Project 2

University ESL reading test for making placement/exemption decisions

SETTING

The director of an ESL program at a large research university in which English is the medium of instruction needs to collect information to make decisions about placing non-native English-speaking students who have been admitted to the university into an ESL reading course or exempting them from study in these courses. For the past several years, she has been using a fairly traditional multiple-choice reading test, administered in a paper and pencil format, for collecting this information. However, the costs of producing copies of the test, of administering it to large groups of students, and of producing test reports and sending these to the students and university administrators have escalated at a time when the university is facing a budget crisis. She is thus under pressure from the administration to find a way to cut the costs of administering this test. In addition, the teachers in the ESL reading course have become increasingly dissatisfied with the test itself. They feel that the test items do not reflect what students actually do in their academic reading courses, and they believe the test does not really measure what students need in order to complete these courses successfully.

Based on their many years of experience, the ESL director, the ESL reading teachers, and the academic course teachers agree that placing students who need developmental instruction in English academic reading into an ESL reading course greatly facilitates their academic coursework and reduces the time to their degree. They also know that exempting students who do not need additional instruction conserves resources for both teachers and students. Finally, they all agree that the most effective way to obtain the information they need is to give students a placement test. Thus, the decision is not to eliminate the use of a test, but whether to adapt an existing one or develop a new one.

Since the decisions the ESL Program Director needs to make are relatively high stakes, she is willing to allocate considerable resources to assessment development, if it is determined that this is needed. As in Project 1, the ‘test developer’ consists of a team of individuals. In this example, the team includes the ESL Program Director and ESL instructors, with all members of the team assuming a collective responsibility for the test development, use, and justification. While the ESL Program Director is the decision-maker, the ESL instructors and academic course instructors can also be considered to be test users.
ASSESSMENT USE ARGUMENT

Consequences

**Claim 1:** The consequences of using the reading assessment and of the placement/exemption decisions that are made are beneficial to the test takers, teachers in the ESL reading courses, the ESL Program Director, and instructors in academic courses at the University who will encounter ESL students in their classes.

**Stakeholders**

1. Non-native English speaking students entering the university
2. ESL course instructors
3. ESL Program Director
4. Instructors in academic courses

**Warrants: Consequences of using the ESL reading test**

A1. The consequences of using the outline reading test that are specific to the test takers, teachers in the ESL reading courses, the ESL Program Director, and instructors in academic courses at the University who will encounter ESL students in their classes will be beneficial.

A2. Assessment reports, which include 1) the scores from the outline reading test and 2) the placement decisions made on the basis of them, are treated confidentially.

A3. Assessment reports, which include 1) the scores from the outline reading test and 2) the placement decisions made on the basis of them, are presented in ways that are clear and understandable to all the test takers.

A4. The ESL Program Director distributes the assessment reports to test takers and authorized university faculty and officials in time for them to be used for the intended decisions.

A5. Use of the outline reading test helps promote good instructional practice and effective learning by linking the assessment task specifications to one of the performance objectives in the ESL reading course into which students are placed.

**Warrants: Consequences of the decisions that are made**

B1. The consequences of the placement and exemption decisions that are made will be beneficial for the students.

B2. The consequences of the placement and exemption decisions that are made will be beneficial for the teachers in the ESL reading course.

B3. The consequences of the placement and exemption decisions that are made will be beneficial for the teachers in academic university courses.

B4. The consequences of the placement and exemption decisions that are made will be beneficial for the ESL program administrator.

**Rebuttal:** The consequences of false positive and false negative classification errors will be different, as follows:

© 2010 Lyle F. Bachman, Adrian S. Palmer, and Oxford University Press
1. **False positive classification errors:** Exempting students from the ESL reading course who actually need the course will have detrimental consequences for students because they will feel overwhelmed, frustrated, may get low grades their academic courses, and may even be dropped from the university. Furthermore, even though these students may have to struggle in their academic courses, they may not report this, and so there is little chance of helping them. Instructors in academic courses may also feel frustrated because their students are not able to keep up with the reading assignments. These instructors may complain to the ESL program administrator who will have to deal with complaints. In this example, therefore, false positive classification errors are regarded as very serious.

2. **False negative classification errors:** Placing students from the ESL reading course who do not need it will have detrimental consequences for students. Students may feel bored with the coursework and resent what they may perceive as time and money wasted for the course. They may also suffer because taking an ESL class prevents them from taking an additional academic course, which may delay their progress toward their degree. In this example, therefore, false negative classification errors are regarded as moderately serious.

**Possible ways of mitigating the detrimental consequences of decision classification errors if they occur**

1. **False positive classification errors:** Teachers in regular university courses will be notified of the existence of the ESL reading courses and asked to be alert for students who seem to be struggling with the reading assignments. They will be asked to advise students to consider taking the ESL reading course even though they were exempted from it on the basis of the ESL reading test.

2. **False negative classification errors:** ESL reading teachers will be alerted to be on the lookout for misplaced students, to collect additional information about the students’ reading ability during the first week of class, and to make recommendations to the ESL Program Director for exempting those students whom they consider do not need the class.

**Decisions**

| Claim 2: | The decisions to place or exempt students from the ESL reading course reflect relevant existing educational and societal values and relevant university regulations and are equitable for those students who are placed or exempted. |
|

The decisions, stakeholders affected by decisions, and individuals responsible for making the decisions are provided in Table P2.1 below.
<table>
<thead>
<tr>
<th>Decisions</th>
<th>Stakeholders who will be affected by the decision</th>
<th>Individual(s) responsible for making the decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place students in ESL reading course.</td>
<td>Students, instructors in ESL reading course, teachers in students’ academic courses</td>
<td>ESL Program Director</td>
</tr>
<tr>
<td>Exempt students from ESL reading course.</td>
<td>Students, instructors in ESL reading course, teachers in students’ academic courses</td>
<td>ESL Program Director</td>
</tr>
<tr>
<td>Adjust instruction in ESL reading course.</td>
<td>Instructors in ESL reading course, students in ESL reading course</td>
<td>Instructors in ESL reading course</td>
</tr>
</tbody>
</table>

Table P2.1  The decisions, stakeholders affected by decisions, and individuals responsible for making the decisions

Warrants: Values sensitivity

A1. Relevant educational values of the university community and legal requirements of the university itself as an institution are carefully considered in the placement and exemption decisions that are made.

A2. Existing educational values of the ESL teachers and academic instructors, as well as relevant university regulations are carefully considered in determining the relative seriousness of false positive and false negative classification errors.

A3. Cut scores are set relatively high so as to minimize the most serious classification errors, that is, false positives.
   a. Relative seriousness of classification decision errors: False positive classification errors are more serious than false negative ones.
   b. Policy-level procedures for setting standards: The standard for exemption from the ESL reading course was set by the ESL Course Director in consultation with ESL Reading course instructors and academic course instructors.
   c. Standard for exemption from the ESL reading course: Students must demonstrate that they have sufficient knowledge of multilevel rhetorical organization of introductory academic texts to enable them to successfully participate in introductory level academic courses.

Warrants: Equitability

B1. The same cut score is used to classify all students taking the ESL Outline Reading Test, and no other considerations are used.
B2. Test takers, ESL reading teachers, and other individuals within the University community are fully informed about how the decision will be made and whether decisions are actually made in the way described to them.

B2. No decisions are made involving achievement or certification. The only decisions made are ‘placed’ or ‘exempt’. Therefore, this warrant is not needed.

Interpretations

Claim 3: The interpretations about the students’ ‘knowledge of multilevel rhetorical organization of written texts’ are meaningful in terms of an analysis of texts used in introductory level academic university courses, impartial to all groups of test takers, generalizable to reading tasks in texts used in introductory level university courses, and relevant to and sufficient for the placement decisions that are to be made.

Warrants: Meaningfulness

A1. The interpretations about the students’ ‘knowledge of multilevel rhetorical organization of written texts’ are meaningful with respect to introductory level academic university courses, generalizable to introductory level academic university courses, and relevant to and sufficient for the placement decisions that are to be made.

- The definition of the construct is ‘knowledge of multilevel rhetorical organization of written texts’.

- This definition is based on a needs analysis that included input from instructors in introductory level academic university courses, students who had completed these courses, and teachers in the ESL courses. It clearly distinguishes the construct from other related constructs, such as ability to draw inferences, ability to read critically, or knowledge of figures of speech.

A2. The assessment task specifications clearly specify that the test takers will read a passage taken from an introductory academic text and then complete an incomplete outline that represents the rhetorical organization of this passage.

A3. The procedures for administering the incomplete outline test enable the test takers to perform at their highest level on the ability ‘knowledge of multilevel rhetorical organization of written texts’.

A4. The scoring key and procedures for using the key focus on elements of multilevel outline structure that, in essence, define the construct ‘knowledge of multilevel rhetorical organization of written texts’.

A5. The incomplete outline task engages the ability ‘knowledge of multilevel rhetorical organization of written texts’.

A6. Scores on the incomplete outline test are interpreted as ‘knowledge of multilevel rhetorical organization of written texts’.

A7. The ESL Program Director communicates the definition of the construct in non-technical language via the instructions for the outline completion task and examples of
multilevel outlines. The construct definition is also included in non-technical language in the assessment report for test takers and other stakeholders.

**Warrants: Impartiality**

B1. The outline completion task does not include response formats or content that may either favor or disfavor some test takers.

B2. The outline completion task does not include content that may be offensive to some test takers.

B3. The procedures for producing an assessment record for the incomplete outline test are clearly described in terms that are understandable to all test takers.

B4. Test takers are treated impartially during all aspects of the administration of the assessment.
   a. Test takers have equal access to information about the assessment content and assessment procedures.
   b. Test takers have equal access to the assessment, in terms of cost, location, and familiarity with conditions and equipment.
   c. Test takers have equal opportunity to demonstrate their knowledge of multilevel rhetorical organization of introductory university academic reading passages.
   d. Not needed, since the test is administered by computer over the internet.

B5. Interpretations of the test takers’ ‘knowledge of multilevel rhetorical organization of written texts’ are equally meaningful across students from different first language backgrounds and academic disciplines.

**Warrants: Generalizability**

C1. The characteristics of the outline completion task correspond closely to those of tasks of reading introductory academic texts both in the academic domain (non-ESL introductory-level university courses) as well as in the language teaching domain (the ESL reading course).

C2. The criteria and procedures for evaluating the responses to the outline completion task correspond closely to those that are typically used by teachers in assessing performance in reading introductory academic texts.

**Warrant: Relevance**

D. The interpretation ‘knowledge of rhetorical organization of written texts’ provides the information that is relevant to the ESL Program Director’s decisions about placement or exemption.

**Warrant: Sufficiency**

E. The assessment-based interpretation of ‘knowledge of multilevel rhetorical organization of written texts’ provides sufficient information to make the placement/exemption decisions.
Assessment records

Claim 4: The scores from the incomplete outline test are consistent across different forms and administrations of the test, across students from different academic disciplines, and across different groups of international students entering the university.

Warrants: Consistency

1. The incomplete outline test is administered in a standard way every time it is offered.
2. The scoring criteria and procedures for the computer scoring algorithm are well specified and are adhered to.
3. The computer scoring algorithm was developed through extensive trialing and comparison with multiple human ratings.
4. The computer scoring algorithm was developed through trialing with several different groups of test takers.
5. Scores on different items in the incomplete outline test are internally consistent.
6. Not needed, since performance is not rated.
7. Not needed, since performance is not rated.
8. Scores from different forms of the incomplete outline test are consistent.
9. Scores from different administrations of the incomplete outline test are consistent.
10. Scores on incomplete outline test are consistent across different groups of students.
DESIGN STATEMENT

1. Description of the test takers and other stakeholders (in Table P2.2)

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Test takers</td>
<td>English Language Learners (ELLs); varied languages and cultures; upper level ESL proficiency, adults (ages mostly in upper teens and early 20s, with some older students)</td>
</tr>
<tr>
<td>2 ESL reading teachers</td>
<td>Teaching Assistants (TAs) in M.A. and Ph.D. Applied Linguistics Program, many with several years of experience teaching in university ESL program, some with many years of experience teaching ESL prior to involvement in the graduate program in Applied Linguistics</td>
</tr>
<tr>
<td>3 Teachers in university academic classes</td>
<td>Regular faculty teaching introductory general education courses</td>
</tr>
<tr>
<td>4 ESL Program Director</td>
<td>Professor of Linguistics, 25 years’ experience directing the ESL program in a large US university; extensive experience in language test design and development</td>
</tr>
</tbody>
</table>

Table P2.2 Attributes of stakeholders
### Intended beneficial consequences (in Table P2.3)

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Intended beneficial consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Test takers</strong></td>
<td>Of using the assessment: Test takers placed in the ESL reading class will realize that the test tasks are similar to instructional tasks, and thus relevant to their target language use (TLU) needs. Test takers placed in the ESL reading class will benefit from using the test by being tested in a way that is consistent with ways in which their performance in the ESL reading course is being evaluated.</td>
</tr>
<tr>
<td><strong>2 ESL reading teachers</strong></td>
<td>ESL reading teachers will benefit from using a test in which the criteria for making placement decisions are similar to those used in making decisions about the effectiveness of their instruction. ESL reading teachers will benefit from being able to focus their instruction on a group of students who are relatively homogeneous in their reading ability.</td>
</tr>
<tr>
<td><strong>3 Teachers in regular university academic classes</strong></td>
<td>Teachers in regular academic courses will be aware of the fact that students placed in their class, by virtue of having taken the test, will have been sensitized to the need to pay attention to the rhetorical organization of material they read. Teachers in regular university courses will benefit from having students in their classes who are prepared to read and understand the rhetorical structure of introductory texts.</td>
</tr>
<tr>
<td><strong>4 ESL program administrator</strong></td>
<td>The ESL program administrator will benefit from using a test whose scoring criteria are consistent with the performance objectives for the course they supervise. The ESL program administrator will have to deal with fewer complaints from bored or frustrated students and frustrated teachers.</td>
</tr>
</tbody>
</table>

*Table P2.3  Intended beneficial consequences*
2. Descriptions of the decisions to be made, the stakeholders affected by decisions, and individuals responsible for making the decisions (in Table P2.4).

<table>
<thead>
<tr>
<th>Decision</th>
<th>Stakeholders who will be affected by the decision</th>
<th>Individual(s) responsible for making the decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place students in ESL reading course.</td>
<td>Students, instructors in ESL reading course, teachers in students’ academic courses</td>
<td>ESL Program Director</td>
</tr>
<tr>
<td>Exempt students from ESL reading course.</td>
<td>Students, instructors in ESL reading course, teachers in students’ academic courses</td>
<td>ESL Program Director</td>
</tr>
<tr>
<td>Adjust instruction in ESL reading course.</td>
<td>Instructors in ESL reading course, students in ESL reading course</td>
<td>Instructors in ESL reading course</td>
</tr>
</tbody>
</table>

Table P2.4 The decisions, stakeholders affected by decisions, and individuals responsible for making the decisions

3. The relative seriousness of classification errors, policy-level decisions about standards, the standards themselves

- **Relative seriousness of classification errors**: False positive classification decisions are relatively more serious than false negative classification decisions.

- **Policy-level procedures for setting standards**: The standard for exemption from the ESL reading course was set by the ESL Course Director in consultation with ESL Reading course instructors and academic course instructors.

- **Standard for exemption from the ESL reading course**: Students must demonstrate that they have sufficient knowledge of multilevel rhetorical organization of introductory academic texts to enable them to successfully participate in introductory level academic courses.

4. Definition of the construct(s) (Claim 3, Warrant A1)

The definition of the construct for the University example is ‘knowledge of multilevel rhetorical organization written texts.’

This definition of the construct to be assessed is based upon a needs analysis of reading tasks in courses fulfilling students’ general education requirement, and this is the domain within which the construct is meaningful.

5. Description of the TLU domain

The TLU domain in the University example is ‘academic readings in textbooks in introductory courses’.

6. Tasks selected as a basis for developing assessment tasks

© 2010 Lyle F. Bachman, Adrian S. Palmer, and Oxford University Press
The TLU task that was selected as a basis for developing assessment tasks is ‘reading a passage from a text in an introductory academic course and developing an outline of this.’

7. Description of the characteristics of TLU task that has been selected as a basis for assessment tasks (Table P2.5)

<table>
<thead>
<tr>
<th>Characteristics of outlining TLU Task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Setting</strong></td>
</tr>
<tr>
<td>Physical characteristics: home, library, classroom, computer, course text</td>
</tr>
<tr>
<td>Participants: the student; possibly teacher and other students</td>
</tr>
<tr>
<td><strong>Rubric (all implicit in the TLU task)</strong></td>
</tr>
<tr>
<td><strong>Instructions</strong></td>
</tr>
<tr>
<td>Target language: written, visual, or internally generated by the student; specification of procedures and tasks based upon students’ prior instruction in and experience with outlining</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
</tr>
<tr>
<td>Number of parts: one part (outline) per reading.</td>
</tr>
<tr>
<td>Salience of parts and tasks: the student used the language, but it was not free from error (either syntactically or semantically)</td>
</tr>
<tr>
<td>Sequence of tasks: determined by sequence of information in reading</td>
</tr>
<tr>
<td>Relative importance of tasks: tasks associated with main headings may be relatively more important than tasks associated with relatively minor details in subheadings</td>
</tr>
<tr>
<td>Number of tasks: depends upon length of reading passage, number of levels in outline, and number of entries under each level</td>
</tr>
<tr>
<td><strong>Time allotment</strong></td>
</tr>
<tr>
<td>Highly variable</td>
</tr>
<tr>
<td><strong>Recording method</strong></td>
</tr>
<tr>
<td>Criteria for recording: degree of association of organization of outline with organization of reading text; meaning of entries far more important than grammatical accuracy</td>
</tr>
<tr>
<td>Procedures for recording the response: student’s internally generated feedback based upon usefulness of the outline in helping the student interact with the material in the reading; possible written or spoken feedback from instructor</td>
</tr>
<tr>
<td>Explicitness of criteria and procedures for recording the response: fairly explicit if student has been given formal instruction in outlining</td>
</tr>
<tr>
<td>Recorders: variable (the student, other students, instructors)</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>---</td>
</tr>
</tbody>
</table>
| **Language characteristics** | Grammatical: complex  
Textual: highly organized  
Functions: most frequently ideational, can also be manipulative (persuasive)  
Genre: introductory academic course text  
Dialect: academic  
Register: formal  
Naturalness: natural  
Cultural references: variable  
Figures of speech: variable |
| **Topical characteristics** | Variable: topics consistent with introductory general education course content |
| **Expected response** | Written, English, extended production, relatively unspeeded |
| Format | Written, English, extended production, relatively unspeeded |
| **Language characteristics** | Grammatical: somewhat complex  
Textual: outline  
Functions: ideational  
Genre: outline  
Dialect: standard academic  
Register: formal  
Naturalness: natural  
Cultural references: variable, depending upon course content  
Figures of speech: variable |
| **Topical characteristics** | Variable: topics consistent with introductory general education course content |
| **Relationship between input and expected response** |  |
| **Type of external interactivity** | Non-reciprocal interaction between test taker, reading passage, incomplete outline as provided in test, and developing outline emerging as test taker completes the test |
| **Scope** | Narrow |
| **Directness** | Direct |
Table P2.5 Characteristics of the selected TLU task
BLUEPRINT

I. Assessment specifications
   A. Number of parts: one, consisting of one written passage and one incomplete outline
   B. Number of tasks per part: approximately 10–15 blanks in the outline per reading passage. These blanks include at least 1 blank at highest level of information, at least 4 blanks at the second level of information, at least 3 blanks at the third level of information.
   C. Sequence of parts/tasks: Reading passage was presented first, then the incomplete outline could be viewed alongside the reading passage in a separate column.
   D. Relative importance of parts/tasks: all parts of same importance
   E. Time allotment: generous time limit
   F. Instructions: general and for each part
      1. General
         a. See example tutorial below.
         b. Complete the outline based on your reading of the passage. (See example tutorial below.)
         c. When you have completed the outline, click the ‘submit’ button to send your answers to the testing center and continue the test.
      2. For each part: instructions are the same for each part.

II. Task specifications
   A. Definition of ability to be assessed: knowledge of multilevel rhetorical organization
   B. Setting
      1. Physical characteristics: computer lab on campus, quiet, comfortable
      2. Equipment: each test taker seated at a PC with headphones
      3. Attributes of participants
         a. Test takers: incoming students who are non-native speakers of English
         b. Degree of familiarity with PC: all quite familiar with PC
         c. Test administrator: trained and experienced in computer-based testing and having a positive attitude toward the test takers
      4. Time of task: by appointment within a fixed time period during the day
   C. Characteristics of the input, expected response, and relationship between input and response
      1. Input
         a. Format
            1) Channel: visual (written text on the computer)
2) Form: language, but some passages may also contain non-language

3) Language: English (target)

4) Length
   a) Passage: 1 long passage (approximately 450-500 words)
   b) Incomplete outline: 1 page, based on the passage

5) Vehicle: reproduced

6) Degree of speededness: relatively unspeeded—generous time limit provided

b. Type: input for interpretation

1) Text: passages

2) Incomplete outline following the structure of the passage with a word/phrase/sentence for each level of information

2) Language of input

   a) Organizational characteristics: as occurs in passages from textbooks/course readers/website assigned in the syllabus
      (1) Grammatical
         (a) Passages
            (1) Morphology and syntax: wide range of organized structures
            (2) Vocabulary: wide range of general and technical vocabulary
            (3) Graphology: typewritten
         (b) Incomplete outline: grammar and vocabulary similar to that in passages, but in outline format
      (2) Textual (cohesion and organization)
         (a) Passages: wide range of cohesive devices and rhetorical organizational patterns, including narration, description, definition, classification, comparison and contrast, and argumentation
         (b) Incomplete outline: word/phrase/sentence level, all related to be an outline
   b) Pragmatic characteristics: as occurs in actual passages from textbooks/course readers/website assigned in the syllabus
      (1) Functional
         (a) Passage: ideational, heuristic, and manipulative, possibly some imaginative
         (b) Incomplete outline: ideational and heuristic
      (2) Sociolinguistic (passage and incomplete outline): standard dialect, formal/informal register, natural, varied on cultural references and figurative language
d. Topical characteristics (passage and incomplete outline): academic, drawn from textbooks from introductory courses meeting students General Education requirement

2. Characteristics of the expected response
   a. Format
      1) Channel: visual
      2) Form: language
      3) Language: English (target)
      4) Length: 1–3 words/phrases per blank in the incomplete outline
      5) Type: limited production
      6) Degree of speededness: relatively unspeeded—generous time limit provided
   b. Language characteristics
      1) Organizational characteristics: vocabulary similar to that in texts and items; morphology and syntax standard English, but not scored; graphology: typewritten
      2) Pragmatic characteristics: mostly same as passages and item stems, plus some need for appropriate register use; little need likely for manipulative functions
   c. Topical characteristics: same as passages and item stems

3. Relationship between input and expected response and type of interaction
   a. Type of external interactiveness: nonreciprocal
   b. Scope of relationship: both broad and narrow—narrow in that specific pieces of information must be provided, and broad in that the relationship of those pieces to the text as a whole must be kept in mind by test takers
   c. Directness of relationship: direct, aside from ability to organize and categorize rhetorically

D. Procedures for recording test takers’ performance (Scoring method)
   1. Criteria for correctness: partial credit based on acceptability of content, not linguistic accuracy. These criteria were applied to each individual item as described below.
   2. Scoring procedures: scored by computer matching regular expressions with list of acceptable answers compiled by human raters. (See example scoring algorithm below.)
   3. Explicitness of criteria and procedures: criteria explicitly explained, scoring procedure described in general terms, with instructions
   4. Scorers or raters (during tryout): instructors and research assistants (RAs) in ESL reading program assigned to test development team. Computer scoring algorithm (during operational use)
E. Instructions for responding to the assessment task (See example test task below.)

III. Procedures and formats for reporting assessment records

A. Procedures

1. Assessment reports are emailed to test takers’ secure university email addresses within 24 hours after they have taken the test.

2. Assessment reports are reported on a secure password-protected website accessible only to university administrators.

B. Formats

1. Exempt/non-exempt decision

2. For those students who were not exempted from the ESL reading course, the report included the URL for a webpage that provided a complete description of the course, along with procedures for registering, and procedures for challenging or appealing the decision.

Discussion

With the earlier paper and pencil test format, students were sent written reports that included their score, the placement decision, and a description of the ESL reading course. The assessment reports from the web-based test, on the other hand, include only the exempt/non-exempt decision.

EXAMPLE ASSESSMENT TASK

The incomplete outline task was presented via the web in secure computer labs on the university campus. Each test taker sat at a computer terminal and viewed the instructions and the task on-screen. The instructions for completing the tasks were given in two parts. First a general introduction to the types of blanks test takers would be asked to complete were provided, as in Figure P2.1 below.
Web-Based ESL Placement Exam

General Directions

This test is intended to measure your academic English ability. It will test both your reading and listening ability.

When you see text with a green background, as on this page, it contains directions.

Most of the questions on this test will require you to type your answers into a text box. Some boxes will be only one line long, like this:


Others will be several lines long, like this:


You will also see text boxes for one-word answers, which will appear in the middle of a line of text.

Finally, there may be some multiple choice items on the test which require you to click on one of several radio buttons.

- A. First answer
- B. Second answer
- C. Third answer
- D. Fourth answer

Click here to continue.

Figure P2.: Initial instructions for the ESL incomplete outline task

This was followed by a short tutorial which included an example of what the headings in an outline look like, as in Figure P2.2 below.
Outline Tutorial

Some parts of this test will involve outlines. Because not all students are familiar with outlines, we have included a brief explanation of what outlines are and how they work.

An outline is a way to organize the important information in a passage. The most important information is marked with a Roman numeral (I., II., III., IV., etc.):

I. First important point
II. Second important point

Secondary points are put under these main points, and are marked with capital letters (A., B., C., etc.):

I. First important point
   A. Secondary point #1
   B. Secondary point #2
II. Second important point
   A. Secondary point #1
   B. Secondary point #2

Details supporting the secondary points are marked with Arabic numerals (1., 2., 3., etc.). Thus:

I. First important point
   A. Secondary point #1
      1. Supporting detail #1
      2. Supporting detail #2
   B. Secondary point #2
      1. Supporting detail #1
      2. Supporting detail #2
II. Second important point
   A. Secondary point #1
      1. Supporting detail #1
      2. Supporting detail #2
   B. Secondary point #2
      1. Supporting detail #1
      2. Supporting detail #2

Figure P2.2  Example outline headings in tutorial for the incomplete outline task

This was followed by a short passage and the corresponding completed outline, as illustrated in Figure P2.3 below.
**Outline Tutorial Directions:** Read the information on the right. The reading passage is included to help demonstrate how an outline can be made of a reading passage.

<table>
<thead>
<tr>
<th>Types of Marketing Research</th>
<th>If we make an outline for the short passage on the left, it will probably look like this:</th>
</tr>
</thead>
</table>
| 1. When the researcher knows relatively little about the issue being investigated, exploratory research designs are most appropriate. Exploratory research designs provide the marketing researcher with ideas and insights about a broad or relatively vague problem. Such designs allow a more precise statement of the problem to be formulated, which in turn will allow causal or descriptive research designs to be used. | I. Exploratory research designs  
A. Appropriate when the researcher knows relatively little about the issue being investigated  
B. Provide the marketing researcher with ideas and insights about a broad or relatively vague problem  
   1. Allow a more precise statement of the problem to be formulated  
   2. Thereby allows causal or descriptive research designs to be used |
| 2. When the researcher knows something about the problem being addressed, descriptive research designs are often employed. Descriptive research designs generally involve attempts to determine the frequency with which something happens or the extent to which two or more variables are related. | II. Descriptive research designs  
A. Appropriate when the researcher knows something about the problem being addressed  
B. Generally involve attempts to determine the frequency with which something happens or the extent to which two or more variables are related |
| 3. When the researcher wants to test specific hypotheses about those factors thought to contribute to the problem being investigated, causal research designs are often employed. Causal research designs generally involve attempts to determine the extent to which changes in one variable cause changes in some other variable. | III. Causal research designs  
A. Appropriate when the researcher wants to test specific hypotheses about those factors thought to contribute to the problem being investigated  
B. Generally involve attempts to determine the extent to which changes in one variable cause changes in some other variable |

Because this tutorial uses a very short reading passage, almost all of the details in the passage are included in the outline. Outlines of longer passages generally do not include every detail, and only give the more important points.

[Click here to take a practice reading section.](#)

**Figure P2.3 Example reading passage and the corresponding completed outline**

The last screen in the tutorial presented test takers with a short passage and an incomplete outline, which they could complete, but their answers were not recorded. After the tutorial, test takers were presented with an incomplete outline task, as illustrated in Figure P2.4 below.
In order to complete the incomplete outline tasks, students were able to size their own computer screen to their desired size, and to scroll either the reading passage or the incomplete outline up and down while filling in the blanks. When they hit the ‘Submit’ button, they were shown the following message:

If you wish to go back and check your answers, click the ‘Go back’ button. If you want to submit your answers and go on to the next part of the test, click the ‘Submit’ button. You will not be able to go back to this part of the test once you submit your answers.

EXAMPLE SCORING ALGORITHM

For the example assessment above, the regular expression for scoring items II. C and II. C.1 and 2, using the computer scoring algorithm, were developed in several steps. First, the test
developers provided complete answers for these three items, as follows. (Note that test takers could simply copy the relevant parts of the second paragraph into the blanks. However, the fact that they could select these specific parts and put them into the correct level in the incomplete outline was viewed as evidence of knowledge of rhetorical organization.)

II. C.  *when the researcher knows relatively little about the issue being investigated, exploratory designs are most appropriate.*

1. *provide the marketing researcher with ideas and insights about a broad or relatively vague problem*

2. *Such designs allow a more precise statement of the problem to be formulated.*

Next, the development team needed to pare down these answers to what they believed to be the minimal information required to indicate comprehension. This was necessary because the criterion for correctness was accuracy of the content with respect to the reading passage, and not linguistic accuracy. This led to the following shortened answers:

II. C.  *exploratory designs; appropriate.*

All the other information in the ‘complete’ answer was regarded as redundant (*knows relatively little about the issue being investigated* essentially repeats what is given in *research problem is rather broad or vague*).

1. *provide ideas and insights*

   (The phrases *the marketing researcher* and *about a broad or relatively vague problem* were also redundant with other information in the passage.)

2. *allow precise statement of the problem to be formulated.*

   (The phrase *such designs* repeats what is in II. C)

The next step was to determine the exact words that would be considered correct, and how many points to count for each answer. This led to the following final scoring key:

II. C.  *exploratory* (1), *designs* (1), *appropriate* (1) Maximum: 3 points

1. *provide ideas* (1), *provide insights* (1), *provide ideas insights* (2) Maximum 2 points

2. *allow precise statement problem* (1 point)

In order to accommodate a whole range of possibly correct answers, the development team then created the following regular expressions.

<table>
<thead>
<tr>
<th>Item number</th>
<th>Regex</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>II. C</td>
<td>[Ee]xploratory\b</td>
<td>1</td>
</tr>
<tr>
<td>II. C</td>
<td>[Dd]esign\b</td>
<td>[Dd]esigns\b</td>
</tr>
<tr>
<td>II. C</td>
<td>[Aa]ppropriate\b</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>([Pp]rovide).*((idea\b)</td>
<td>(ideas\b))</td>
</tr>
</tbody>
</table>
Since responses were not scored according to grammatical accuracy, these regular expressions allow both singular and plural forms as an acceptable answer. For example, *design* and *designs* will both be scored as acceptable answers.

**Reference**

Project 3

University elementary Modern Chinese speaking assessment for progress, learning, and instruction¹

SETTING

Native English-speaking students at a large US university are studying Elementary Modern Chinese. There are multiple sections of this course, and multiple Teaching Assistants (TAs) as instructors. The instructors on the course are native Chinese-speaking graduate students. The course supervisor is a native speaker of Modern Chinese, an experienced language teacher and supervisor, and has a Ph.D. in Chinese linguistics. She would like the instructors to test their students’ ability to speak Chinese in order to assign course grades and provide feedback to students on their achievement of course objectives.

The supervisor believes that a formal speaking test administered on a program-wide basis several times each term will provide the program with a way to obtain this information systematically in all sections of the course. She also believes that such assessments will have a positive impact on the way instructors teach and students study. She thus needs to decide how best to obtain this information for these purposes. In this example, the ‘test developer’ consists of a team of individuals including the Chinese course supervisor and the classroom instructors, with all members of the team assuming a collective responsibility for the test development, use, and justification. For the formative decisions at the level of the classroom, the students and course instructors are the decision makers. For the summative decisions about students’ grades, course instructors are the decision makers. For decisions about curriculum and staffing, the course coordinator is the decision maker.

ASSESSMENT USE ARGUMENT

Consequences

Claim 1: The consequences of using the mini-speech tests and of the decisions based on them will be beneficial for students, the instructors, and the course supervisor.

Stakeholders

i. Native English-speaking students at a large US university who are studying Elementary Modern Chinese.

ii. Instructors in the course, who are native Chinese-speaking graduate students in the Departments of Asian Languages and Literatures and Applied Linguistics.

iii. The course supervisor, who is a native speaker of Modern Chinese, an experienced language teacher and supervisor, and has a Ph.D. in Chinese linguistics.

¹ This project is based on materials provided by Hongyin Tao, Michelle Fu, and Huan Wang, on the Elementary Modern Chinese classes at UCLA.
Warrants: Consequences of using the checklist

A.1. The consequences of using the assessment that are specific to each stakeholder group will be beneficial. Specifically, the consequences for each stakeholder group will be as follows:

i. Students: The spoken Chinese of students in Elementary Modern Chinese classes will improve.

ii. Instructors: The classroom teaching of the instructors will improve.

iii. Course supervisor: The overall effectiveness of the course will improve.

A.2. The scores from the mini-speech tests and the course grades of individual students are treated confidentially.

A.3. The scores from the mini-speech tests and the course grades are reported in ways that are clear and understandable to students, teachers and the course supervisor.

A.4. Scores from the mini-speech tests and the course grades are reported in a timely manner.

A.5. The mini-speech test helps promote good instructional practice and effective learning, and the use of this is thus beneficial to students, instructors, the course supervisor, and the program. Because the mini-speech test requires students to speak Chinese:

i. *students* will take the course’s learning objectives in speaking seriously and organize their study accordingly, so as to optimize the time and effort spent learning to speak Chinese.

ii. *instructors* will devote more time to activities that engage students in speaking Chinese. Instructors will use the experience of giving the mini-speech test and the students’ performance on this to make adjustments in their teaching, and hence make this more effective.

iii. the *course supervisor* will observe greater attention to speaking in the classes. The course supervisor will use scores to make decisions about retaining and hiring instructors, and about changes in the syllabus that will improve the course as a whole.

Warrants: Consequences of the decisions that are made

B1. The consequences of using scores from the mini-speech test to make progress/not progress decision about students in the course are beneficial to students, instructors, and the course supervisor.

i. Because the scores from the mini-speech test will be used to determine whether students pass to the next higher course, students will take the learning objectives of the course seriously and organize their study accordingly, so as to optimize the time and effort spent studying Chinese, and hence mastering the course objectives and improving their spoken Chinese.

ii. Because the scores from the mini-speech test will be used to determine whether students pass to the next higher course, and because these will be part of their evaluation, instructors will make adjustments in their teaching, and hence make this more effective.
iii. Because the scores from the mini-speech test will be used to determine whether students pass to the next higher course, and because these will be part of the instructors’ evaluations, the course supervisor will observe greater attention to speaking in the classes.

**Rebuttal:** The consequences of making progress/not progress decisions about students will be detrimental for students who are erroneously assigned the incorrect grade.

i. *False positives:* (allowing students who should not progress to progress) can result in students receiving unrealistic feedback on their abilities; students may move on to courses for which they are unprepared.

- Possible remedies for false *positives*:
  1. Use additional information, such as class participation, and other speaking assignments, to assign grades.
  2. Provide additional tutoring or small group work to help students practice their speaking.

ii. *False negatives:* (not letting students progress who should progress) can result in students feeling demotivated.

- Possible remedies for false *negatives*:
  1. Use additional information, such as class participation, and other speaking assignments, to assign grades.
  2. Provide additional tutoring or small group work to help students practice their speaking.
Decisions

Claim 2: The decisions to be made reflect the university’s regulations and values, and the common practice and values of the academic community, including students, teachers, and administrators, and are equitable to all students in the course.

The specific decisions to be made, the stakeholders who will be affected by these, and the individuals responsible for making these decisions are given in Table P3.1 below:

<table>
<thead>
<tr>
<th>Decision</th>
<th>Individual(s) responsible for making the decision</th>
<th>Stakeholders who will be affected by the decision</th>
</tr>
</thead>
</table>
| Decide which students progress to the next course partly on the basis of scores on the mini-speech test. | Instructors  
Course supervisor | Students |
| Adjust individual learning activities to focus on speaking. | Students  
Instructors | Students, instructors |
| Adjust teaching style, focus and activities. | Classroom instructors | Instructors, students |
| Revise curriculum. | Course supervisor  
Instructors | Instructors, students |
| Maintain or change instructional staff. | Course supervisor | Course supervisor, instructors, students |

Table P3.1  Decisions, decision-makers, and affected stakeholders

Warrants: Values sensitivity

A1. University and departmental regulations, ‘customary practice’ among the faculty in the Department, and the opinions of students were carefully and critically considered in the kinds of decisions that are to be made.

A2. University and departmental regulations, ‘customary practice’ among the faculty in the Department, and the opinions of students were carefully and critically considered in determining the relative seriousness of false positive and false negative classification errors.

A3. Because the scores on the mini-speech tests comprise just one part of students’ course grades, no cut scores are needed.

Warrants: Equitability

B1. Decisions about which students will progress to the next course are made only according to the procedures that have been established, and not on the basis of other considerations.
B2. Students and instructors are fully informed of the procedures for making decisions about which students will progress to the next course whether these procedures are actually followed in making these decisions.

B3. Students have equal opportunities to learn or acquire speaking ability in Modern Chinese.

Interpretations

Claim 3: The interpretations about language ability in Modern Chinese are meaningful with respect to the course syllabus, impartial to the students in the course, generalizable to tasks in the language instructional domain, relevant to the different kinds of decisions to be made, and sufficient for the different kinds of decisions to be made.

Warrants: Meaningfulness

A1. The definition of the construct includes 1) freedom from grammatical and pronunciation errors, 2) organization, 3) spontaneity, creativity, and fluency, 4) vocabulary and sentence patterns used, and 5) ability to stay within allotted time. These abilities differ from other possible construct definitions, such as knowledge of the Chinese writing system. This definition of the construct is based on the course syllabus.

A2. The test task specifications clearly specify the mini-speech test procedure that will elicit speaking performance from which we can make inferences about 1) freedom from grammatical and pronunciation errors, 2) organization, 3) spontaneity, creativity, and fluency, 4) vocabulary and sentence patterns used, and 5) ability to stay within allotted time.

A3. The procedures for administering the mini-speech test are followed consistently across classes and different occasions.

A4. The rating scales that are used to score the students’ performance on the mini-speech test reflect the five parts of the definition of language ability in Chinese as articulated in Warrant A2.

A5. The mini-speech test tasks engage the students in speaking Chinese.

A6. The scores from the mini-speech test can be interpreted as indicators of the students’ language ability in Chinese as articulated in Warrant A2.

A7. The instructors communicate the definition of language ability in Chinese articulated in Warrant A2 in terms that are clearly understandable to the students.

Warrants: Impartiality

B1. The mini-speech test tasks do not include response formats or content that may either favor or disfavor some students.

B2. The assessment tasks do not include content that may be offensive (topically, culturally, or linguistically inappropriate) to some test takers.

B3. Students are treated impartially during all aspects of the administration of the mini-speech tests.
B4a. Students have equal access to information about the test content and procedures, and have equal opportunity to prepare for the mini-speech test.

B4b. Students have equal access to the test. It is given as part of the course at no extra cost, in the Chinese classrooms, under conditions with which the students are familiar.

B4c. Students have equal opportunity to demonstrate their Chinese speaking ability.

B5. Interpretations of language ability in Chinese based on the mini-speech tests are equally meaningful across different classes in the Modern Chinese course.

Warrants: Generalizability

C1. The characteristics (e.g. input, expected response, type of external interaction) of the mini-speech test tasks correspond closely to those of instructional tasks.

C2. The criteria and procedures for evaluating the responses to the mini-speech test tasks correspond closely to those that instructors have identified as important for assessing performance in other speaking tasks in the instructional setting.

Warrant: Relevance

D. The interpretations of speaking based on the mini-speech tests provide information that is needed by the course director, instructors, and students to make the decisions articulated in Table P3.1 above.

Warrant: Sufficiency

E. The interpretation of language ability in Chinese, based on the mini-speech test scores, is not sufficient to make the decisions listed in Table P3.1. That is, this interpretation needs to be supplemented with other information gathered by the class instructors.

Assessment records

Claim 4: Scores obtained from students’ performance on the mini-speech tests are consistent across different tasks, administrations, instructors/raters, and classes of students.

Warrants: Consistency

1. Administrative procedures are followed consistently across different occasions, and for all classes and students.

2. The criteria and procedures for rating students’ performance on the mini-speech task are well specified and adhered to.

3. Instructors undergo training at the beginning of each school term.

4. Instructors are trained to avoid bias for or against different groups of students.

5. (Not relevant to this test, since the mini-speech test is essentially a single task.)

6. Ratings of different instructors are consistent.

7. Different ratings by the same instructor are consistent.

8. Scores from different forms of the mini-speech test are consistent.
9. Scores from different administrations of the test are consistent.
10. Ratings of students’ performance on the mini-speech test are of comparable consistency across different classes and subgroups of students in the elementary Modern Chinese course.

Example Oral “Mini Speech” Tests

Oral Tests
You will have five oral tests in Chinese.

GROUP Tests
In the first three oral tests you will be working with another student to create dialogues by adapting the current dialogues you are learning. Each dialogue will be three minutes in length. You are expected to read aloud the created dialogues in front of the class. The teacher will then ask you a few questions related to your dialogue. The suggested topics for the first three mini speech are as follows:
   a. The weather
   b. Transportation
   c. Dining

Individual MINI SPEECH
In the last two oral tests you will work alone. You are required to give a two-minute mini speech in each test. The suggested topics for the last two mini speech tests are:
   a. At the Library
   b. Asking Directions
   c. Birthday Party

You may also choose your own topic for your speech. No matter what topic you may choose, your speech should contain the current vocabulary and sentence structures covered in class up to the time you do your mini speech.

When you present your mini speech in class, please do not read directly from your notes. You are expected to deliver the speech in a natural manner with only occasional help from some cue cards.

To prepare for the oral tests, you need to listen to the audio files on the e-campus regularly, practice reading aloud the dialogues you are learning on a daily basis, and seek help from your instructors by attending the classes and going to their office hours.

Please refer to the course syllabus for the criteria to be used to assess your oral work.
Grading criteria for GROUP test (equal weights for the following 7 aspects)

1. collective efforts (every member gets equal chance to speak)
2. grammaticality (correct use of vocabulary and sentence structures)
3. pronunciation (initials, finals, tones, intonations, voice is clear and loud enough for everyone in the room to hear)
4. content (well organized, use of rich vocabulary and a variety of patterns, creative, interesting & clear)
5. naturalness, fluency (well rehearsed, no hesitation, not constantly looking at the notes)
6. time management (5-6 minutes for a group of 4-5)
7. script (title page, names of members, no typos/wrong sentences, etc)

Grading criteria for MINI SPEECH (equal weights for the following 7 aspects)

1. pronunciation
   a. initials
   b. finals
   c. tones
   d. intonations
2. correct use of the vocabulary Learned in this quarter
3. correct use of grammatical structures learned in this quarter
4. content
   a. clear and well organized
   b. rich vocabulary
   c. a variety of sentence structures from different lessons covered in this quarter
   d. creative and interesting
5. naturalness
   a. able to speak fluently
   b. able to speak without constantly referring to the notes
   c. Voice is clear and loud enough for everyone to hear.
6. time management (about 2 minutes, give or take a few seconds)