 4. Appropriateness | a. Cognitive | 1 | 62 | 1 | 95 |
| | b. Content relativeness | 2 | 44 | 2 | 62 |
| | c. Difficulty level | 5 | 106 | 5 | 196 |
| | d. Length | 2 | 76 | 2 | 138 |
| | e. For specific student populations | 1 | 28 | 1 | 37 |
| | | **11** | **149** | **11** | **423** |
| 5. Impact | a. Available resources | 2 | 16 | 2 | 17 |
| | b. Instruction | 7 | 93 | 7 | 137 |
| | c. Students' psyche | 3 | 74 | 3 | 104 |
| | | **29** | **135** | **30** | **233** |
| 6. Logistics | a. Amount of time | 1 | 59 | 1 | 81 |
| | b. Overlap with other tests | 0 | 31 | 0 | 34 |
| | c. Educator training | 1 | 5 | 1 | 5 |
| | d. Personnel | 2 | 42 | 2 | 54 |
| | e. Physical space | 0 | 35 | 0 | 45 |
| | f. Scoring process | 2 | 40 | 2 | 52 |
| | g. Test materials | 0 | 59 | 0 | 87 |
| | | **5** | **137** | **5** | **285** |
| | Total | ***86*** | ***216*** | ***132*** | ***1015*** |

*Note.* Totals may not add up because subcategories may overlap with other subcategories; thus, a single educator's comment may be coded multiple times under a category, but will count as only one educator comment in the main category. Pos. = positive; Neg. = negative.

young children struggled with directions and that test administrators were not allowed to explain them. The educators mentioned that young learners had difficulties scoring high on the speaking exam because they exhibited behaviors that unexpectedly affected the scoring process. For example, the use of two-question prompts (which were designed to elicit two-part responses) resulted in low scores for some young ELLs because they answered only the second question. Other educators noted that

because in some cases the tests were administered not by students' teachers but by unfamiliar adults, young language learners may not have felt secure in the testing environment. The following comments exemplify key opinions expressed.

> **Example 1.** Having given the ELPA grades K-4, I feel it is somewhat appropriate at levels 1–4. However I doubt many of our American, English speaking K students could do well on the K test. This test covered many literacy skills that are not part of our K curriculum. It was upsetting for many of my K students. The grade 1 test was very difficult for my non-readers who often just stopped working on it. I felt stopping was preferable to having them color in circles randomly. *Educator 130, ESL teacher, administered Levels I , II & III, kindergarten-5th grade.*
>
> **Example 2.** When 5–8 year olds have to do that much reading, they shut down, especially when it's in another language. K-2nd do better reading out loud . . . they gave up! . . . A bubble test for an elementary student is ABSURD, they are more worried about coloring one than making sure it is correct!!!! Students 5–8 like to shout out answers. . . . It was a disaster!!! *Educator 133, ESL and bilingual teacher, administered Levels I & II, kindergarten-2nd grade.*

Another teacher wrote that the test was not appropriate for special education students and that some of those students "scored poorly due to their overall [cognitive] ability level, not to their English Proficiency level" (Educator 14, ESL teacher, administered all 5 Levels, kindergarten through 12th grade).

### *The length of the exam*

Table 5 shows that 76 educators provided 138 comments mentioning that the exam or parts of it were too long. One repeated comment was that the listening, reading, and writing subsections were unable to hold the students' attention because they were repetitive or had bland topics. Some representative comments included:

> **Example 3.** Some of the parts seemed to be a lot harder than the students' abilities—even for students who normally get really good grades. Reading for example for the first graders-the stories were very lengthy and it was quite a big section. Most of my first graders got bored after the second page. *Educator 195, ESL teacher, administered Levels I, II & III, kindergarten-5th grade.*
>
> **Example 4.** [W]riting activities were ok, but it was much too long for one sitting; too many essays were required; by the time students got to the last two they were tired of writing and didn't put forth much effort especially on the

3–5 test. *Educator 150, ESL teacher, administered Levels I, II & III, kindergarten-5th grade.*

**2. The Exam's Impacts**

The comments about the ELPA's impacts were both negative and positive. Concerns about the negative impacts of the ELPA tended to focus on two areas: The impact the test had on instruction, and the impact the test had on students psychologically. On the other hand, the largest portion (23%) of the positive comments about the ELPA were categorized under the theme of "impact." Each of these (the negative comments related to the impacts on instruction, the students' psyches, and the positive ones) are addressed in turn below.

### *The negative impacts on instruction*

The educators wrote that the ELPA did not directly affect the content of the ESL curriculum. However, many of the educators reported that the administration of the ELPA significantly reduced the amount of ESL classes and services offered in the spring of 2007 because the ESL teachers were tied up in testing and were thus not available for instruction. They reported that some ELLs did not receive any ESL services during the testing window at all, and in some cases missed out on mainstream and elective classes. Examples 5 through 7 illustrate these comments.

**Example 5.** For 2 groups, it took 5 weeks out of instruction because it take so long to administer; close to impossible to maintain any reasonable teaching schedule during the ELPA time span. *Educator 142, ESL teacher, administered Levels I , II & III, kindergarten-5th grade.*

**Example 6.** Since the ESL teacher did all of the testing, the ESL classes were left to the parapros for nearly a month—no substitute teacher was provided. Students missed nearly a month of quality instruction to accommodate testing. *Educator 77, ESL teacher, administered Level V, 9th–12th grade.*

**Example 7.** During ELPA (and also MEAP[5]) instruction comes to a halt. My ESL students are losing about 8 weeks of instruction per year. Needless to say that this issue impacts their progress negatively. *Educator 204, ESL teacher, administered all 5 Levels, kindergarten-12th grade.*

[5] This stands for the Michigan Educational Assessment Program (MEAP), which is used to test students in math, reading, and science as required by Title 1 of the NCLB Act.

### *The negative psychological impacts*

Seventy-eight educators wrote 109 comments on the psychological impact of the test on students. Educators wrote that the test stressed and frustrated some students, made them feel inadequate, humiliated, or embarrassed, or led them to question their self-worth. One educator reported that "beginning students, who did not even understand the directions, were very frustrated and in some cases crying because they felt so incapable" (Educator 109, ELL and resource room teacher, administered Levels II & III, 1st–5th grade). Another wrote that the test "adds more pressure" and described children "shak[ing] a little" (Educator 48, ESL teacher, administered Levels I , II & III, kindergarten-5th grade). Other illustrative comments included:

> **Example 8.** Sometimes, these poor kids come in a week, day or even during the test and the first thing we hit them with is this test. I find that very unfair. Often you can't even explain to the student that they don't need to worry that they can't complete it because they don't understand you. I think the ELPA is a good tool but should have wider parameters to allow for this type of circumstance. *Educator 30, ESL teacher, administered Levels III & IV, 3rd–8th grade.*
>
> **Example 9.** I think new arrivals should be exempt from taking the test. If they do have to take the test, then there should be a cut-off point that we can stop administering the test, so as not to humiliate them. *Educator 61, ESL teacher, administered all 5 Levels, kindergarten-12th grade.*

Educators also noted that some students protested the testing. Some students felt singled out; others apparently strongly resisted being identified as ELLs:

> **Example 10.** I feel it makes those students stand out from the rest of the school because they have to be pulled from their classes. They thoroughly despise the test. One student told me and I quote, "I wish I wasn't Caldean so I didn't have to take this" and another stated she would have her parents write her a note stating that she was not Hispanic (which she is). *Educator 71, Title 1 teacher, administered Levels I, II, III & IV, kindergarten-8th grade.*

### *The positive impacts*

Although individual students may have resisted being identified as ELLs, the typical positive comments about the test's impacts focused on the increased visibility of ELLs as a result of testing. As one teacher wrote, the "ELPA puts ELLs 'on the radar' so they are not as invisible" (Educator 156, ESL and other-subject teacher, administered Level V,

9th–12th grade). Some educators suggested that the ELPA can thereby increase funding and educational opportunities for ELLs in Michigan. For example, one teacher responded that "[i]n the big picture it keeps ELL's visible and districts accountable for ELL's" (Educator 182, ESL teacher, administered Levels II & III, 1st–2nd & 6th–8th grade). Other representative comments were:

> **Example 11.** It [the ELPA] gives the ESL and classroom teachers confirmation about their assessment of the student and sometimes new information that we can use to better instruct the individual. *Educator 80, ESL teacher, administered Levels I, II & III, kindergarten-5th grade.*
> **Example 12.** I think that this test has forced school districts that have small populations to take ESL seriously. *Educator 262, ESL teacher, administered Levels I, II & III, kindergarten-5th grade.*

### 3. The Logistics of Administering the Exam

Only five of the educators surveyed wrote specific comments stating that the administration of the ELPA went well. More vocal were those who experienced problems. Many voiced concerns that the administration was overly complex, required large amounts of educator time, and pulled key personnel from ESL instruction. For example, one teacher wrote that administering tests to 700 students "takes a lot of time from teaching. Teachers must give up teaching, use lunch hours and prep times to get this all done" (Educator 18, school testing coordinator, administered Levels I, II, III, & IV, kindergarten through 8th grade). Some students were tested in rooms that educators described as noisy, cramped, or full of distractions. The educators complained that other standardized tests were being administered shortly before, after, or during ELPA administration, with all tests competing for the educators' and students' time and resources. Other educators remarked that test materials or prerecorded audio often arrived late, or never came at all, and teachers had to read listening materials out loud to students.

In sum, the qualitative data relevant to Research Question 1 (What are teachers' opinions about the ELPA?) show that the majority (1,015 of the 1,147) of free-response comments were negative across all four groups of questions (particular skills, appropriateness, impact, and logistics). Positive comments included remarks that the ELPA had improved over the previous year's administration, is helping ELLs become more visible, and will help with placement and serve diagnostic purposes. The qualitative data relevant to Research Question 2 (Do educators' opinions vary according to demographic or teaching environment in which the ELPA was administered?) indicate that the educators most likely to provide free-response comments on the ELPA

were those who administered the exam to kindergarteners, first-graders, and second-graders. The 29 educators who mentioned the ELPA's effect on particular subgroups identified the very young, the recent arrivals, and ELLs who were also special education students.

## DISCUSSION

Traditionally, validity has been viewed and measured narrowly. Evaluations of validity have considered "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Test scores are interpreted to measure how consistent a test is with similar tests (concurrent validity) and how well a test predicts future performance (predictive validity). That narrow, score-based concept of validity is what official reviews have used to conclude that the ELPA is reliable and valid (Harcourt Assessment, 2006). Validity, however, should also include value judgments and social consequences (Messick, 1980, 1989, 1994, 1995; see chapter 2 in McNamara & Roever, 2006, and chapter 3 in Norris, 2008, for detailed explanations of Messick's theory of validity). Tests should be not only statistically reliable and valid, but also developmentally appropriate, fair, feasible, and practical (Hughes, 2003; Kane, 2002; Lane & Stone, 2002; Linn et al., 1991; Menken, 2008; Messick, 1989; Shohamy, 2000, 2001).

The results of this study illustrate the importance of considering validity in this broader sense. The educators' responses to the survey confirm that the 2007 ELPA had some of the collateral curricular and psychological effects that researchers have predicted. The data are largely consistent with Shohamy's (2001) argument that testing can impact instruction and the allocation of services and funds—both positively and negatively. A few educators commented that, by focusing attention on ELLs, the ELPA can increase funding and education opportunities for those students. On the other hand, several free-response comments indicated that administering the test kept experienced educators out of the classroom for as much as several weeks.

The free-response survey results were also consistent with research showing that language tests can affect students' psychological well-being by mirroring back to students a picture of how society views them and how they should view themselves (Menken, 2008; Schmidt, 2000). In this study, of the 107 comments made about the effect of the ELPA on students' psyches, 104 were coded as negative. As described by some

educators, the ELPA provides beneficial symbolic recognition of ELLs in Michigan, but it may also cause particular ELLs to be *misrecognized* (represented in inaccurate or harmful ways) and *nonrecognized* (ignored or made invisible; terms used by Taylor, 1994). It may seem difficult to harmonize the negative comments about the psychological impacts when, overall, the educators responded that generally they were pleased with the exam's impacts (see Figure 3). The qualitative comments help us understand this (see Examples 11 and 12 in particular). Regardless of whether the majority of educators liked the impact the ELPA had on the visibility of ELLs and ELL programs, the comments that the ELPA had a negative psychological impact on some students raise a red flag that suggests that further investigation into the test's psychological effects is needed.

Determining validity in a broad sense requires more than simply analyzing test scores. The perspectives of teachers and school administrators are indispensable for validating mandatory tests (Chapelle, 1998, 1999) and can contribute to understanding how the tests should be interpreted and used (Crocker, 2002; Lane & Stone, 2002; Ryan, 2002). Additionally, teachers and school administrators can spot unintended consequences of tests on teaching and learning (Norris, 2008). Much of what this study reveals about the administration and collateral impacts of the ELPA could not have been learned from looking at scores. The results of the study thus provide evidence that surveying the perspectives of educators is an important way to evaluate the broad validity of a test. This study strongly suggests that future official validations of the ELPA and other mandated ELL tests, in the United States and other countries where standardized tests are used to measure English language proficiency, should anonymously survey the educators who administer them. Furthermore, because states (in the case of the United States) and for-profit test agencies often have an incentive to avoid criticizing the tests they manage, it might be best to have an outside evaluator construct and conduct such a validity survey, as has been suggested by past researchers (Cronbach, 1989; Kane, 2002; Norris, 2008; Shepard, 1993). An outside evaluator who is neutral about the test may be the key to any large-scale test's broad validation process: The evaluator would summarize results and present them to the public, and suggest ways to improve the test (Ryan, 2002). This way of disseminating qualitative as well as quantitative validity data may also increase trust in the states (i.e., the organizations responsible for administration) and any corporate testing agencies they have hired and may encourage discussion about the uses of the test and the inferences that can be drawn from it.

The educators' opinions about the ELPA should not be dismissed as uninformed. Some criticisms the educators expressed are consistent

with current language-testing research. For example, many of the educators opined that the tests were too long for kindergarten, first, and second grade students, opinions supported by research concluding that the attention span of young children is as short as 10 to 15 minutes (McKay, 2006). Other educators commented that the testing format was not appropriate for the developmental level of their very young students. Those comments were not surprising, given research that states children cannot read silently until between the ages of seven and nine years (Pukett & Black, 2000). Moreover, children between 5 and 7 years of age are still developing their gross and fine motor skills (McKay, 2006) and so may not even be able to fill in bubble-answer-sheets.

When policy makers and test creators evaluate the broad validity of a widely administered language test such as the ELPA, they should consider the possibility that broad validity can vary across subgroups. In this respect, the broad conception of validity adopted in this study is different from the narrow conception, which includes only reliability, concurrent validity, and predictive validity. Those narrower measures of validity are inherent to the test and so should logically be uniform, regardless of who is taking the test. Because broad validity takes into account the social context and social impacts of a test, however, it should be expected to vary depending on where the test is administered. The Likert-scale responses on the test administration factor provide an example. Although Figure 3 illustrates that educators were significantly more critical of the administration of the test than any other factor, the ANOVA analysis summarized in Figure 4 reveals an important distinction that is hidden in the overall mean scores: Educators at schools where less than 25% of the students were ELLs were more than twice as critical of the test administration than educators at schools with more than 25% ELLs. That difference in attitude may exist because, in schools and districts with very low concentrations of ELLs, the burden of testing can fall on a few teachers, or even a single teacher, who must balance testing and teaching at several different remote schools. When policy makers and test creators learn that the validity of a widely administered test is not uniform, they will face difficult decisions about whether and how to alter the test program. Still, those decisions should be based on the fullest possible information about the effects of a test.

Ultimately, the data collected in this survey suggest two conclusions. First, the administration of the ELPA has some collateral effects on teaching and on students' psychological well-being that need to be further investigated. Second, the data about the ELPA provide empirical evidence for some theories about broad validity that apply to similar testing conditions in many countries. Broad validity is

important, because language tests can have collateral effects. Broad validity is not a fixed property of a test, but can vary depending on the context in which a test is administered. This study has shown that anonymously surveying educators about their experiences with a test provides critical information about whether the test is valid. Teachers' voices should be heard and not be ignored. By taking teachers' perceptions into consideration, standardized testing programs can improve their ESL or English as a foreign language tests and better ensure they are broadly valid—that the tests produce reliable scores and have positive impacts.

## LIMITATIONS AND DIRECTIONS FOR FURTHER RESEARCH

The study described here had two main limitations. First of all, the responders were self-selected and so were not a representative random sample. Educators with strong opinions, particularly those with negative opinions, were probably more likely to respond. Nonetheless, this study had a reasonably large sample size and used a mixed-research design to gather both quantitative and qualitative data. Together these two design features may allow for a meaningful sampling of educators' perceptions when random sampling is not possible.

Second, the study group was limited to educators. The broad validity of NCLB-mandated English language proficiency tests could be more thoroughly evaluated by triangulating statistical and qualitative data from educators with data from other stakeholders, including students and parents. Future studies with a broader and more representative sample should examine whether different stakeholder groups hold different attitudes regarding the validity of high-stakes tests such as the ELPA. At the same time, it would be helpful to conduct similar studies in other countries where standardized tests are as consequential as the ELPA. Doing so would help us gain a more comprehensive view of stakeholders' important contributions to such tests' validity arguments.

## CONCLUSION

The 267 Michigan educators who responded to an online survey concerning the large-scale, ELPA they recently administered expressed a variety of opinions related to the broad validity of the test. They described (a) the test's sometimes problematic administration, (b) the both positive and negative impacts the test had, and (c) how the test's validity waned at schools with lower concentrations of ELLs, with

special education and recent immigrant ELLs, and, in particular, with younger test takers. No findings like these were reported in the validity section of the ELPA technical manual (Harcourt Assessment, 2006). Also no results like these could have been gleaned from test scores. This study thus shows that surveying educators about a widely administered test such as the ELPA is worthwhile because their responses provide valuable information about the psychological, ethical, and social consequences of a test, information that might not otherwise be available.

The key goal of establishing a validity argument is to improve the assessment instrument (Chapelle, 1999; Norris, 2008). Although it is admittedly difficult to change a test once it is fully operational (Kane, 2002), this article's results shed light on how the Michigan ELPA could be improved. More importantly, this article demonstrates that collecting qualitative data from stakeholders provides rich and important information about the broad validity of a testing program. Such information can be used to ensure that large-scale testing programs like the ELPA are accountable not only to the entities that mandated them, but also to those for whom the tests are intended to serve—students, educators, and the public at large.

## ACKNOWLEDGMENTS

This project was supported by a Michigan State University College of Arts and Letters Public Humanities Grant. I would like to thank Li Shaofeng, Vineet Bansal, and Ashley Romanowski for their help with this project. I am also grateful to Lucy Pickering, Shawn Loewen, Debra Friedman, and the anonymous *TESOL Quarterly* reviewers for their instructive comments on earlier versions of this article. Any mistakes, however, are my own.

## THE AUTHOR

Paula Winke is an assistant professor of second language studies in the Department of Linguistics and Germanic, Slavic, Asian, and African Languages at Michigan State University, East Lansing, Michigan, United States. Her research focuses on second language testing and individual differences in second language acquisition.

## REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* Upper Saddle River, NJ: Prentice Hall Regents.
Canale, M. (1987). The measurement of communicative competence. *Annual Review of Applied Linguistics, 8*, 67–84. doi: 10.1017/S0267190500001033.
Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267–296. doi: 10.1016/S1060-3743(03)00038-9.
Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge, England: Cambridge Applied Linguistics.
Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254–272. doi: 10.1017/S0267190599190135.
Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2007). Children left behind in AYP and non-AYP schools: Using student progress and the distribution of student gains to validate AYP. *Educational Measurement: Issues and Practice, 26*, 21–32. doi: 10.1111/j.1745-3992.2007.00098.x.
Crawford, J. (2000). *At war with diversity: US language policy in an age of anxiety.* Clevedon, England: Multilingual Matters.
Crocker, L. (2002). Stakeholders in comprehensive validation of standards-based assessments: A commentary. *Educational Measurement: Issues and Practice, 22*, 5–6. doi: 10.1111/j.1745-3992.2003.tb00132.x.
Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438–481.
Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement theory and public policy* (pp. 147–171). Urbana, IL: University of Illinois Press.
Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing.* Mahwah, NJ: Lawrence Erlbaum Associates.
Evans, B. A., & Hornberger, N. H. (2005). No Child Left Behind: Repealing and unpeeling federal language education policy in the United States. *Language Policy, 4*, 87–106. doi: 10.1007/s10993-004-6566-2.
Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.
Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 22*, 16–22.
Harcourt Assessment, Inc. (2006). *2006 MI-ELPA technical manual* [Electronic version] Retrieved from http://www.michigan.gov/documents/mde/MI-ELPA_Tech_Report_final_199596_7.pdf
Hill, R. K., & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice, 22*, 12–20. doi: 10.1111/j.1745-3992.2003.tb00133.x.
Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.
Jöreskog, K. G. (1999). *How large can a standardized coefficient be?* Retrieved from http://www.ssicentral.com/lisrel/techdocs/HowLargeCanaStandardizedCoefficientbe.pdf
Kane, M. J. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*, 31–41. doi: 10.1111/j.1745-3992.2002.tb00083.x.
Kearney, K. A., Hopkins, R. H., Mauss, A. L., & Weisheit, R. A. (1984). Self-generated identification codes for anonymous collection of longitudinal questionnaire data. *Public Opinion Quarterly, 48*, 370–378. doi: 10.1086/268832.
Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* New York, NY: McGraw-Hill.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*, 6–14. doi: 10.1111/j.1745-3992.2004.tb00160.x.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 22*, 23–30.

Lazarín, M. (2006). *Improving assessment and accountability for English language learners in the No Child Left Behind Act*. Washington, DC: National Council of La Raza.

Linn, R. L., Baker, E. L., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 5–21.

McKay, P. (2006). *Assessing young language learners*. Cambridge, England: Cambridge University Press.

McNamara, T. (2000). *Language testing*. Oxford, England: Oxford University Press.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

Menken, K. (2008). *Standardized testing as language policy: English language learners left behind*. Clevedon, England: Multilingual Matters.

Messick, S. (1980). Test validation and the ethics of assessment. *American Psychologist, 35*, 1012–1027. doi: 10.1037/0003-066X.35.11.1012.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York, NY: American Council on Education & Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13–23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8. doi: 10.1111/j.1745-3992.1995.tb00881.x.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256. doi: 10.1177/026553229601300302.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*, 6–12. doi: 10.1111/j.1745-3992.1998.tb00826.x.

Norris, J. M. (2008). *Validity evaluation in language assessment*. Frankfurt, Germany: Peter Lang.

Porter, A. C., Linn, R. L., & Trimble, C. S. (2005). The effects of state decisions about NCLB Adequate Yearly Progress targets. *Educational Measurement: Issues and Practice, 24*, 32–39. doi: 10.1111/j.1745-3992.2005.00021.x.

Pukett, M. B., & Black, J. K. (2000). *Authentic assessment of the young child*. Upper Saddle River, NJ: Prentice-Hall.

Roberts, M., & Manley, P. (2007). *Planning for the English Language Proficiency Assessment (ELPA)* Retrieved from http://www.michigan.gov/documents/mde/S07_Planning_For_The_ELPA_184193_7.ppt

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 22*, 7–15.

Schmidt, R. (2000). *Language policy and identity politics in the United States*. Philadelphia, PA: Temple University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Shohamy, E. (2000). Fairness in language testing. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (Vol. 9, pp. 15–19). Cambridge, England: Cambridge University Press.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Essex, England: Pearson Education Limited.

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. New York, NY: Routledge.

Stansfield, C. W., Bowles, M. A., & Rivera, C. (2007). *Case studies of state policies and practices concerning test translation.* Mahwah, NJ: Lawrence Erlbaum.
Stansfield, C. W., & Rivera, C. (2001). *Test accommodations for LEP students.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
Taylor, C. (1994). *Multiculturalism: Examining the politics of recognition.* Princeton, NJ: Princeton University Press.
U.S. Department of Education. (2002). *PUBLIC LAW 107–110—JAN. 8, 2002.* Retrieved from http://www.ed.gov/policy/elsec/leg/esea02/index.html
U.S. Department of Education. (2004a). *Fact sheet: NCLB provisions ensure flexibility and accountability for limited English proficient students.* Retrieved from http://www.ed.gov/nclb/accountability/schools/factsheet-english.html
U.S. Department of Education. (2004b). *Testing: Frequently asked questions.* Retrieved from http://www.ed.gov/nclb/accountability/ayp/testing-faq.html
Wallis, C., & Steptoe, S. (2007, June 4). How to fix No Child Left Behind. *Time,* 35–40.
Wiley, T. G., & Wright, W. E. (2004). Against the undertow: Language-minority education policy and politics in the "age of accountability." *Educational Policy, 18,* 142–168. doi: 10.1177/0895904803260030.
Zehr, M. A. (2006). 'No Child' effect on English-learners mulled. *Education Week, 25, 1,* 14–15.
Zehr, M. A. (2007). A balancing act: NCLB's renewal, English-learners. *Education Week, 26,* 9.

## Appendix A

## ELPA Survey Demographic Questions

1. How are you involved with the school? I am a/an . . . (Check all that apply.)

   ○ English as a Second Language (ESL) teacher
   ○ Language Arts teacher
   ○ English literature teacher
   ○ Teacher of other subjects (i.e., biology, physical education)
   ○ School Principal
   ○ School Administrator
   ○ Parent of a student in the school who took the ELPA
   ○ Student
   ○ Volunteer at the school
   ○ Other (Please specify: _____)

2. With what level of the test were you involved? (Check all that apply.)

   ○ Level I: Kindergarten
   ○ Level II: Grades 1–2
   ○ Level III: Grades 3–5
   ○ Level IV: Grades 6–8
   ○ Level V: Grades 9–12

3. Who administered the ELPA at your school? (Check all that apply.)

- ○ English as a Second Language (ESL) teachers
- ○ Language Arts teachers
- ○ English literature teachers
- ○ Teachers of other subjects (i.e. biology, physical education)
- ○ School Principal(s)
- ○ School Administrator(s)
- ○ Parent(s) of students who took the ELPA
- ○ Volunteers from outside the school
- ○ Teachers' aids
- ○ Other (Please specify: _____)

4. What portions of the ELPA did you administer? (Check all that apply.)

- ○ Listening
- ○ Speaking
- ○ Reading
- ○ Writing

5. How would you describe your school? (Check all that apply.)

- ○ Urban
- ○ Rural
- ○ Suburban
- ○ Public
- ○ Magnet
- ○ Charter
- ○ Private
- ○ Religious-affiliated

6. Approximately what percentage of your school is made up of English Language Learners (ELLs)?

- ○ Less than 5 percent
- ○ 5 percent
- ○ 10 percent
- ○ 15 percent
- ○ 20 percent
- ○ 25 percent
- ○ 30 percent
- ○ 35 percent
- ○ 40 percent
- ○ More than 40 percent

# Appendix B

## Factor Analysis Pattern Matrix

| | Factor | | | | |
|---|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 | 5 |
| 30. The second part of the writing test (essay writing) is a positive feature of the test. | 0.94 | | | | |
| 31. I feel the writing test adequately measured the students' true writing ability. | 0.91 | | | | |
| 28. The writing test is well designed. | 0.89 | | | | |
| 27. I feel the reading test adequately measured the students' true reading ability. | 0.71 | | | | |
| 29. The first part of the writing test (about writing conventions) is a positive feature of the test. | 0.63 | | | | |
| 24. The reading test is well designed. | 0.58 | | | | |
| 4. The school had enough personnel to administer the test smoothly. | | 0.86 | | | |
| 5. Overall, the administration of the ELPA ran smoothly. | | 0.84 | | | |
| 6. The teachers had enough support in administering the ELPA. | | 0.80 | | | |
| 3. The school had enough physical space and equipment to administer the test smoothly. | | 0.67 | | | |
| 9. Overall, the ELPA is a beneficial test for the English language learners (ELLs). | | | 1.00 | | |
| 8. English as a Second Language (ESL) instruction at the school was positively impacted by the ELPA. | | | 0.83 | | |
| 17. The ELPA has had a positive impact on the students' English language ability. | | | 0.64 | | |
| 10. Overall, the ELPA materials were well designed. | | | 0.63 | | |
| 11. Overall, I feel the ELPA test results will be reliable and valid. | | | 0.56 | | |
| 33. The rubric for the speaking test was well designed. | | | | 1.02 | |
| 34. The rubric for the speaking test was easy to follow. | | | | 0.89 | |
| 35. I feel the speaking test adequately measured the students' true speaking ability. | | | | 0.68 | |
| 32. The speaking test procedures worked well. | | | | 0.66 | |
| 22. The listening portion of the listening test was easy for the students to understand. | | | | | 0.86 |
| 23. I feel the listening test adequately measured the students' true listening ability. | | | | | 0.84 |
| 20. The listening test is well designed. | | | | | 0.73 |
| 21. The administration of the listening test was easy. | | | | | 0.66 |

*Note.* Extraction Method: Maximum Likelihood. Rotation Method: Promax with Kaiser Normalization. Rotation converged in 7 iterations.

# Appendix C

## ELPA survey open-ended items

1. How did the students in your school prepare for the ELPA?
2. Were there any special circumstances at your school that affected the administration of the ELPA? If so, please describe.
3. Does the ELPA affect instruction in your school, and if so, it is positive, negative, or both? Please describe below how it affects instruction at your school.
4. What effect does the ELPA have on the English language learners (ELLs) at your school?
5. Is there anything else you would like to say about Michigan's ELPA?