Table of Contents

Spaan Fellowship Information.................................................................iv

Previous Volume Article Index ...............................................................v

Beverly A. Baker
Conflicting Genre Expectations in a High-Stakes Writing Test
for Teacher Certification in Quebec.........................................................1

Ute Knoch
Collaborating with ESP Stakeholders in Rating Scale Validation:
The Case of the ICAO Rating Scale ......................................................21

Atta Gebril
Lia Plakans
Investigating Source Use, Discourse Features, and Process
in Integrated Writing Tests..................................................................47

Tian Song
Investigating Different Item Response Models in Equating
the Examination for the Certificate of Proficiency in English (ECPE)....85

Luke Harding
Kerry Ryan
Decision Making in Marking Open-Ended Listening Test Items:
The Case of the OET.................................................................99
The University of Michigan

SPAAN FELLOWSHIP

FOR STUDIES IN SECOND OR FOREIGN LANGUAGE ASSESSMENT

In recognition of Mary Spaan’s contributions to the field of language assessment for more than three decades at the University of Michigan, the English Language Institute has initiated the Spaan Fellowship Fund to provide financial support for those wishing to carry out research projects related to second or foreign language assessment. The Spaan Fellowship has been created to fund up to six annual awards of $4,000 each.

These fellowships are offered to cover the cost of data collection and analyses. Fellows will submit a paper at the end of the year to be published in the *Spaan Fellow Working Papers*. Fellows may make use of the English Language Institute’s resources to carry out a research project in second or foreign language assessment. These resources include the ELI Testing and Certification Division’s extensive archival test data (ECCE, ECPE, and MELAB) and the Michigan Corpus of Academic Spoken English (MICASE).

Applications are welcome from anyone with a research and development background related to second or foreign language assessment, especially those interested in analyzing some aspect of the English Language Institute’s suite of tests (MELAB, ECPE, ECCE, or other ELI test publications). Spaan Fellows are likely to be second or foreign language assessment specialists or teachers who carry out test development or prepare testing-related research articles and dissertations; doctoral graduate students from one of Michigan’s universities who are studying linguistics, education, psychology, measurement, or related fields; and doctoral graduate students in foreign language assessment or measurement from elsewhere in the United States or abroad.

For more information about the Spaan Fellowship, please visit our website:

[www.lsa.umich.edu/eli/research/spaan](http://www.lsa.umich.edu/eli/research/spaan)
<table>
<thead>
<tr>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigating the Construct Validity of the Grammar and Vocabulary Section and the</td>
<td></td>
</tr>
<tr>
<td>Listening Section of the ECCE: Lexico-Grammatical Ability as a Predictor of L2 Listening</td>
<td>5, 37–78</td>
</tr>
<tr>
<td>Ability. Liao, Yen-Fen</td>
<td></td>
</tr>
<tr>
<td>Investigating the Invariance of the ECPE Factor Structure across Different Proficiency</td>
<td>6, 29–55</td>
</tr>
<tr>
<td>Levels. Römhild, Anja</td>
<td></td>
</tr>
<tr>
<td>Lexical Diversity in MELAB Writing and Speaking Task Performances. Yu, Guoxing</td>
<td>5, 79–116</td>
</tr>
<tr>
<td>Ratings of L2 Oral Performance in English: Relative Impact of Rater Characteristics</td>
<td>6, 181–205</td>
</tr>
<tr>
<td>and Acoustic Measures of Accentedness. Kang, Okim</td>
<td></td>
</tr>
<tr>
<td>Switching Constructs: On the Selection of an Appropriate Blueprint for Academic Literacy</td>
<td>2, 135–155</td>
</tr>
<tr>
<td>Assessment. Van Dyk, Tobie</td>
<td></td>
</tr>
<tr>
<td>Toward a Cognitive Processing Model of the MELAB Reading Test Item Performance. Gao,</td>
<td>4, 1–39</td>
</tr>
<tr>
<td>Lingyun</td>
<td></td>
</tr>
<tr>
<td>Underlying Factors of MELAB Listening Constructs. Eom, Minhee</td>
<td>6, 77–94</td>
</tr>
<tr>
<td>Validation and Invariance of Factor Structure of the ECPE and MELAB across Gender. Wang,</td>
<td>4, 41–56</td>
</tr>
<tr>
<td>Shudong</td>
<td></td>
</tr>
</tbody>
</table>
Conflicting Genre Expectations in a High-Stakes Writing Test for Teacher Certification in Quebec

Beverly A. Baker
McGill University

ABSTRACT All preservice teachers in faculties of education in the province of Québec are required to demonstrate proficiency in language of instruction (either English or French) as part of their certification by the Quebec Government’s Ministère de l’éducation, du loisir et du sport (MELS). In the English sector, this proficiency is currently measured by university-prepared tests that concentrate on writing skills.

I will report on an investigation of one such high-stakes English-language proficiency test written by hundreds of native and nonnative-English BEd undergraduates every year at a Montreal university. More specifically, I will discuss the written composition part of the test, focusing on the construct of writing competence it is perceived to be measuring. My analysis has revealed conflicting genre expectations (i.e., conflicting assumptions about what type of text is required to fulfill the test requirements and/or demonstrate competent writing).

Taking a social-constructivist view of genre that includes, but is not limited to, text-based linguistic analysis, I report on conflicting perceptions of the construct of writing competence—conflicts that were identified while comparing genre expectations (as revealed by test prompts, interviews with test administrators, and focus groups with test takers) with elements of successful and unsuccessful student compositions. One notable finding is that these conflicts have not necessarily impacted pass/fail decisions, which seem to be made almost exclusively on accuracy of language structures. These findings contribute to previous work on high-stakes standardized writing assessments and the role of genre knowledge in the construct of writing competence.

This study is part of a larger project that includes the development of a new common English proficiency test for preservice teachers across all English universities in Quebec.

As in many North American contexts, preservice teachers in Quebec universities form an increasingly heterogeneous group, representing many cultures and often not teaching in their first language. Partially for this reason and partially to address wider concerns about
language standards in society, the Quebec government’s Ministère de l’éducation, du loisir et du sport (MELS) has mandated that all students in faculties of education in Quebec are required to demonstrate proficiency in their language of instruction (either English or French) as part of the teacher certification process. In Quebec’s three English-medium universities, this language proficiency is currently measured by three separate university-wide or faculty-wide writing tests. Students may be granted permission to rewrite these tests a limited number of times if they fail. Ultimately, failure can result in students being streamed into additional writing courses and in some cases being expelled from their university programs altogether. Therefore, passing these tests is a requirement for both graduation and teacher certification. Such tests would therefore be considered “high stakes” tests (Ryan, 2002). They could also be considered “exit tests,” which are defined by Ruetten (1994) as tests “used to verify the writing competence of students...to gain upperclassman status or to graduate from the university” (p. 88).

Governments are increasingly mandating assessments within educational settings to further social policies (McNamara & Roever, 2006). Leki, Cumming, and Silva (2006) argue that the current trend in accountability in education has led to the practice of explicitly defining standards that are expected of students completing their degrees. These standards often include quality writing standards, and success on writing tests often serves as the proof that these standards are met. Penrod (2005) has also discussed what she refers to as increasingly “rampant” and “rabid” calls for accountability within the university system. This phenomenon of accountability testing reflects two prevalent notions in today’s educational contexts: that the state should be intimately involved in educational decision making, and that tests have the power to affect changes in learning, behavior, and public opinion (for further discussions of these issues, see Mazzeo, 2001, and Shohamy, 2001). According to Janopoulos (1992), writing proficiency tests are a response to widespread concerns that literacy levels on university campuses are declining, and that the tests themselves are designed “to reflect an institutional consensus of what minimum writing competence should be” (p. 110).

The MELS recently requested that a common English proficiency test for preservice teachers be developed to replace the three separate tests currently in use in the three English-medium Quebec universities.¹ (A common French proficiency test has been in development for some time as part of the certification process for preservice teachers in the French university system.) The development of this common English test began by evaluating one of the English proficiency tests currently being used (see Figure 1).

The Present Report: Evaluation of a Current Test

The present report is part of this first step, the evaluation of one current test, as well as a discussion of how insights from the evaluation will aid in the replacement of this current test with a new common test. The evaluation involved first ascertaining whether the current test adheres to best test practices as established by recent international standards and guidelines on language assessment (The Joint Committee on Testing Practices, 2004; International Language Testing Association, 2005). The evaluation also involved a critical comparison of the reports of all interested stakeholders regarding their perceptions of the goals of this particular assessment and the construct of language competence that the assessment had originally been put in place to measure.
This particular English proficiency test is written by all bachelor of education undergraduates at one Montreal university, usually at the end of their first semester. It has been in place for approximately four years. This test is usually written in the regular exam period, and consists of an essay component with a large choice of prompts, and a sentence-editing task. The test is graded by the test coordinator and a few colleagues together in a one-day marathon marking session. The essay component is marked by one grader, who gives it a pass or fail score. Papers who receive a failing grade on the first reading are read by a second reader, and a final pass/fail decision is made.

Approximately 10% of those who take the test fail it the first time. The majority of these students who fail take a writing course in the semester following the test, and take the test again the following fall. If students fail the test a second time they must withdraw from their program. A withdrawal from the program does not prevent students from reapplying at a later date.

Genre Perceptions and Expectations

Several issues have arisen during this current test evaluation having to do with the construct of writing competence as understood by various stakeholders involved with the test. These stakeholders include test takers, test graders, and administrators who make decisions based on the results of these tests. The bulk of the discussion here centers on the construct of writing competence in terms of the genre of text expected by all of these stakeholders.

Beck and Jeffery (2007) have examined the role that genre knowledge plays as part of the construct of writing competence. They examined the high-stakes writing assessments of high school students in three U.S. states, comparing the genre demands of the test prompts and the genre characteristics of benchmark texts used in test preparation. Their results suggested a mismatch between the two. This study has contributed to the literature in more than one fruitful way. Firstly, it was conducted under the assumption that a common understanding of the genre elements of a text cannot be separated from a common understanding of the underlying test construct. Another way this study advances knowledge in this area is through its acknowledgement of the need to analyze prompts and responses as
pairs of texts when considering genre. Freadman (2002) has stated that the term *genre* may be more usefully applied to the interaction of a pair of texts—in this case, the prompt and the response. Devitt (2000) discusses how genres do not exist independently but only in response to and in interrelation with other genres.

In their study, Beck and Jeffery (2007) analyzed prompts and responses through “identifying and tallying the rhetorical processes signaled by key words” (p. 64). In their analyses, they identified key rhetorical features as displayed by the tests. However, despite performing a genre analysis based entirely on textual elements, they acknowledged the essentially social nature of genre: “The way prompts are interpreted by students and by the teachers who prepare students to take the tests is ultimately what kind of genre a student will produce in a test situation” (pp. 64–65). This idea is intuitive to anyone involved in writing instruction and assessment. In making decisions about the type of writing to produce, we know that test takers are influenced by much more than the question in front of them, and in assigning a grade, a grader is influenced by much more than the text in front of them.

A textual analysis of prompts and responses is not the only way to establish genre expectations. It can be illuminating to incorporate direct and indirect comments by test administrators, graders and takers as they discuss the processes of creating, responding to and writing the tests. This is what a sociocultural view of genre (Miller, 1984) can bring to this study: “Genre, as redefined in rhetoric-composition in complex and myriad ways, is defined by its situation and function in a social context” (Devitt, 2000, p. 698). Viewing genre in this way means that we must consider the social context for text production, or else our understanding of the resultant text(s) will always be incomplete.

**A Sociocultural View of Genre: Towards a Working Definition**

The current study makes use of a sociocultural framework for defining genre put forward by Paré and Smart (1994) which considers that a genre profile is established through the consideration of four dimensions:

1. a set of texts
2. the composing processes used in creating the texts
3. the reading practices used to interpret them
4. the social roles performed by the writers and readers (p. 147).

The first dimension includes regularities of textual elements. As Paré and Smart (1994) affirm, “Repeated patterns in the structure, rhetorical moves and style of texts are the most readily observable aspects of genre” (p. 147). Perhaps for this reason, genre analysis often begins and ends here. However, if writing is viewed as a social action or construction, then consideration of the next three dimensions becomes necessary. In including the composing process, an analysis of genre can begin long before the prompt is even seen, and can include the assumptions made by the writer about how his or her text will be read and interpreted (Hyland, 2002). In addition, many insights can be gained in questioning how the reader (in this case, the grader) defines appropriate writing for this situation, and also in examining the social roles the writers and readers play. In this situation, for example, the evaluative function of the reading constitutes an element of genre. An important decision will be made as a result of reading this text (passing or failing the student). The decisions which the reader must make after reading the text affect the way the text is read from the outset.
Research Questions

The research questions for this part of the study address the role of genre expectations in the construct of writing competence—using the extended view of genre described above—and whether there is a common understanding by all parties of this construct of writing competence.

**Research Question 1:** Is the construct of competent writing the same as suggested by:

- the test prompts?
- the passing compositions?
- the student comments?
- the grader and administrator comments?

**Research Question 2:** If there is not a common understanding of this construct, what are the implications for the parties involved?

Research Design and Method

In order to create the complete genre profile as described above, as well as to answer the research questions, multiple sources of quantitative and qualitative data needed to be collected. Therefore, the study followed a mixed method design (Creswell & Plano Clark, 2007). Four different sources of data from different stakeholders were collected: test prompts, a sample of test compositions, student focus group interviews, and grader and administrator interviews. These four sources of data were analyzed independently and then converged during reporting of the data. However, data were not collected simultaneously, so later collection phases were affected by themes that emerged during the analysis of earlier collection phases. For example, themes revealed by initial administrator interviews informed the guiding questions for later student focus groups, and student comments influenced the textual elements chosen to examine. Therefore, it was beneficial to use elements of a triangulation as well as an exploratory mixed-methods design (Creswell & Plano Clark, 2007). This data collection and analysis process is summarized in Figure 2. All data were collected in 2008. While mostly beneficial in this study, some drawbacks to the use of this design will be covered in the discussion section.

![Figure 2: Summary of data collection and analysis](image-url)
Procedures: Data Collection and Analysis

Data Collection: Data Sources and Participants
Data for this study included:

- interviews with test graders and other administrators;
- focus group meetings with students who had already taken the test; and
- one test prompt and a sample of 11 passing and 11 failing compositions that responded
to this particular prompt.

The four data sources for this study will be presented here in the order in which they were
collected. All data were collected from May to September, 2008. Each of the data sources was
chosen in order to fully answer the research questions as well as to collect information on
each of the four dimensions of genre, as outlined in Table 1.

Table 1. Outline of Data Sources for the Study as They Relate to Genre

<table>
<thead>
<tr>
<th>Dimension of genre (Paré and Smart, 1994)</th>
<th>Data source used to examine this dimension in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>a set of texts</td>
<td>11 failing and 11 passing compositions</td>
</tr>
<tr>
<td>the composing processes used</td>
<td>student focus groups; prompts</td>
</tr>
<tr>
<td>in creating the texts</td>
<td></td>
</tr>
<tr>
<td>the reading practices used to interpret</td>
<td>interviews with administrators/graders; prompts</td>
</tr>
<tr>
<td>them</td>
<td></td>
</tr>
<tr>
<td>the social roles performed by</td>
<td>student focus groups and interviews with</td>
</tr>
<tr>
<td>the writers and readers</td>
<td>administrators/graders</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prompts; Failing and Passing Compositions
The analyzed task, which carries almost the entire grade of the test, is a 500-word
composition based on a choice of six or seven prompts. The prompt in this case consists of
two separate texts: One set of general instructions for completing the composition, located on
the top of the paper, and the instructions for each of the seven choices. Below is the prompt
chosen for analysis:

Write an essay of at least 500 words in response to one of the following topics.
Answer the question or questions you are asked and discuss the significance of
your answer. In other words, relate your personal experiences to larger social,
moral, or cultural issues.

1. Discuss ways in which one person’s actions can help to improve the world.

In this prompt, the thesis of this essay is provided (one person’s actions can change
the world) and the test taker is expected to provide a list of how one person’s actions can
change the world. Test takers also seem expected to provide an anecdote or some other link to
personal experience. Therefore, the uptake of the prompt is presumably to discuss the idea in
personal, specific terms and in more general, abstract forms relating to society in general.
Of the seven prompts of the test, this particular prompt was chosen because it was a popular choice (chosen by 18% of the sample), yet it had the highest failure rate of all seven prompts (18% of all those who chose this prompt failed it, compared to the average failure rate of 12%). A sample of 11 failing compositions and a matching sample of 11 randomly chosen passing compositions were transcribed to control for any possible handwriting effects.

**Interviews and Focus Groups**

Focus groups and interviews with student test takers, both those who pass and those who fail, are common in current studies involving high stakes literacy testing (Luce-Kapler & Klinger, 2005). Semistructured interviews (Gubrium & Holstein, 2002) were conducted with administrators and instructors at the faculty who have direct involvement with the test. The interviews were audio-recorded and transcribed. Nine participants were recruited:

- two graders and one administrator of the test;
- one member of the faculty who deals with the government body that coordinates the certification of Quebec teachers (*le Comité d’agrément des programmes de formation à l’enseignement* or CAPFE); and
- six other members of the staff and faculty who deal with the administration of the test. These administrators include those responsible for informing students about the test, those who advise students who have failed the test, and those who receive complaints about the language quality of preservice teachers. These complaints come primarily from student teacher supervisors and cooperating teachers who observe student teachers during their field experiences.

Preliminary analysis of themes in the interview data were later explored in focus groups with the student test takers (see Denscombe, 2003 and Puchta & Potter, 2004 for descriptions of focus groups used in this way).³ The focus group meetings lasted approximately one hour each.

The participants for the focus groups were preservice teachers who have written the test at least once in the previous two years. There were two groups and 11 students in total:

- three students who failed the test and were waiting to retake it;
- three students who failed the test, retook it, and subsequently passed it; and
- six students who passed the test on the first attempt.

See Appendices A and B for the guiding questions developed for the interviews and focus groups. Although many issues were addressed, discussion here will be limited to the responses that provide further insight into the research questions, namely:

- references to the type of writing or genre which is expected or produced; and
- the conceptualization of the test construct (i.e., what the test is perceived as measuring by all parties).

**Data Analysis**

The instruments and procedure for the quantitative analysis of the passing and failing compositions will be presented first, followed by the instruments and procedure for the
qualitative analysis of the prompts, interview data and focus group data. The results will follow the same order of presentation.

Analysis of Student Papers

The sample, consisting of 11 failing and 11 passing papers, was coded for the elements described below. The choice of elements for the textual analysis was made based on several criteria: elements that have been used previously in Beck and Jeffrey (2007), elements that have been identified in the literature on academic writing and assessment, and elements that were suggested by the interviews and focus groups. Four elements were chosen in total: text length, rhetorical moves, academic vocabulary, and percentage surface error. The goal was to identify independent elements that, while not exhaustive, showed evidence of contributing to the construct of quality writing.

1. **Text Length** (word count) was chosen as an element for two reasons. Firstly, during interviews, sufficiency of text was mentioned by the test coordinator as an important element to consider in grading. Because this coordinator was the final grader and made all pass/fail decisions, it is reasonable to assume that text length may indeed play a role in the construct of quality writing for this context. In addition, previous research has suggested that longer texts in these situations may receive higher marks (Cohen, 1994).

2. **Rhetorical Moves** were classified based on the same keywords and structural/register features as Beck and Jeffrey (2007), themselves making use of a description by Glasswell et al. (2001) of common features in the traditional genres associated with school writing. Of the many rhetorical moves identified in this previous work, only three were identified in the student compositions studied here: explanation, narrative, and argument. A summary of this classification, with a few examples from the compositions, can be seen in Table 2 on the following page.

3. **Academic Vocabulary**: Morris & Cobb (2004) found that vocabulary profiles of beginning preservice teachers in English as a second language (ESL) predicted their future academic performance. Their data suggest that access to higher numbers of academic words by students can result in better performance on academic tasks; thus, vocabulary level may be a key part of the construct of writing competence for this context. As Cobb and Horst (2004) suggest, “It is arguable that AWL [academic wordlist] items are not only important because they increase text coverage, but also because of the intellectual work they do in academic texts” (p. 21).

Student texts were therefore run through *Vocabprofiler* (Cobb, 2002), a lexical profiler computer program which classifies each word in a text according to its frequency or its inclusion in the academic wordlist or AWL (Coxhead, 2000). The number of academic word types, as a percentage of total words, was compared for each text. Word types (unique instances of a word) were chosen for analysis, as opposed to word tokens (total word counts including repetitions). For example,
one text contained the word *positive* and its variants 43 times, or 7.5% of the total words of the text. As this word is on the AWL, counting each instance of this word family would have had the effect of making a text seem more academic than it actually was: her percentage of word *tokens* was 14.5% but her percentage of word *types* was only 3.6%.

<table>
<thead>
<tr>
<th>Rhetorical move</th>
<th>Use in this context</th>
<th>Structural/register features with examples from compositions</th>
</tr>
</thead>
</table>
| Explanation     | Explaining or describing how one person’s actions can improve the world | • The use of *how*: “This is how you can improve the world as an individual…”  
• The use of steps: “first…second”  
• The imperative and modals for advice and suggestions: “To improve the world you must…”; “You have to excel as far as humanly possible…” |
| Narrative       | Recounting a first- or second-hand anecdote as support for how one person’s actions can improve the world | • Use of past tense: “I made positive changes for myself as a young adult…”  
• Use of introductory elements such as “for example,” or “personally”: “For example, each year I give fifty dollars to Christmas baskets.”  
• Orientation to characters and chronological organization: “I remember my third grade teacher, ‘T,’ had a demon child in her class: me!” |
| Argument        | Thesis statement (*one person’s actions can improve the world*) plus evidence, then restatement of thesis | • abstract concepts, especially about groups of people: “society,” “our world,” “leaders,” “individuals”  
• Thesis in declarative: “People’s actions can contribute positively to society”  
• modals and other hedges as associated with academic discourse: “Leaders can be influential such as Gandhi…”; “If a person must drive, it would be better for the environment if he or she chose to drive a smaller car…”  
• Explicit mention of essay conventions: “For the purposes of this essay I will draw from my experiences…to illustrate the claim that I have made above.” |

4. *Percentage Surface Error* was included for many reasons. In Hamp-Lyon’s (2002) discussion of the history of large-scale language tests, she outlines how the focus of writing has shifted from its use as a rhetorical tool to a way of demonstrating an accurate control of the structures of academic language. Salient surface errors (mostly word level errors) in writing tests are suspected of playing too great a role when papers are graded holistically (Hamp Lyons, 1995).
grading process used here was *general impression marking* (Weigle, 2002), which provides even less guidance than holistic grading schemes and has been widely discredited in the field of writing assessment. As over four hundred papers were corrected in one day by a small team, they were clearly read quickly. Refer to the graders’ own descriptions of their grading process below:

**Interviewee 8 (grader):** There are just reams of these little booklets to read and [you] just read an essay and put your impression on it.

**Interviewee 6 (grader):** I was the guy with the red light switch and as I was going through if I saw one, I’d turn the red light on and put it in the stack.

With this grading procedure, it is reasonable to assume that grammar errors, by their very saliency, may indeed be playing a major factor in decision making by graders. In addition, comments by test graders indicated that the absence of surface error was synonymous to them with quality writing (at least, the quality being considered in this test). All comments made about language quality seemed to be accompanied with examples of structural accuracy:

**Interviewee 6 (grader):** Imagine this person up in front of a crowd of young people, writing on the board, filling that board with all kinds of grammatical errors and spelling mistakes…The teachers must have a basic competency… sometimes it’s just glaring. You know the syntax is so bad, the thought structure, the progression is so awful, you can’t follow it. It’s laced with you know, spelling errors or you know, the person really does not have a command of the language.

**Interviewee 7 (administrator):** It seems to me that (students) do run into problems that seem to be either that their written English is full of errors, they make mistakes in spelling...they’re making all these mistakes in messages home to parents, when they make comments to students on their work, not picking up on student errors in work, or preparing classroom (sic) say an overhead or PowerPoint and it’s full of errors.

Therefore, grammar errors were counted and divided by the total number of words in the text. This error density calculation was used because of the large range of word length in the papers (346 to 778 wds). The types of errors were chosen because they were frequent in the students’ compositions, they were mentioned as examples by the graders in interviews, they were circled or otherwise identified by graders in the texts, and they were uncontroversial errors (not differences in style). The errors counted were in spelling, misused homophones or other misused words, punctuation (misplaced comma, apostrophe), capital letters, missing words, verb forms, subject-verb agreement, pronoun-antecedent agreement, and article use.
Multivariate analysis of variance (MANOVA) was then used for analysis of these four dependent variables (test length, AWL types, rhetorical moves, and percentage surface error) on the chosen sample of passing and failing papers, to examine to what extent each variable contributed to a common underlying construct of quality writing. The qualitative data from the interviews and focus groups informed further data collection for the quantitative textual analysis of the compositions. This procedure is typical of an exploratory model for combining methods in research (Cresswell & Plano Clark, 2007).

Analysis of Prompt Language and Interview and Focus Group Data

This analysis was performed with several techniques associated with grounded theory (Glaser & Strauss, 1967). Coding categories were not decided ahead of time but instead emerged from the data (Glaser & Strauss, 1967; Strauss & Corbin, 1998). Category labels were created based on specific language used by participants (e.g., essay, grammar) and more descriptive sociological constructs (e.g., perceptions, beliefs).

The data analysis procedure involved reading the transcripts several times and discussing possible themes with another reader (the transcriber), then reading again while coding each turn/response, in an attempt to consistently apply the criteria of selection (Berg, 2007). Some categories were merged using the “constant comparative method” (Glaser & Strauss, 1967), which means constantly reviewing category make-up by what it contained and by new elements as they were added, in an iterative process. For example, an initial category, entitled “our future teachers,” included all references by all participants to test takers as future members of the teaching profession. Another category included depictions of test takers as future models of proper English in society. Some comments clearly overlapped between the two categories. Following discussion with the transcriber and rereading, these categories were collapsed and then redivided. Comments involving the inherent responsibility of future teachers to be models of proper English were given one label, and another category was created for comments about future teachers without reference to language quality.

Results

Quantitative analysis of the passing and failing compositions will be presented first, followed by the results for the qualitative analysis of the prompts, interview data, and focus group data. Results here provide insight into each of Paré and Smart’s (1994) four dimensions of genre: textual features, writing processes of the test takers, reading processes of the graders, and social roles played by all parties.

Quantitative Results

A one-way MANOVA was conducted in order to determine the grade effect (i.e., “pass” or “fail”) on each of the textual characteristics identified. In other words, the MANOVA was conducted to compare group means across all the dependent variables: text length, rhetorical moves (argument, narrative, or explanation), academic vocabulary, and percentage surface error.

Significant differences were found between the two groups for only one variable: percentage language errors. Wilk’s $\Lambda = .32$, $F(6, 13) = 4.54$, $p = .011$. The multivariate $\eta^2$
based on Wilk’s Λ was 0.68. Means and standard deviations on dependent variables for pass and fail groups are presented in Table 3.

Table 3. Means and Standard Deviations on Dependent Variables for Pass and Fail Groups

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Error</th>
<th>AWL types</th>
<th>argue</th>
<th>explain</th>
<th>narrative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M  SD</td>
<td>M  SD</td>
<td>M  SD</td>
<td>M  SD</td>
<td>M  SD</td>
<td>M  SD</td>
</tr>
<tr>
<td>PASS</td>
<td>588</td>
<td>52</td>
<td>.006*</td>
<td>.004*</td>
<td>3.94</td>
<td>1.46</td>
</tr>
<tr>
<td>FAIL</td>
<td>574</td>
<td>123</td>
<td>.036*</td>
<td>.029*</td>
<td>4.40</td>
<td>1.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.30</td>
<td>4.32</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.30</td>
<td>2.36</td>
<td>5.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.94</td>
<td>2.10</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.30</td>
<td>2.38</td>
<td>2.38</td>
</tr>
</tbody>
</table>

*p < .05

In summary, text length was not seen to be a contributor to pass/fail decisions, and neither was percentage academic vocabulary or differences in rhetorical patterns. Passing papers could be short or long. One passing paper was just 416 words—well below the 500-word minimum stated in the prompt. The passing papers did not make use of a more academic level of vocabulary than the failing papers. Overall, the average percentage of AWL words was 6.31% for the passing papers and 7.31% for the failing papers. This is well below the 12 to 15% level associated with academic level texts as discussed in the literature on vocabulary profiles, and more at the level of expository writing intended for a general audience (Cobb & Horst, 2004). This may have to do with the nature of the prompt, where students are expected to discuss personal information in the form of anecdotes, or possibly to provide the reader with instructions on how to change the world as an individual. It could be argued that these functions do not require an academic register.

In addition, no pattern emerged with either the passing or the failing papers in terms of the rhetorical functions employed. Data suggests that the rhetorical structure of the passing papers did not necessarily have to resemble the structure suggested by the prompt. For example, although the prompt explicitly instructs test takers to relate their personal experiences to their arguments in their responses, compositions which failed to do so (i.e., which contained no narrative elements) did not necessarily fail.

Qualitative Results

Regarding the student focus groups, of the total of 32 comments in reference to the genre or the writing construct of the assessment, only one discussed rhetorical elements of composing (such as the successful transmission of information, or of appropriate argumentation, etc). In fact, when asked what was expected of them in their compositions, their 32 comments fell into three categories:

1. Fifteen discussed expectations in terms of accurate control of language structures. Of these, only two comments mentioned organisation or vocabulary variety; the rest mentioned the words grammar, punctuation, or spelling, or used words like perfect or correct. Below are two examples:

   **Student 5**: Well, the big thing, just grammar.

   **Student 3**: Like you’re writing a paragraph and it has to be perfect writing.
2. Eight of these comments were unable to be classified as they made vague reference to “good writing” without further explanation. Students appeared to be making an appeal to a common understanding of what good writing is:

   **Interviewer:** Well, I guess you could imagine for the grammar exercises what they would be grading, but if you’re writing a text, what do you think they were going to grade it on?

   **Student 4:** Well, to see if you can write.

3. Nine of the responses made explicit reference to the genre of the essay or mentioned structures normally associated with a five-paragraph essay structure (like introductions or conclusions). However, in a couple of these instances students were explaining how they didn’t need to worry about using these structures for this test, because that wasn’t what graders would be looking for. This interpretation contrasts with the description of the composition as an academic essay made by all three graders:

   **Administrator 6 (grader):** I mean, I think it was five paragraphs, just a standard academic essay, that most kids in high school English should know how to write, should know how to do an introduction with a thesis statement, three paragraphs of development and a conclusion.

   This mismatch is demonstrated in the following exchange, where a student recounts when she realized that her own perception of what was required of her diverged from a grader, and how this ostensibly led to her failure:

   **Student 8:** I don’t know, and instead of writing an essay, I just thought okay, they’re not going to look at if I do an intro, like the paragraphs and a conclusion, so I just wrote and wrote and wrote…I never did an intro, three paragraphs and a conclusion. (The correctors of my test) started and they saw that there was no intro, so it’s a fail.

   **Interviewer:** The instructor of the (remedial) course told you that?

   **Student 8:** Yup. She’s the one who corrects it, I guess.

   In response to questions about what was expected and measured in this assessment, students and administrators did not only discuss specific elements of performance but discussed wider goals of the university system, of the teaching profession, and of society. One theme that was common across students and administrators was the depiction of future educators as models of English to their students, and the responsibility for English competence that this implies:

   **Student 3:** I mean, we’re future teachers, so I mean we have to master the language in some way.
Administrator 6 (grader): Imagine that person up in front of a crowd of young people, writing on the board, filling that board with all kinds of grammatical errors and spelling mistakes?

Administrator 2: They’re going to be our future teachers. Would you want your child to be taught by somebody who can’t spell correctly? Or, makes the same grammatical errors over and over again? This is what they’re teaching your child.

Another prevalent theme, but one not shared by students and administrators, is the idea that there is a system of testing in place because students are required to constantly prove to the university establishment that they belong there and they deserve to remain. The inherently confrontational nature of these comments is evidenced by the following statements from the student focus groups only:

Student 5: So, okay, so what do I do? Can I stay? …Okay, so then I’ll just prove you wrong that I belong here.

Student 5: I just went out and decided to work hard to prove that I belonged here.

Student 2: I failed the first exam, the English Requirement Test, I felt that I had to prove myself to, that I belonged, that I was in the right program and I had to prove not only to myself, but to…all the faculty.

At no point do the graders or other administrators refer to this assessment as a means by which students must prove that they deserve to remain. In fact, all comments about the goals of the test by administrators have to do with objective measurements of quality, with the composition being a neutral sample of performance.

Discussion

Research Question 1: Is the construct of competent writing the same as suggested by:

- the test prompts
- the passing compositions?
- the student comments?
- the rater comments?

In response to this question, there is evidence that these four sources of information do not reveal a common conception of the construct of quality writing. Firstly, the prompt language suggests genre requirements that are not necessarily taken up in passing papers, which suggests that rhetorical structure is not necessarily as valued a component of these compositions as rater comments imply. One might very well ask if the prompt is playing the role we typically ascribe to it, or if the true prompt consists of the preconceived notions of what is typically expected in this testing situation on the part of both the writers and the readers.
Secondly, what the students believed was most valued in the compositions—accurate control of language structures—seems to be confirmed by the textual analysis of the passing and failing compositions, which only showed significant differences in terms of language error. The decisions made during the composing process by the students were directly affected by the perceptions these students had of how the compositions would be read, and it would seem that their intuitions were correct. However, the administrators of the test did not characterize their rating behavior, to the students or in interviews, as consisting of looking for and counting language errors. There is evidence that the raters believed that they were looking at appropriate rhetorical elements associated with an academic essay. Textual analysis did not indicate that passing papers needed to be an essay or to have any particular rhetorical profile.

How graders characterize their grading practices is perhaps not as illuminating as how they discuss the social role of teachers in society as models of good language. At this point, the fundamental importance of error-free composition comes through, and more clearly indicates what affected graders when they turned on their “red light” during grading.

**Research Question 2:** If there is not a common understanding of this construct, what are the implications for the parties involved?

It would be expected that students’ papers would fail if they held fundamentally different perceptions than the graders of the genre required to demonstrate quality writing. This did not seem to be the case. This may be because the conception of quality writing as described by graders is not put into action in the pass/fail decisions they make for these compositions. Of course, while language errors appear to be playing a major role, there could be other criteria playing a factor in pass/fail decisions. However, there is no evidence here that rhetorical structure (an element graders mention) plays a role, and whatever any additional criteria may be, these criteria are not made explicit and may vary from one grader to the other.

There is another important implication arising from these conflicting perspectives: the surprising genre profile that emerges from compositions that represent competent (i.e., passing) writing. Student goals, perceived audience, and the roles played by writers and readers do not operate as in a traditional view of the academic essay. Evidence suggests that test takers approach the task of writing not with rhetorical goals in mind but with goals related to their social role at the university. They have stated that they are writing to prove their worthiness. The audience for these compositions is not perceived as interested or even collegial. They are people looking for errors, perceived by students as skeptical of their writing competence. Any characterization of the genres produced here would be incomplete if it did not include this inherently confrontational relationship between writer and reader.

Limitations of this current work involve the potential bias introduced through the data collection process. Additional textual characteristics may be differentiating between the passing and failing compositions but were not evident in this sample. There could have been other elements of student writing that graders use to make pass/fail decisions, but that they are unaware of or neglected to mention in the interviews. As the focus groups were self-selected, students may have volunteered because they saw this as a chance to speak about their experiences in the university in general, not only their specific experiences with this test. These findings indicate that more work must be undertaken to further explore the relationship between genre expectations and the construct of writing competence in high-stakes university writing tests.
Conclusions and Future Implications

The next phase of this project involves making use of these insights in the development of a common proficiency test across all English Quebec universities. Every effort will be made to ensure that the new common assessment has the following elements:

**The Input of Multiple Stakeholders, Including Test Takers, in Development and Validation**

The consideration of multiple sources of evidence from multiple stakeholders has enhanced the validation of this current test (Bachman, 1990) and will inform the development of the new common test. In particular, it was especially valuable to include the writers’ perceptions of the writing required of them in assessment situations (Leki & Carson, 1997). This valuing of the role of the test takers has also been encouraged both in traditional education measurement (Ryan, 2002) and as part of Critical Language Testing (Shohamy, 2001). The importance of test-taker input goes beyond the ethical imperative to consult test takers during the test review and development process. As seen in this study, this input, in addition to input from administrators and graders, contributes to an expanded view of genre which enhances our understanding of writing for high-stakes university gatekeeping tests.

**A Complex Generic Representation which is Common among all Stakeholders**

This expanded view of genre complicates our conception of the construct of proficient English writing, but this conception should be understood by all stakeholders. The first step in eliminating genre misunderstandings is to ensure that any writing task in the test be given an instrumental, not just an evaluative, goal, to “go beyond a ritual display of knowledge for assessment” (Shaw & Weir, 2007, p. 71).

A complex construct will go hand in hand with grading procedures which are capable of capturing this complexity. This means using rating scales rather than impressionistic grading. Certainly, the simple incorporation of a rating scale does not necessarily ensure that all graders suddenly have a common understanding of the construct or will grade in the exact same way (see Lumley, 2005). However, there is much to be gained from providing more guidance during grading as well as deflecting from knee-jerk accuracy judgements. One should hesitate in making generalisations about writing success in either academic or professional contexts if it is suspected that too much of the construct of quality writing is based solely on the absence of error. This may be especially true in settings such as these, with mixed first and second language writers. The types of errors made by second language writers may be more salient or disturbing for first language graders.

There is an additional benefit to establishing an acceptable common description of the genre profile of successful performance on these tests. As Miller (1994) reminds us, “genres serve as keys to understanding how to participate in the actions of a community” (p. 39). Professional communication is part of identity-building within a community of educators, so we are doing a disservice to future teachers if our assessment of language doesn’t help them to greater understand and access membership in that community.
Notes

1 Quebec has separate English and French school boards. The English proficiency test is intended for preservice teachers in English medium universities, which includes those who will be teaching in the English school system as well as those who will be teaching English as a Second Language in the French school system.

2 See Artemeva, 2004, for a detailed discussion of the trend from primarily text-based analyses of genre to the New Rhetorical approach, reconceptualising genre as a form of social action.

3 See Luce-Kapler and Klinger, 2005, for another example of the use of focus groups with students who have passed and failed a high-stakes high-school exit test in Ontario, Canada.

Acknowledgements

I wish to thank the English Language Institute at the University of Michigan for granting me the funds to enable me to present this work. I also wish to thank Dr. Carolyn Turner, Heike Neumann, Alison Crump, Candace Farris, and Myra Lepp for their valuable assistance on this project.

References


Appendix A: Interview Guiding Questions

1. How did you come to be involved with the English Proficiency Test (EPT)? (What is your connection to the EPT?) Who are the other people involved with this test?

2. (For those that know the history of the test or helped to develop it) How did this test come about? (Who created it?) Did anything exist before this test?

3. Why do we have this test? If you had to explain this test to an undergraduate student, how would you describe it?

4. If language ability is mentioned) You mentioned the language ability/competence of the students here. How would you describe the language abilities of preservice teachers here? How would you describe the language quality of teachers in Quebec schools?

5. Look 10 years into the future. Will we still have this test? What will the quality of language of our preservice teachers be? What will the language quality of our teachers be in Quebec schools?

6. Can you give me the details of a time when you had to discuss this test with someone else (such as a student, a professor, a member of CAPFE, etc.)? What did they want to know about the test? What did you tell them?

Appendix B: Focus Group Guiding Questions

1. What did you do on this test? What did they expect of you?

2. Who gave you this test? What is the test for? Why do you have to write it? What are the results used for?

3. What did you know about this test before you took it? How did you find out? (What did you read, what did people tell you?)

4. What feelings did you have about the test before you wrote it? (Were you concerned about it?)

5. What did the test consist of? Why do you think the test is structured the way it is?

6. Can you remember what you were thinking as you were writing the test? When you got the results?

7. If someone asked you about your language abilities, what would you say?
Collaborating with ESP Stakeholders in Rating Scale Validation: The Case of the ICAO Rating Scale

Ute Knoch
University of Melbourne

ABSTRACT In response to a number of high-profile aviation accidents in which language proficiency was cited as a major contributing factor and after consultation with the PRICE Study Group, ICAO stipulated that, as of March 5, 2008, pilots, air traffic controllers, and flight crew involved in flight operations in international airspace need to have their English language proficiency endorsed on their license. ICAO published a detailed document entitled Manual on the Implementation of ICAO Language Proficiency Requirements (International Civil Aviation Organisation, 2004) that includes an analytic rating scale designed to assess speaking ability. Each criterion (fluency, structure, pronunciation, vocabulary, interaction, and comprehension) is assessed on a six-point scale ranging from Level 1 (pre-elementary) to Level 6 (expert). Candidates are certified at the level that corresponds to their lowest performance on a criterion. It was stipulated that all pilots and air traffic controllers engaged in international flights should have a certified English ability of at least Level 4. However, little information is available about the development and validation of this rating scale. In fact, it is understood that the PRICE Study Group commissioned to develop the scale had to work with few resources. Validation work is clearly needed.

This study set out to explore the utility of using different groups of stakeholders in such validation work. Two stakeholder groups were targeted: (a) users of the ICAO rating scale and (b) pilots. In the first part of the study, an online survey was sent out to elicit the views of test developers, administrators, and raters of the criteria and descriptors on the rating scale. Fifty-five responses were received from professionals around the world. The second part of the study involved focus-group interviews with pilots. Twelve operational pilots as well as pilot trainers listened in groups of two or three to speaking performances of test takers taking a variety of aviation English tests. The performances were chosen to represent a range of ICAO levels, but were mainly selected to be around the Level 3 / Level 4 cutoff. The focus-group participants were asked to rate the acceptability of the pilot’s language ability for (a) making announcements to passengers, (b) communicating with other pilots, and (c) radiotelephony communications with air traffic control. The aim of these focus groups was twofold: (1) to establish the “indigenous”
assessment criteria (Jacoby, 1998) pilots use when assessing the language ability of peers (and therefore validate the criteria in the scale), and (2) to establish whether Level 4 is, in the pilots’ opinion, sufficient as operational level.

The results of the survey show that there is criticism of some of the ICAO scale descriptors among those stakeholders who responded. The paper reports on the range of problems mentioned by the survey participants. In the focus groups, pilots focused on some of the criteria that are included on the rating scale (e.g., pronunciation and fluency), but they paid less attention to others (e.g., structure). Whilst listening to the performances, they also paid a lot of attention to the pilots’ technical knowledge, which for them could not be disconnected from the test takers’ language performance. In fact, in certain borderline performances, pilots used their judgment of the candidates’ technical knowledge to influence their decision of the candidates’ language ability.

The paper concludes by discussing the usefulness of involving stakeholders in ESP rating scale validation work and shows that a wide range of stakeholders, even if not language testing experts, should be included in post hoc validation work.

**Background**

Rating scales are used in objectively scored assessment contexts because they are a representation of the test construct (McNamara, 2002; Turner, 2000). To best represent the test construct at hand, rating scales should be grounded in some sort of theory that describes the type of language used in the target language use domain. However, rating scales are often not developed in a way that accounts for such a theory. In fact, several authors have pointed out that often very little information is available on how rating scales were developed (e.g., Brindley, 1998; McNamara, 1996; Turner, 2000; Upshur & Turner, 1995).

Several methods of rating scale development have been described in the language testing literature. Fulcher (2003) points out that most rating scales are based on the intuitions of the test developers or are based on preexisting rating scales (which may or may not have been taken from a different context). A number of studies have described empirically based methods of scale development (e.g., Fulcher, 1987; Knoch, 2007). Here, descriptors are created through an empirically verifiable procedure and are based on observable learner behavior. Knoch (2007) and others have argued that the categories included in a rating scale should be based on a theory of language, language development, or language use, and that the rating scale descriptors should be empirically based.

In the context of LSP testing, Douglas (2001) writes that the content of the target language use (TLU) domain that serves as the basis for the content of the test tasks is usually fairly well understood; however, the way assessment criteria should be developed is not through an analysis of the TLU situation, but rather through an understanding of what it means to know and use a language in the specific context (Jacoby & McNamara, 1999). He further argues that in the development of LSP assessment criteria, theoretically based
approaches should be supplemented by taking into account the criteria that experienced professionals in the relevant field use when evaluating communicative language use. A number of studies have been done in this area of indigenous assessment criteria (Douglas & Myer, 2000; Jacoby, 1998).

Using subject specialists’ judgments of language performance adds to the validity of the resulting assessment criteria. Usually, however, these indigenous assessment criteria are used during the rating scale development process. The aim of this study is, therefore, to explore the utility of using stakeholder input in the context of post hoc validation of an LSP rating scale. More specifically, two aspects of post hoc stakeholder input are explored: (a) using scale users’ opinions of the ICAO rating scale in the validation process and (b) establishing whether canvassing pilots’ indigenous assessment criteria can be used to validated and potentially improve the ICAO scale.

**Research Questions**

The study attempted to answer the following three research questions:

1. What are stakeholders’ opinions of the ICAO rating scale?
2. Do raters and industry professionals agree on the appropriate proficiency level for operational flying?
3. What criteria do pilots use when evaluating the effectiveness of speech samples of other pilots?

**Methodology**

The study was undertaken in two phases:

1. A questionnaire completed by a range of aviation and aviation English experts using the ICAO rating scale (to answer Research Question 1)
2. Focus group interviews with pilots (to answer Research Question 2)

As both phases are distinct, the methods used during each will be reported separately.

**Phase 1: ICAO Rating Scale Questionnaire**

*Instruments*

A questionnaire eliciting stakeholders’ perceptions of the ICAO rating scale was designed. The questions focused on the categories in the scale (i.e., fluency, structure, vocabulary, pronunciation, comprehension, and interaction) as well as the descriptors in each scale. Some more general questions were also included and respondents were also asked to complete some basic questions about their background. Wherever possible, respondents were provided with the opportunity to provide open-ended answers to questions.

There were two versions of the questionnaire. In the first version, the questionnaire formed part of a larger questionnaire designed by Alderson (2008). In the second version, the same questions were sent out by themselves. The questionnaire was loaded onto
Surveymonkey so that it could be completed in electronic form, but a version in Word format was also available.

Participants
Fifty-five people responded to the questionnaire. They were from at least 23 countries (not everyone listed their country). To illustrate the wide range of origins of the respondents, the list of 23 identified countries is shown below according to broader geographical locations:

- Europe: Italy, Hungary, France, Switzerland, Finland, Macedonia, Belgium, UK, Sweden, Germany, Estonia, Ireland
- North and South America: USA, Canada, Argentina, Brazil, Mexico, Peru
- Australasia: Australia, New Zealand, China, Indonesia
- Africa: South Africa

The respondents were from a wide range of professional backgrounds. Some worked for their countries’ civil aviation authorities (e.g., in charge of flight crew licensing), others worked for airlines or the air force (e.g., as simulator supervisors or in charge of training quality), others were pilots or air traffic controllers. Another group of respondents were either language testers or language instructors. People in this group were either academics working at universities, language testers in charge of specific English for Aviation tests, raters of Aviation English tests, graduate students interested in Aviation English (testing), or they were involved in Aviation English instruction at training companies, civil aviation authorities, or airlines. Two of the respondents had been members of the original ICAO PRICE study group. It is fair to say that the respondents were from a wide range of backgrounds including language experts and aviation experts and neither group was overrepresented.

Procedures – Data Collection
The questionnaire was sent out in two phases. During the first phase, the questionnaire was emailed to aviation test providers as part of a larger study (Alderson, 2008). Twenty-two responses were received to this larger questionnaire. Because responses from stakeholders other than those targeted during the first phase were also of interest, the questionnaire was sent out again (without the questions of the larger study). Thirty-three responses were received in this second phase. The link to the online questionnaire was emailed to contacts (who were also asked to pass it on to anyone working in the area of aviation English) and it was posted on two Aviation English discussion boards: Flight English and ICAEA World. Participants were also provided with the opportunity of completing the questionnaire in Word format and emailing it back to the researcher. Four participants selected this option. The responses to both questionnaires were received between April and October 2008.

Procedures – Data Analysis
The questionnaire responses were analyzed using two methods: (a) wherever possible responses were quantified (e.g., how many people selected “yes” or “no” to a question), (b) for open-ended responses, common themes emerging from the data were recorded.
Phase 2: Focus-Group Interviews

During the focus-group interviews, pilots were played extracts of recordings from three different types of Aviation English tests. Pilots were asked to decide whether they felt the English proficiency of each candidate was sufficient to work as a pilot. The aim of this phase was to investigate two points: (1) whether pilots agree with the Level 3/4 cutoff as being the appropriate level for operational flying and (2) which aspects of the test takers speech the pilots were most likely to focus on.

Instruments

Speech Samples.

Eight speech samples were selected for this phase of the study. They were selected from three different types of aviation tests, thought to be representative of the types of tests used in different countries. The three tests were:
- a semidirect speaking test
- a structured interview
- an Oral Proficiency Interview (OPI)

The reason why more than one test type was selected was to avoid the criticism that the test is not eliciting the type of speech representative of the target language use domain. This criticism could of course be leveled at all three test types, however this could not be avoided. The speech samples were selected to represent a range of proficiency levels and L1 backgrounds. A summary of the eight speech samples can be found in Table 1, below:

<table>
<thead>
<tr>
<th>Speech sample</th>
<th>ICAO Level</th>
<th>L1 background</th>
<th>Type of test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>Indian</td>
<td>Semidirect</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>Japanese</td>
<td>Direct - scripted</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Korean</td>
<td>Semidirect</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Arabic</td>
<td>Semidirect</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>Chinese</td>
<td>Direct - scripted</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>Japanese</td>
<td>Direct - scripted</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>Chinese</td>
<td>Direct - OPI</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>Indian</td>
<td>Direct - scripted</td>
</tr>
</tbody>
</table>

As can be seen from Table 1 above, most speech samples were around the Level 3/4 cutoff, as these levels were deemed to be the most interesting for the purpose of the focus-group interviews. Two performances at Level 5 were included as some questionnaire respondents had argued Level 5 to be a more appropriate level for operational flying. The speech samples had previously been rated by two trained Aviation English raters. Only speech samples on which the raters had agreed were selected for the purpose of this study. Two speech samples from the “ICAO rated samples” were included. The length of the speech samples ranged from 3 minutes to 4.5 minutes.
Speech Sample Questionnaire.

While listening to each speech sample, pilots were asked to complete a short questionnaire about each speaker. The questionnaire questions can be found in Figure 1, below:

<table>
<thead>
<tr>
<th>Do you think this pilot would be able to communicate effectively with:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) other pilots ☐ Yes ☐ No</td>
</tr>
<tr>
<td>b) ATC  ☐ Yes ☐ No</td>
</tr>
<tr>
<td>c) passengers ☐ Yes ☐ No</td>
</tr>
</tbody>
</table>

Give reasons:
___________________________________________________________________________
___________________________________________________________________________
___________________________________________________________________________

Would you be comfortable working with this pilot?
☐ Yes
☐ No

Figure 1. Questionnaire

As can be seen, the participants were asked to judge whether the speaker in each speech sample would be able to communicate effectively with (a) other pilots, (b) ATC (air traffic control), and (c) passengers. These three groups were selected because they would be the main groups pilots would communicate with when in the air; ineffective communication with passengers is of course less likely to pose a risk to flight safety.

Participants

Ten pilots participated in the interviews. They were all very experienced pilots, which the summary of the background information in Table 2 below shows. The majority of pilots were from a major airline in Australia, while others were recruited through personal networks of the researcher. About half the pilots taking part in the study were at the time of the research working in the capacity of pilot trainers.

Table 2. Background Information – Pilot Participants

<table>
<thead>
<tr>
<th>Pilot</th>
<th>Years of flying experience</th>
<th>Domestic/international experience</th>
<th>Experience working with colleagues from non–English speaking backgrounds?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>Both</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>Both</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>Both</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>Both</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>Domestic</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>Domestic</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>Both</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>Both</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
<td>Both</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>Both</td>
<td>Yes</td>
</tr>
</tbody>
</table>
To take part in the study, pilots had to have sufficient experience working with air traffic control, as it is conceivable that less experienced pilots mainly flying in visual flight rule conditions would not have much opportunity to talk to ATC. Because pilots in Australia have very little opportunity to meet speakers from backgrounds other than English in their professional life, great care was taken to recruit pilots with some overseas flying experience. Pilots 1 and 2, who indicated not having any colleagues from non–English speaking backgrounds, did have extensive experience flying into Asia and therefore had encountered ATC whose first language was not English. Of the ten pilots taking part in the study, six were from the same major airline, three were working for a flight training school training overseas student pilots, and one was a retired pilot working in pilot training in Asia. While it would have been preferable to also include pilots from a wider range of backgrounds (including pilots from nonnative speaking backgrounds), this was not possible at the time of the study.

 Procedures

 Data Collection.
 Pilots were recruited through personal networks of the researcher and with the help of two senior staff members at an Australian airline who helped with sending out emails asking for research volunteers. Although the intention was to perform the focus group interviews in small groups of three to four pilots, only seven pilots were interviewed in this manner. The others volunteered to listen to and comment on the speech samples in a one-on-one manner. All interviews were undertaken in a quiet room. All interviews were recorded using a digital recorder.

 Data Analysis.
 The interviews were transcribed by a research assistant and then checked by the researcher. The next step was to identify coding categories. It was assumed that some of the categories in the ICAO rating scale would be mentioned during the interviews, therefore these categories formed the starting point for the analysis. Careful reading and rereading of the interview data also resulted in a number of other categories, which will be described in the results section of Phase 2. This analysis is in line with Douglas’ (2001, p. 182) and Douglas and Selinker’s (1994) recommendation on how such commentary about primary data (in this case the speech samples) should be analyzed.

 Results – Phase 1: Questionnaire Data

 The responses to the questionnaire data presented below aim to answer the first research question: What are the opinions of stakeholders of the ICAO rating scale?

 The first questionnaire question asked respondents whether they thought that the scale was designed to be used by both language (testing) experts and aviation experts. This question was included because it could be argued that industry professionals are in the best position to judge the communicative effectiveness of performances of pilots and air traffic controllers and that engaging such professionals in the rating process should therefore be encouraged.

 The responses were divided on this question. Table 3 below presents the summary of responses, and Table 4 presents some of the open-ended responses.
Table 3. Quantitative Results Questionnaire Question 1

<table>
<thead>
<tr>
<th>Is the ICAO rating scale equally useful for language experts and aviation experts?</th>
<th>Yes</th>
<th>No</th>
<th>Undecided/depends</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 22 (42%)</td>
<td>N = 22 (42%)</td>
<td>N = 8 (15%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Open-Ended Responses to Questionnaire Question 1

I think the ICAO rating scale is more useful for language experts than for aviation experts.

Not useful for aviation experts. I have to explain some of the terminology in the rating scale to aviation experts so that they know how to rate or how they are being rated.

No. I think it would be difficult for aviation experts who lack a language training background to use the scale effectively. At the same time, although the interview should not contain technical material, language experts should have sufficient knowledge of the controller-pilot communicative environment to ensure the interview context is relevant to the context in which controllers and pilots will be communicating.

No. Most aviation experts have a limited knowledge of the rating scale, and possibly an unrealistic view of it.

Yes, but only if the aviation expert is familiar with the linguistic terms (e.g., discourse markers etc).

No. We developed additional explanatory materials for aviation experts. This “translation” proved much more accessible to those aviation experts who will be administering and scoring.

Not without training and clarification of some of the terminology. In the training sessions we had, there were questions about the meaning of some more linguistic terms in the scale, such as “idiomatic,” which were not clear to many aviators.

It was further of interest whether scale users thought that Level 4 is the adequate level for operational flying (Table 5). Again, the respondents did not agree on this. Half of the respondents thought it was adequate, while half thought it was not. A number of representative qualitative responses are listed below (Table 6).

Table 5. Quantitative Results Questionnaire Question 2

<table>
<thead>
<tr>
<th>Do you think ICAO Level 4 is the appropriate operational level for pilots and air traffic controllers?</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 26 (48%)</td>
<td>N = 28 (51%)</td>
<td></td>
</tr>
</tbody>
</table>
The remainder of the questionnaire was devoted to more specific questions about the categories and the descriptors on the ICAO rating scale. First, participants were asked whether they felt satisfied with the categories and the descriptors at different band levels on the ICAO scale.

As was noted in the other questions, the respondents were divided in their answers. The exact breakdown can be seen in Table 7 below. The summary shows that the respondents were generally satisfied with the categories.

Table 7. Quantitative Results Questionnaire Question 3

<table>
<thead>
<tr>
<th>Category</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>33 (75%)</td>
<td>11 (25%)</td>
</tr>
<tr>
<td>Structure</td>
<td>31 (70%)</td>
<td>13 (30%)</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>30 (68%)</td>
<td>14 (32%)</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>31 (70%)</td>
<td>13 (30%)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>33 (75%)</td>
<td>11 (25%)</td>
</tr>
<tr>
<td>Interaction</td>
<td>30 (68%)</td>
<td>14 (32%)</td>
</tr>
</tbody>
</table>

However, even though the summary of the quantitative responses points to a general satisfaction of stakeholders with the rating scale, research participants added a wealth of open-ended responses to this question that point to the fact that revisions might be necessary. The responses are summarized below, focusing on each scale category in turn.

Fluency

Although most respondents were satisfied with the category of fluency, there were a number of suggestions for improvements. One participant noted that discourse markers are less relevant in the TLU domain and therefore should not feature in the rating scale. For example, one respondent wrote the following:

*We are asked to do things which are never going to be used at work, for example [we are not going to say]: As a matter of fact, it’s not possible for me to clear you for decent, even though you are in an emergency, since you have traffic below you. That is not the way we are going to talk!!*
Similarly, another participant noted the following:

*I have encountered speakers with high fluency who do not use discourse markers/connectors. Many speakers from Eastern European countries do not vary their speech flow at all, but they are fluent speakers. Also, I have experienced fillers as being distracting, when my rating partner has not. There seems to be a strong subjective element involved...*

One respondent noted that there was no mention of the transition from phraseology to plain English. Another common response (in fact one noted across all categories), was a criticism of the vague language used. One respondent, for example, criticized the use of words that could lead to ambiguity, such as “natural language flow,” “distracting,” and “appropriate.” Another participant commented that formulaic speech was left undefined in the scale. He was wondering if this refers to the use of phraseology. Finally, one participant was wondering whether “variation of speech flow for stylistic effect, e.g., to emphasise a point” should be mentioned in the pronunciation scale rather the fluency scale.

**Structure**

Participants generally commented positively about the descriptors in the trait scale “structure”; however, there was some confusion about the descriptor for Level 3, which reads: *Basic grammatical structures and sentence patterns associated with predictable situations are not always well controlled. Errors frequently interfere with meaning.* One respondent, for example, commented that:

*There seems to be the assumption that unusual or unexpected situations elicit complex structures and routine situations elicit basic grammatical structures, however this is not always the case.*

Just as with the fluency scale, there was some confusion about the use of adverbials such as “usually” and “consistently.” For example, one participant commented that “the degree of manipulation of basic and complex structures and the degree and density of grammatical error at Level 4 to Level 6 is unclear.” Overall, however, the comments relating to this trait scale were positive. The scale was described as “well-designed and reasonable to distinguish test takers” and that “it covers the differences quite satisfactorily.”

**Vocabulary**

The main criticism leveled at the vocabulary trait scale had to do with the use of idiomatic language in the descriptors. For example, one respondent wrote:

*The [ICAO] manual recommends that idioms are not used (Chapter 3: 3.2.1), but descriptors state the use of idioms are basically an element of being highly proficient speakers, which is contradictory in this particular context.*

On a similar line, another participant wrote: “*Idiomatic language shouldn’t appear in safety critical communication.*” This was also mentioned in the following quote: “*I don’t*
agree that the use of idiomatic expressions should be rewarded, as they are strictly forbidden in the Aviation environment.”

Another contentious issue in the trait scale for “vocabulary” is the use of the word “register.” For example, one respondent wrote:

Sensitivity to register is difficult to discern as it is culturally influenced. Some cultures are extremely formal, as opposed to other very relaxed cultures where speakers use a lot of colloquial language.

Finally, the reference in the descriptors to “common, concrete, and work-related topics” was unclear to three respondents. One wrote: “The meaning of ‘common, concrete, and work-related topics’ is unclear—is this one set of topics or three?”

**Pronunciation**

A number of common themes emerged from the comments about the pronunciation scale. One common thread related to the feeling that the ratings very much depended on the background of the raters and that the adverbs of frequency were the only distinguishing factor between the different band levels. Three such comments are reproduced below:

This is my specialization and perhaps for this reason I find this skill the most ludicrously vague. Who can know what the difference between almost never and sometimes is in real terms? “Understanding” for whom, the examiner who is familiar with the accent of his students or a NS or a NNS.

The problem with this descriptor is that raters who work with multi-cultural candidates, understand more than those raters who only work with native speakers. Also, if the rater has an aviation background, some of the aviation terms are easier to determine if spoken incorrectly because of context-based information. The term “understanding” needs to be redefined.

I honestly feel that this is the most difficult area to rate accurately, as it is so subjective and determined by the rater’s own language and prior experience with the language that is being tested.

Finally, it was also criticized that there is a reference to first-language influence in the descriptors.

**Comprehension**

Several topics were mentioned for the rating scale of comprehension. The most commonly recurring theme was that comprehension could not be accurately measured in a scale designed to assess speaking performance. Below are some selected comments relating to this theme:

- It’s not clear how the comprehension scale would be operationalized—how would one determine that “comprehension maybe be slower or require clarification strategies”?
- Comprehension should not appear in a scale for speaking
• Cannot be assessed effectively in speaking performance
• I have also noticed that some speakers just naturally ask a lot of questions, to be absolutely clear about instructions. It is a personality rather than a speech characteristic. In terms of the rating scale, this could then be construed as uncertainty and slow comprehension
• Some candidates comprehend the questions asked of them but are unable to respond due to poor vocabulary; therefore, the comprehension score is impacted
• It is problematic that comprehension actually occurs in a scale for assessing spoken language. This means that the presence of clarification strategies means that the measurement of comprehension must be undertaken in a “live” test event.

A number of people involved in test development mentioned that it was difficult to select speakers who have an “accent” that is sufficiently intelligible for international users. Some mentioned that this was nearly impossible to achieve and caused a threat to the validity of the scale.

Another respondent criticized the lack of congruence across scale levels. For example, “a range of accents” is mentioned at Level 5, but not at other levels.

Interaction
There were three major concerns about the interaction rating scale. Each of these is summarized below.

The first concern was the inclusion of descriptors relating to nonverbal clues. A number of respondents mentioned that nonverbal clues are not a feature of radiotelephony speech and should therefore not be assessed.

A second problem raised by the participants was that the descriptors for Level 5 and Level 6 seemed to have been switched with the descriptor for Level 5 being higher than the one for Level 6.

Finally, as was the case with the rating scale for comprehension, some features of speech are mentioned in only a small number of band levels and not across all levels in the scale. For example, one participant wrote: “Even a Level 2 can deal with misunderstanding by checking, confirming and clarifying by saying ‘What do you mean by ...?’ I think that’s fairly basic interaction especially in the context of air traffic communication. This should be mentioned across levels.”

Rating Scale as a Whole
The last section of the questionnaire provided opportunity for responses not covered in the other questionnaire items. A large number of issues are raised, some going beyond the actual scale. Nevertheless, this section attempts to summarize these issues because all raise interesting ideas. An attempt is made to first cover issues raised about the scale itself and then move to more general concerns.

A number of participants mentioned that Level 6 across all scale categories seems irrelevant and not much different to Level 5. Similarly, there was a criticism of Level 6 in that it requires test takers to demonstrate abilities not necessarily expected in the target language use domain. One respondent, for example, wrote:
Level 6 is problematic in that it requires people to demonstrate a level of language proficiency that should not be happening on the frequency (using idioms, comprehension of cultural and linguistic subtleties, sensitivity to nonverbal clues, etc). I believe Level 6 should not be included as a test for operational pilots/controllers—it describes language we do not want to hear on the frequency.

Similarly, there was concern that the scale is not always consistent with what is advocated in the ICAO manual. That criticism was also related to those aspects mentioned by the previous participant (i.e., idiomatic language, nonverbal clues, etc):

Related to the previous point was a concern that the scale is not based on a verified theoretical framework. In fact, rating scales are designed to represent the test construct, there was concern that this scale currently does not represent an accurate portrayal of what happens in the target language use domain.

Other respondents lamented the fact that the scale was an inadequate basis for test development. For example, one respondent noted that operationalizing concepts such as “unusual or unexpected situations” in a listening or speaking test was difficult and that the scale and the manual together did not provide sufficient examples of what is meant by this.

Finally, a number of participants were concerned that different test providers were interpreting the scale levels differently and that therefore a Level 4 on one test could not be equated to a Level 4 on another test. Several solutions were offered for this; including an international rater network as well as the suggestion that ICAO should audit the tests. Some of these issues and others surrounding aviation language testing are also discussed in Alderson (forthcoming).

Results – Phase 2: Focus-Group Interviews

RQ 2: Do raters and industry professionals agree on the appropriate proficiency level for operational flying?

Another contentious issue surrounding the new ICAO policy is what the appropriate operational level for pilots and air traffic control should be. ICAO stipulated that to be allowed to work, pilots and ATCs should at least reach Level 4 in all criteria. The purpose of the next section of this study is to establish whether pilots agreed with this level as being sufficient for operational flying.

A group of ten pilots were played test performances of eight candidates of different proficiency levels. The pilots were asked whether they thought the language proficiency of each candidate was sufficient to talk to (a) other pilots, (b) ATC, and (c) passengers.

Table 8 presents a summary of the results. The first column shows the number of the speech sample. The rating provided by two language-trained raters can be found in column 2. Columns 3–5 show how many of the ten pilots thought that the language proficiency of the speakers was high enough to talk adequately to other pilots, ATC, and passengers.
Table 8. Pilots’ Speech Sample Questionnaire Responses (collated)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Language raters</th>
<th>Work with other pilots</th>
<th>Work with ATC</th>
<th>Talk to passengers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High 4</td>
<td>90%</td>
<td>80%</td>
<td>60%</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>30%</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>20%</td>
<td>30%</td>
<td>10%</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>80%</td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>100%</td>
<td>90%</td>
<td>80%</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>60%</td>
<td>70%</td>
<td>30%</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

In terms of general trends, the pilots seemed to agree with the ICAO ratings. However, there was much disagreement around candidates who were borderline around the Level 3 and 4 cutoff. All pilots thought that the candidate with the highest language proficiency, Speaker 9, is proficient enough to work as a pilot. They also generally agreed that Speaker 1, the candidate rated the second highest, should be allowed to work. However, they were less in agreement when it came to the appropriacy of this candidate talking to passengers, something that is, of course, less of a risk to flight safety. The pilots also agreed with the ICAO rating of Speaker 7. Almost everyone deemed him as proficient enough to fly. However, the other speaker rated as Level 4—Speaker 4—was generally rated very low. Hardly any of the pilots thought that this speaker was proficient enough to work as a pilot. This speaker will be examined in more detail later. Four speakers were rated as being at Level 3 by the language-trained raters. However, the pilots were divided when making their ratings. They generally agreed that Speakers 2 and 3 were not proficient enough to work as pilots where English is required; however, nearly all thought that Speaker 6 was proficient enough to work in such an environment, and more than half considered Speaker 8 proficient enough. Speaker 8 was a sample selected from the ICAO website, and a group of Aviation English test raters that rerated this sample thought that he should actually have been rated at Level 4. Therefore this discrepancy can be explained. This leaves two speakers where the language-trained raters and the pilots participating in this study did not agree: Speaker 4 and Speaker 6.

Before turning to a more detailed analysis of these speakers, it is also interesting to note that in the majority of cases the pilots were harsher in their judgments when “talking to passengers” was concerned. This was an unexpected finding, as inappropriate speech or low proficiency when addressing passengers does not directly lead to a threat to flight safety.

The following section presents a closer analysis of the two speech samples on which the language raters and the pilots could not agree.

Speaker 4

Speaker 4 was rated at Level 4 by the language-trained raters, but only a minority of pilot participants rated this speaker as proficient enough to work as a pilot. Because of this discrepancy between the ratings and the pilots’ judgement, a further scrutiny of the pilots’ remarks about this speaker was undertaken.

Pilots’ remarks focused on a number of linguistic features in the test taker’s speech. Four pilot participants commented on the lack of vocabulary in Speaker 4’s speech. For example, Participant 7 said: “It seems to me his vocabulary is not great, if you just gotta explain going around ... aircraft in front has a technical problem, so you don’t panic
Collaborating with ESP Stakeholders in Rating Scale Validation: The Case of the ICAO Rating Scale

everybody by telling them it’s on fire ....” Similarly Participant 5 said: “He just didn’t have the... he didn’t have the words to explain it....”

Four pilots mentioned in the interviews that they thought Speaker 4 also lacked comprehension. Participant 1, for example, said: “Again, I thought he had trouble actually understanding the question.”

Pilot 9 was concerned with Speaker 4’s lack of English language ability. He particularly thought that he would have trouble communicating with other L2 speakers or with certain L1 English speakers less familiar with foreign accents. Below are some excerpts from the interview:

[Working with] other pilots, it would depend on who the other pilot is, but if it was someone who was marginal in English again, communicating in English would be a problem. He couldn’t express... couldn’t get the words to express the situation. I think that’s gonna depend a little bit on the country people come from. Australians now are getting quite adept at hearing a lot of English variation. Someone from America... well, it’s a while ago now and probably not as bad [when I flew there], but an American still has a little bit of difficulty even with me. ... American ATC, they would understand him on the set script, and it’s gonna depend on who the person is with as to how they get by. If you’re gonna say, “can he just go with anybody?”— no, I don’t think so. Could I work with him? Maybe because of where I’ve been, yes, but that’s a qualified on. ... Other pilots, no but depends on background. (Participant 9)

There was concern among the pilot participants that although this speaker would be fine operating using standard phraseology, that his language ability would not be sufficient when using plain language. The excerpt below by Participants 7 and 9 are exemplifications of this concern:

Communicating effectively with ATC, he’s doubtful on that one, because it’s scripted what they say. But if they step off the script, this is the problem. If he just followed everything, as per SOP, he could probably get by, but the moment they step off that... so I’d have to say, you know, from a safety angle, “no.” (Participant 7)

In the area where you use set script stuff, I’m sure [he] would be fine. So if he had to respond to ATC in a standard fashion, I don’t think there’s be a problem. Ah ... communicating in plain language with other pilots ... difficult. (Participant 9)

But the main concern among the pilots seemed to be Speaker 4’s lack of technical knowledge. This concern was raised by nine of the ten pilots participating in the study. This seemed to be the one of the main reasons, the pilot was rated as communicatively ineffective. Participant 4 below exemplifies other pilot’s comments:

I got the impression that he didn’t know the answer. That airspeed indicator example ... he certainly had communication issues getting it out but I don’t think he knew what to say even in his own language, to be honest. It just sounded like he was largely making it up. (Pilot 4)
Similarly, referring to the same test item, Pilot 5 noted:

_He was trying to explain why the air speed was basically low. ... I don’t think he knew the actual aerodynamics of why._ (Pilot 5)

Summing up, it seems that the reasons why most of the pilots did not rate Speaker 4’s language ability as sufficient for operational flying had to do with two concerns: (a) the speakers lack of technical knowledge and (b) the concern that the speaker would struggle in the transition from standard phraseology to plain language.

_Speaker 6_

Speaker 6 was the opposite case to Speaker 4. He was rated at Level 3 by the language-trained raters but more than half of the pilot participants considered him adequate to work in an operational environment. Again, the interview comments were scrutinized to arrive at an explanation for this disparity.

Just as was reported for Speaker 4 above, the pilot participants commented on a range of linguistic features in Speaker 6’s speech sample. In fact, they mentioned a wider range of linguistic features than they had done in response to Speaker 4’s sample.

Seven of the pilots commented on Speaker 6’s lack of fluency. For example, Participant 6 said: “I think there’s hesitancy there, too, so he doesn’t have the fluency we are looking for.”

There was no consensus among the participants whether they thought Speaker 6 lacked comprehension or not. For example, Pilot participant 7 said: “It seems to me he’s listening to what he’s being asked, and picking out the words he knows. Therefore, his comprehension is gone.” Similarly, Pilot 5 commented: “He hasn’t quite grasped ... he hasn’t got comprehension ... of what’s being asked of him.” On the other hand, Pilot 8 said: “And he seemed to understand the questioner as well.” Similarly, Pilot 3 pointed out that it was positive that Speaker 6 asked for clarification if he didn’t understand a question.

Pilot 6 thought that the speaker lacked sentence structure and vocabulary. Five pilots commented on the speaker’s accent and pronunciation. Pilots 7 and 9 mentioned that the speaker pronounced the words clearly and was therefore easy to understand. Pilot 2 noted, however, that he thought Speaker 6 had a very strong accent, but that he would be able to get used to that. Interestingly, Pilot 3 said that it was possible to hear a certain confidence in the speech sample, and that this was related to the tone of voice of the speaker.

Again, moving from standard phraseology to plain English was a concern. Below is an excerpt from Pilot 2 on that topic:

_Possibly with standard phraseologies, he might be okay, but if there’s anything non-standard like we heard with, um, US ATC, he’d be lost. So, I think that with ATC’s phraseology, with pure standard phraseology, I think he’d be okay, but that’s not always the case, so I say... he’d be okay with pilots, cause we know what we’re talking about, but with ATC, I don’t think so._
Similarly to Pilot 2 above, Pilot 3 also made the distinction between other pilots, passengers and ATC.

*He can communicate with other pilots effectively, I think ... again with the visual clues ... clues that he’d be issuing ... but I don’t think he’d be good at ATC and passengers.*

Just as technical knowledge seemed to be the deciding factor in the case of Speaker 4 described above, seven of the pilots commented on Speaker 6’s technical knowledge. In this case, they thought that the speaker had sufficient knowledge to work operationally. Below are some excerpts to exemplify the pilots’ opinions:

*I don’t know if there was a bit of a misunderstanding with the VFR on the first questions, but ... yeah, they seemed to sort of explain all the stuff that they would do ... they understood what they were talking about [...] and all the relevant aviation type terms were there. Yeah, it would’ve been interesting to hear, for example, that overshoot question ... and see what they would come up with then. But yeah, all the aviation related stuff’s all there.* (Pilot 8)

*I could understand him and he put enough information out to give me confidence...* (Pilot 3)

Summing up this section on Speaker 6, it seems that although the language raters thought that his language ability was below operational level, the majority of the pilots thought that he should be allowed to fly. It seems that their ratings of his technical knowledge seemed to have been the deciding factor in this judgment.

**RQ 3: What criteria do pilots use when evaluating the effectiveness of speech samples of other pilots?**

To answer the third research question, the data were coded according to recurring themes. The categories on the ICAO rating scale were automatically included as possible coding categories. These were:

- Pronunciation including accent
- Structure
- Vocabulary
- Fluency including “speed” or “rate of delivery”
- Comprehension
- Interaction including “immediacy” or “delay” in responses

As was anticipated at the outset of the data analysis, references to each of the above categories were found in the interviews. Beyond the categories identified from the ICAO rating scale, the following other themes were also found in the data:

- Pilots’ technical knowledge, experience and level of training
- Overall evaluation of level of speech
- Transition from standard phraseology to plain language
- Visual cues
- Appropriacy of answer
Altogether, therefore, eleven thematic categories were identified and each of these will be exemplified in tables below. Following that, an attempt is made to quantify these results to show which of these categories were mentioned most frequently by the pilots.

Table 9 below presents some sample quotes on the topic of pronunciation. These comments mostly focus on the comprehensibility of the speakers’ accents.

<table>
<thead>
<tr>
<th>Table 9. Comments on Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• I wouldn’t put the pronunciation down as … as, ah, brilliant—and that’s what I mean by saying like he had marbles in his mouth. (Pilot 10 about Speaker 3)</td>
</tr>
<tr>
<td>• I found his accent a little off-putting at first when I first listened to the first few seconds of it, I thought—is he actually speaking English, but then, I’m picking up most of what he’s saying (Pilot 8 about Speaker 3)</td>
</tr>
<tr>
<td>• He’s got a good grasp of the English language … his accent is what’s difficult (Pilot 9 about Speaker 1)</td>
</tr>
</tbody>
</table>

Vocabulary (or problems with vocabulary) was also frequently mentioned in the focus group interviews. Some sample extracts can be seen in Table 10, below.

<table>
<thead>
<tr>
<th>Table 10. Comments on Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>• This one is marginal … again, it’s his vocabulary (Pilot 9 about Speaker 3)</td>
</tr>
<tr>
<td>• She certainly didn’t find the words to get the message across (Pilot 3 about Speaker 2)</td>
</tr>
</tbody>
</table>

Comments relating to speakers’ fluency related to the speed of speech (i.e., the rate of the delivery) and the number of hesitations and pauses. Sample extracts are in Table 11, below.

<table>
<thead>
<tr>
<th>Table 11. Comments on Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>• I think that was fine. Just needs to slow the rate of speech down, again. (Pilot 10 about Speaker 4)</td>
</tr>
<tr>
<td>• He spoke continuously, rather than stopping and going (Pilot 8 about Speaker 6)</td>
</tr>
</tbody>
</table>

Comprehension was often mentioned by the pilots. There was concern that some speakers did not understand the questions posed to them or that although they had understood what was asked, were not able to answer the question due to a lack of English proficiency. Table 12 shows some sample responses.
Table 12. Comments on Comprehension

- So then … he really got onto the runways, so for me, comprehension wasn’t there, ‘cos he missed the point of the question (Pilot 7 about Speaker 4)
- He seemed to lack understanding (Pilot 8 about Speaker 8)
- Again, I thought he had trouble actually understanding the question (Pilot 3 about Speaker 4)

Although not directly mentioned in the focus group interviews, there were several comments that touched on the interaction of the interviewer and the speakers. These focused on two areas: (a) the immediacy of the response and (b) whether speakers asked for clarification if they did not understand a question. Table 13, below, shows some examples.

Table 13. Comments on Interaction

- She couldn’t respond to the questions in a reasonable time (Pilot 9 about Speaker 2)
- That can be fairly upsetting thing when you know you can’t communicate but there’s ways you can react to that. You can confidently say, “I’m sorry, I don’t understand your question, “or you can … get very nervous and upset, which she did. Like, rattled, and make stuff up. So she’s not responding in a very sound way (Pilot 2 about Speaker 2)
- He responds slowly […] there is an impatience over the airwaves … so, you will find that people sort of jump in to get the answer out. So there is that impatience of being able to respond. (Pilot 4 about Speaker 8)

The pilot participants relied on many occasions on their judgment of a candidates’ technical knowledge to decide on a level. Interestingly, they also repeatedly discussed the level of pilot training a candidate must have. As mentioned above under Research Question 2, the absence or presence of such technical knowledge was used as the final deciding factor if the candidate was perceived to be borderline. Some quotes can be seen in Table 14, below.

Table 14. Comments on Pilots’ Technical Knowledge

- I think he spoke in some aviation-related terms, and so, with the question, you’re gonna go, “here’s the answer I hope to hear,” and you go, “yep—that’s what I heard, that’s what I heard, that’s what I heard,” so… that was good. That’s why I thought he was effective (Pilot 10 about Speaker 7).
- I could fly with him. I would, um … he wouldn’t be a captain level in my opinion … er, he may be, but … I wouldn’t feel particularly comfortable going to a busy international airport [with him] (Pilot 7 about Speaker 1)
- I got the impression he didn’t know the answer. That airspeed indicator example … he certainly had communication issues getting it out but I don’t think he knew what to say even in his own language, to be honest. It just sounded like he was largely making it up. (Pilot 1 about Speaker 4)
- He can communicate effectively. He is an airline pilot or something (Pilot 5 about Speaker 9)
Several pilots also made use of an overall evaluation of a candidate’s speech. These comments could not be classified into any categories as such as they pertained to the overall communicative effectiveness of the speech (see Table 15 below).

Table 15. Comments on Overall Speech Performance

- Lack of basic English language skills (Pilot 5 about Speaker 2)
- I wouldn’t work with him and I wouldn’t say that he was communicating effectively for aviation, yeah (Pilot 6 about Speaker 6)

A number of comments were in reference to the point where pilots need to make the transition from using standard phraseology to plain language. This was mentioned again and again as the important point in the TLU domain where lower level pilots would not cope. Some sample extracts are below in Table 16.

Table 16. Comments on Transition from Standard Phraseology to Plain English

- Generally ATC … we use standard phrases. Similarly, on the flight deck,, we use standard calls and procedures… but if something was to go wrong… for example, if she was having an engine stop or engine failure, ah, it may require immediate actions where one pilot—we fly multicrew so we’re not by ourselves—one pilot will do one thing and will ask the other person to do something which may not be standard… something that you don’t do every day … and if they’re then having to think about, a) what did they ask me or b) did I understand it correctly? For example, you may ask him to do something, they may misunderstood and the incorrect thing … […] it might be open to interpretation and then it’s quite difficult perhaps, to ah stress the urgency of what they need done by ATC as well. (Pilot 8 about Speaker 2)
- Possibly with standard phraseologies, he might be ok, but if there’s anything non-standard like we heard with US ATC, he’d be lost (Pilot 4 about Speaker 6)

Interestingly, although visual clues can clearly not be used in radiotelephony communication between pilots and ATC, a number of the pilots involved in the interviews mentioned just how much they relied on visual clues (or nonverbal communication) in multicrew flights. Some typical quotes are in Table 17, below.

Table 17. Comments on Visual Clues

- You have a lot of nonverbal communication in a cockpit as well. There might be times where something’s going on an you’ll both look at it and you’ll pick up on each other’s body language. (Pilot 8)
- As far as a pilot’s concerned, a lot of our communication is not just verbal. It is ... gestures (Pilot 2)
Very occasionally pilots also mentioned that the appropriacy of an answer was important for the judgement of communicative effectiveness. A sample quote can be seen in Table 18, below.

Table 18. Comments on Appropriacy of Answer

- He responded positively in my mind (Pilot 2 about Speaker 7)

Finally, the sentence structure of a speaker’s speech was evaluated. An example extract is in Table 19, below.

Table 19. Comments on Sentence Structure

- It was the way he was being able … he was able to say it with structured sentences. It made sense

**Quantification of Categories**

To be able to suggest changes to the ICAO rating scale, it is important to understand the construct of English used in aviation. Therefore, it is not only important to know which features in the performances were important to the pilot participants, it is also crucial to know which of these features are mentioned more often than others. For this reason, all tokens of each of the categories were counted. Then percentages were calculated for each category. Table 20 below presents each category in order of the number of times it was mentioned in the interviews. The final column indicates whether each category is mentioned in the ICAO rating scale.

Table 20. Focus Group Interview Categories

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>N</th>
<th>Percentage</th>
<th>ICAO Scale?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Technical knowledge, experience, training level</td>
<td>50</td>
<td>23.26</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Pronunciation</td>
<td>45</td>
<td>20.93</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Comprehension</td>
<td>27</td>
<td>12.58</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Overall evaluation</td>
<td>24</td>
<td>11.16</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Fluency</td>
<td>22</td>
<td>10.23</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>Transition from standard phraseology to plain speech</td>
<td>13</td>
<td>6.05</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Vocabulary</td>
<td>12</td>
<td>5.58</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>Interaction</td>
<td>8</td>
<td>3.72</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>Visual clues</td>
<td>6</td>
<td>2.79</td>
<td>Yes</td>
</tr>
<tr>
<td>10=</td>
<td>Appropriacy of answer</td>
<td>4</td>
<td>1.86</td>
<td>No</td>
</tr>
<tr>
<td>10=</td>
<td>Sentence structure</td>
<td>4</td>
<td>1.86</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>215</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
Table 20 above shows that a pilot’s level of technical knowledge, experience, and training level was most often mentioned in the pilot interviews and that this played an important factor in their evaluation of the effectiveness of the speech sample. This was closely followed in importance by the evaluation of a test taker’s pronunciation. Comprehension, overall evaluation and fluency were of similar importance to the pilots (these categories were mentioned 27, 24 and 22 times, respectively, in the transcripts). The point of transition from standard phraseology to plain speech and a test taker’s vocabulary were mentioned almost equally as often. References to interaction, visual clues, appropriacy of an answer and sentence structure were less common in the interviews.

Discussion and Conclusion

The findings are discussed in two sections: (1) recommendations for possible revisions to the ICAO ratings scale and (2) the efficacy of using stakeholders in post hoc rating scale validation.

Recommendations for Possible Revisions of the ICAO Rating Scale

The questionnaire results showed the disparity of opinions among stakeholders about the ICAO scale, however some clear trends in the data were apparent. Firstly, a number of users of the ICAO rating scale were mindful of the fact that the scale was not developed for the use of aviation specialists as raters and that raters from such a background would either need thorough training in the understanding of such terms as “register,” “varying speech flow for stylistic effect,” etc., or would need to use a different rating scale altogether. Secondly, Level 4 was regarded by only about half of the questionnaire respondents as the adequate level for operational flying. Evidence from the focus group interviews suggests that Level 4 is adequate for experienced pilots with sufficient technical knowledge while it might be too low for student pilots. Thirdly, the open-ended questionnaire responses suggest a number of possible areas of improvement of individual scale categories. Possible recommendations are listed below in bullet-pointed format.

Fluency:

- “Varying flow for stylistic effect” should not be included in the category of fluency. This point seems to be already covered under pronunciation.
- Discourse markers were not found to be typical of the type of speech used in the target language domain (TLU).
- Fluency should only contain references to speed of speech, pausing, hesitations and possibly to fluency at the transition point between standard phraseology and plain language.

Structure:

- There were hardly any references to structure in the focus group interviews, so one might conclude that correct structure is less important to the stakeholders in the TLU domain.
Vocabulary:
- Any references to idiomatic language should be deleted as this is not appropriate in the TLU domain.
- “Register” should possibly be changed to “appropriacy.”
- Reference to “common, concrete and work-related topics” is not clear to stakeholders as the whole test should be in the aviation domain.

Pronunciation:
- Reference to L1 influence should be deleted.

Comprehension:
- This cannot really be measured in a speaking test but it was common in the pilots’ reactions.
- Selecting speakers for tests that satisfy the requirement of an “accent sufficiently intelligible for international users” is problematic – it might be better deleting this.

Interaction:
- Nonverbal clues are not typical of the TLU domain (i.e., radiotelephony speech) although pilots in multicrew environments seem to rely on these. It is possible to conclude from this that there should possibly be different tests and scales for pilots at different levels.
- Descriptors for levels 5 and 6 should be switched.
- Immediacy of response and asking for clarification seemed important to pilots.

Overall, Level 6 of the scale seemed problematic and might need revision. Also, there seemed to be some concern that the individual trait scales are not congruent within themselves or with the ICAO manual.

Using Stakeholders in Validation of LSP Criteria

The second part of the discussion now turns to the use of stakeholders in post hoc validation of LSP assessment criteria. The ICAO scale was developed by members of the PRICE study group with little funding. Because the scale, as well as the accompanying manual, are so central to the implementation of the ICAO policy, post hoc validation work is important and this study provides one piece in the larger puzzle. The study showed that through the questionnaire, a wide variety of stakeholders was relatively easy to canvass. These people offered a number of important insights into their problems or successes of using the scale in their respective assessment contexts. The experiences and views these people were able to offer were varied. One might argue that using stakeholders in such validation work is difficult as stakeholders’ views are often considered as naïve, but the background data of the questionnaire show that these people have much to bring to the area of aviation testing.

In fact, making use of a combination of professional and language (testing) experts is central to LSP testing. These two groups of people should be canvassed to provide a broader range of validity evidence. Especially in such a policy-driven situation as the implementation of the ICAO language requirements, it is crucial not to take all aspects of the policy as given. It is clear that the ICAO rating scale needs post hoc validation work and this study has shown that stakeholders can be used successfully to help with some aspects of such work. In fact, the
results show how detailed such feedback can be and that the comments are useful in refining the scale to more appropriately act as the representation of the test construct.

The results from the focus groups that were set up to establish the indigenous assessment criteria used by pilots were able to show that such research is also useful for post hoc validation studies. The findings show which rating scale criteria were less important to pilots (e.g., structure). They also show how much importance the pilots put on the speakers’ technical knowledge and experience and just how difficult it was for them to separate language ability from aviation knowledge. Jacoby and McNamara (1999) make the point that it is not simple to make the transition from the indigenous assessment criteria derived from a TLU situation to criteria that will be employed in a test. It is therefore possible that such post hoc validation work as was employed for this study could be useful to suggest changes to a previously existing scale such as the one developed by ICAO.

The transcripts of the focus groups show just how difficult it is for industry professionals such as the pilots to divide language ability from aviation knowledge, a point that has been made before (e.g., Davies, 2001; Ryan, 2007). Especially in this specific context—namely aviation English testing, which is quite different from, for example, testing English for medical professionals because of the prominence of the use of standard phraseology—it is possible that the testing of language and technical knowledge cannot or should not be separated. This also relates to the use of the policy surrounding raters from aviation and nonaviation backgrounds. ICAO recommends (but doesn’t mandate) the use of both aviation experts and English language experts. This study was able to show that the standards industry specialists and language experts used when rating the speech samples were slightly different. They agreed on the ratings of most candidates, but there were some significant differences. This was also found in other, similar studies (e.g., Douglas & Myer, 2000; Ryan, 2007). The findings of this study suggest that it is important to involve industry specialists in the rating process, especially around the Level 3/Level 4 cutoff as language experts might not have sufficient insight into the TLU domain.

As with most studies, there are a number of notable limitations to this research. Firstly, although every attempt was made to get a large response rate for the questionnaire and to ensure that the respondents are all actively engaged in Aviation English testing, not enough information about the background of the respondents was canvassed and it is not clear if all the stakeholder groups around the world are sufficiently captured. In saying that, the background data that is presented in the methods section points to the fact that there was a good spread of language and aviation specialists and that stakeholders from a number of countries completed the questionnaire.

There are also several limitations with respect to the focus group interviews. Firstly, the sample size was of course small, with only ten pilots taking part in the study. Secondly, most of the pilots were from a very similar background, working in large multicrew environments. Only one of the pilots was from a non–English speaking background and none were private pilots (mainly because private pilots in Australia have very little opportunity to interact with nonnative speakers via radiotelephony). The speech samples that were selected were mainly produced by nonnative speakers from Asian backgrounds. It would have been good to include a wider variety of speech samples, and this could be a suggestion for future research.

It can further be argued that the use of speech samples collected from aviation tests is fundamentally flawed as it is possible that the tests are not actually successfully tapping into
what happens in the target language use domain and therefore representative of the construct. If the test performances played to the pilots are not representative of the target use domain, then of course the indigenous assessment criteria elicited from the pilots might not be generalizable back to the TLU domain. However, the language produced during radiotelephony communications is mostly composed of the use of standard phraseology with small parts of plain language when standard phraseology does not suffice. Because aviation English tests should mainly test candidates’ ability to produce plain English in an aviation context, the speech samples from the selected tests were the closest possible. For this reason, speech samples from a number of tests were chosen.

The study presented above shows an attempt to involve stakeholders on a number of levels in post hoc validation work of a LSP rating scale. The findings show that it is possible to engage different stakeholder groups in such a process. The interpretations we make based on candidates’ test performance have a good chance of being improved and more generalizable if we engage in such validation work. However, it is clear that this study provides just one piece in the puzzle necessary to fully validate the ICAO rating scale.

Acknowledgements
I would like to thank the English Language Institute at the University of Michigan for providing the funding for this study, without which this study would not have been possible.

References


ABSTRACT  As integrated writing tasks appear in more assessments of writing for academic purposes, research on their validity and usefulness is needed to assist users in interpreting scores. This study investigated features of writing from integrated reading-writing tasks as well as the process writers’ used to complete them. One hundred and thirty-one undergraduate students at a United Arab Emirates university completed a reading-writing task on the topic of global warming followed by a questionnaire on their writing process. Their writing was scored and divided into three levels of writing proficiency. The written products were analyzed for discourse features, including lexical sophistication, syntactic complexity, accuracy, and fluency. They were also analyzed for source use by determining the amount of source use, direct (quotation) or indirect integration, and verbatim source use. All of these features became dependent variables in one-way analyses of variance to find if they differed significantly across the writing proficiency levels. The process-questionnaire items that related to source text use were also analyzed across the proficiency levels. The results indicated significant differences across levels for a number of discourse and source-use features; however, follow-up analysis indicated that the differences were greater between the lowest level and the upper two levels. The upper levels were not significantly different in terms of the writing features, which suggests that the writing at the higher levels was distinguished by other aspects such as organization, content, or coherence, which were on the rating rubric. The process results reveal all writers reporting used the source texts for ideas and forming opinions. Differences did appear in difficulties experienced by writers, coursework on integrated writing, and borrowing/citation from the readings. The implications for the study hold that integrated writing may be more distinguished at lower levels for discourse features, but other textual features such as cohesion, content, or organization differentiate higher-level writing. Further, a construct of integrated writing should include reading proficiency and knowledge about integrating reading with writing.
Integrated reading-writing tasks are increasing in popularity and either replacing or complementing writing-only independent tasks used in assessing academic writing. The integrated tasks are seen to have more authenticity (Feak & Dobson, 1996; Weigle 2002, 2004) and may provide test takers with content, lowering anxiety and creativity demands on writing (Plakans, 2008; Read, 1990). These benefits extend to potentially positive washback effects in academic writing classrooms. (Cumming, Grant, Mulcahy-Ernt, & Powers, 2004). However, problems exist for these tasks, such as development, plagiarism, and construct-related validity.

In terms of validity, more is needed to understand what scores from integrated tasks infer about English language writing ability. One component of a validity argument is “explanation,” which requires evidence such as the connection between an expected score and academic language proficiency (Chapelle, Enright, & Jamieson, 2008). Some research has looked at discourse features and process in writing assessment with relation to proficiency (Engber, 1995; Ishikawa, 1995; Jarvis et al., 2003; Machon et al., 2000; Ortega, 2003; Sasaki, 2000); however, most studies addressed the relationship based on impromptu writing tasks. Although holistic scores on independent and integrated tasks have been found to correlate (Brown, Hilgers, & Marsella, 1991; Gebril, 2006; Lewkowicz, 1994), the written products have been shown to have significantly different discourse features (Cumming, Kantor, Baba, Erdosy, Eouanzou, & James, 2005, 2006). Thus, it may be inconclusive to base score interpretations for integrated tasks on research from independent writing tasks. To fill this gap, our research explored the connections between writers’ scores on integrated test tasks and features of the written products as well as the process in completing the tasks with a group of writers from the Middle East.

In a study of TOEFL prototype tasks, Cumming et al. (2005, 2006) investigated the writing features of three kinds of tasks: writing-only, reading-writing, and listening-writing. They analyzed and compared the written products across tasks and across proficiency levels emphasizing the importance of such analyses:

A related issue concerns knowing if and how the written discourse may vary in the written compositions produced by examinees at different score levels on the test. The discourse of written texts cannot be assumed consistent for examinees with differing levels of proficiency in English, so consideration also needs to be given to how the written discourse of examinees varies in particular tasks with their English proficiency. This information is needed to verify, or refine, the scoring schemes being developed to evaluate examinees’ performance on these writing tasks. (2005: 8–9).

In their study, they found that, across proficiency levels, differences occurred in both essay length and number of clauses, indicators of syntactic complexity. The higher proficiency writers also used more variety in their word choice and scored higher on a grammatical accuracy scale. Lastly, the higher proficiency writers wrote longer compositions than the other groups, a finding that has appeared in other studies of writing assessment (Watanabe, 2001).

Due to the use of sources in integrated tasks, score interpretations should also consider how writers integrate the reading material. A few research studies have looked at integration
style in these tasks. Watanabe (2001) studied the use of summary, paraphrase, and quotation in writers’ products. He discovered that quotation was used most, with some instances of paraphrasing and quotation. However, his analysis did not extend to the relation between these occurrences and writers’ proficiency. Cumming et al. (2005, 2006) did consider this aspect of integration style finding summarizing more common for higher proficiency writers, along with better integration of the readings. Those writers who were midrange relied more on paraphrasing and verbatim source use, while the least proficient writers used sources least. The authors suggested that this last group struggled more with comprehension, thus did not use the content from source texts in their essays.

Another line of research has focused on writers’ verbatim source or plagiarism (Campbell, 1990; Currie, 1998; Deckert, 1993; Johns & Mayes, 1990; O’Connor, 2003; Pennycook, 1996; Rinnert and Kobayashi, 2005; Scanlon & Neumann, 2002; Sutherland-Smith, 2005), suggesting that there is a relation between proficiency and verbatim source use. Johns and Mayes (1990) studied the writing of nonnative writers at two proficiency levels on a summary task. The findings showed no significant difference between the two proficiency groups in their textual borrowing in terms of distortion or sentence replication. However, other studies have found that proficiency impacts source use (Campbell, 1990; Currie, 1998; Cumming et al., 2005). Campbell concluded, “From this study it is seen that language proficiency affects the use of information from background reading text in academic writing” (p. 224). Thus, the impact of L2 proficiency on verbatim source use is possible, and the inconsistent results might be attributable to how researchers define and measure proficiency. Studies of L2 writers’ source use outside of the assessment field have focused on either ESL students who study in North America or EFL students studying in South Asia. We are not aware of any studies that discuss the issue of source integration and plagiarism with students from the Middle East.

In addition to uncovering relationships between integrated writing test scores and the written products, studying the test taker’s process in completing these tasks is also important in explaining some of the differences or similarities found in the written products and to provide validity evidence. Studying process allows test developers to understand if the test measures accurately what we want to measure; in other words, validity (Ascensión, 2005; Cohen, 1998, 2007; Cohen & Upton, 2007; Plakans, 2008; Rupp, Ferne, & Choi, 2006). Bachman (2004) made this clear by stating: “There are two aspects of test performance that we need to investigate in our evaluation of test usefulness: the processes or strategies test takers use in responding to specific test tasks and the product of those processes or strategies…. In order to evaluate the usefulness of a given test, we need to investigate both aspects” (p. 5). Many of the recent process studies have analyzed data to learn about writer characteristics and strategies used in the process of writing. Research by Cumming (1991), Cohen (1994), Bosher (1998), and Sasaki (2000) has addressed these issues in L2 writing. In their studies, both writing-only and reading-to-write tasks are used. They include variables related to writer characteristics such as L2 proficiency, writing skill, writing expertise in L1 and L2, and instruction. This research on writer characteristics has clarified some important points in the L2 writing process. First of all, writing expertise has a strong impact on the L2 writing process; this impact is separate from L2 proficiency and perhaps more central. In terms of strategies used in the process of L2 composing, writers with higher L2 proficiency and more expertise and skill use more strategies.
To build on previous research with integrated writing tasks, this study explored how discourse features, source text use, and process interact with writing proficiency in reading-writing tasks for students in a Middle Eastern context. The following research questions guided this study of integrated tasks:

1. Does writing proficiency level affect discourse features?
2. Does writing proficiency level affect use of source texts and verbatim source use?
3. How does writing proficiency level impact writers’ process of source use?

The results hold implications for both writing assessment and instruction. First of all, this study improves our understanding on how Middle Eastern students integrate information from sources in their writing. As mentioned earlier, this population is not given adequate attention in writing assessment research. Secondly, the research yields useful results concerning the writing features and use of sources in the performance of students at different proficiency levels. This information is useful in interpreting test scores from integrated tasks as well as for development of rating rubrics for such tasks. Lastly, the study provides some insight into the processes that L2 test takers follow while writing based on source texts; an area that needs further investigation as suggested by research (Cumming et al., 2005). This information will be helpful in making informed decisions in test development, test use, and writing instruction, such as in creating rating scales or determining areas for curricular focus in teaching academic writing.

Method

This study was conducted with participants from a college of humanities and social sciences in a public university at the United Arab Emirates (UAE). The students who participated in this study represent a number of majors, including applied linguistics, general linguistics, translation studies, communications studies, geography, urban planning, and social work. Two instruments were developed and implemented: an integrated reading-writing task and a process questionnaire. Once the participants completed the tasks, the writing was scored and analyzed for discourse features as well as source use. The items on the questionnaire regarding source use were also processed to look at how writers reported on their source use. This section will detail these methods.

Participants

Students participating in the project were undergraduate students enrolled in humanities and social sciences programs. Courses at the university are taught in English; therefore, participants hold a level of English sufficient for university study because students go through a rigorous language program before choosing a major. Most students had taken the International English Language Testing System (IELTS), a measure of English language proficiency. The average mean score on IELTS was 4.82 with the mode of 4.5 (See Table 1 for participants’ IELTS scores). It is important to mention that those students took the IELTS test after finishing their general education classes and before joining their majors. Most of those students took the IELTS test at least one year before data collection, and some of them took it two to three years earlier. So, because these students’ proficiency level was much better than what the IELTS scores reflect, caution is warranted when interpreting their
language scores. Data were collected from the women’s campus, thus all participants were female.

Initially, 139 students wrote on the tasks, but eight essays were removed from the data set because they were incomplete.

Table 1. Participants’ IELTS Scores

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.82</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.04</td>
</tr>
<tr>
<td>Median</td>
<td>4.75</td>
</tr>
<tr>
<td>Mode</td>
<td>4.50</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Instruments

The Integrated Task

The task for the study was developed to have some characteristics of authentic academic writing. Thus, we chose an argumentative essay prompt and two short reading passages presenting opposing points of view. The argumentative mode was selected based on the recommendations of a number of researchers (Plakans, 2008; Gebril, 2006; Melenhorst, 2006; Cumming et al., 2005). For example, Melenhorst (2006) argues that argumentative texts are approached successfully by students. The topic addressed in this prompt—global warming—was selected with the idea that it would hold students’ interest, holds clear positions that writers can choose, and has possible source texts that included some solid evidence. Once the topic and format were determined, we spent time selecting the reading passages. Authenticity was again attempted, by finding sources from published magazines. One source was written by a credible scientist, while the other source was attributed to Reuters, so an author’s name was added to increase the ease of citation for students. Two passages were selected following Lewkowicz (1994) recommendations to use more than one text. The texts were modified slightly for readability as well as parallel length and difficulty (see Table 2 for more information). In addition, following Carrell’s recommendation (1987) to consider factors such as the nature of students, background knowledge, and cultural issues, the task was given to five experts, who were faculty members at the university where data were collected. They felt that the source texts were too long, some words too difficult, and instructions needed clarity. These issues were addressed in revision.

Table 2. Descriptive Data of the Reading Passages

<table>
<thead>
<tr>
<th></th>
<th>Flesch-Kincaid Grade Level</th>
<th>Flesch Reading Ease</th>
<th>Words</th>
<th>Sentences</th>
<th>Words per sentence</th>
<th>Characters per word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 1</td>
<td>12</td>
<td>47.9</td>
<td>275</td>
<td>11</td>
<td>25</td>
<td>1356</td>
</tr>
<tr>
<td>Text 2</td>
<td>12</td>
<td>33.4</td>
<td>275</td>
<td>15</td>
<td>18.3</td>
<td>1437</td>
</tr>
</tbody>
</table>
This revised version of the task was piloted with 46 undergraduate students from the same university, all of whom had English language levels similar to the research population. They also answered questions following the writing session about the task difficulty. Their responses and writing products were considered for revisions to the tasks. The biggest challenges for writers were difficult vocabulary in the source texts, lacking ideas for their arguments, and organizing ideas. In their writing, some students wrote summaries rather than essays. Time was also attended to in the pilot. Based on this session, the source texts were modified to make the vocabulary less difficult; the instructions and prompt were revised for clarity and offered clearer positions on the issues; and it was determined that the task should have an one-hour limit. The final integrated tasks are included in Appendix A.

The Process Questionnaire

This study employed a questionnaire to elicit writers’ process of composing the tasks. Given the large number of writers and our previous use of concurrent verbal protocols (Plakans, 2008), a questionnaire seemed feasible and had potential to add new information about writers’ process in integrated tasks.

The questionnaire was developed to have a five-point Likert scale format presenting statements about the task and process followed by: 1 (strongly disagree) through 5 (strongly agree). Initially, 55 items were developed with the intention of selecting the best items through a multistage process of piloting. In addition, some questions were included to collect demographic and background information as well as six open-ended questions for qualitative analysis. The statements in the questionnaire were based on research on L1 and L2 writing processes (Anderson, 1991; Esmaeili, 2002; Grabe & Kaplan, 1996; Spivey, 1984, 1990, 1997; Watanabe, 2001) and our own previous research on the writing and reading processes in integrated tasks (Plakans, 2008, in press).

To design the statements, seven interviews of writers who had completed an integrated task were reviewed for process statements. In addition, three writers’ think-aloud transcriptions during an integrated task were analyzed. The questionnaire included statements to consider the research question proposed by this study, but also attempted to provide more information such as writers’ attitudes toward the tasks, task effect, topic effect, planning before writing, and comprehension of reading. Based on these areas of interest and prior research, the main themes of the questionnaire included: reading effect, topic effect, task effect, writing process, validity, reading-to-write vs. writing-only, and source text use. The last area will be the focus of this report.

The next step was to give the questionnaire to four experts: two who have extensive experience with questionnaires in second language acquisition research and two with backgrounds in academic writing. Their feedback provided suggestions on revisions and additions of items as well as formatting and scoring advice leading to a revision of the questionnaire.

This draft of the questionnaire was completed by one student as a think-aloud protocol task following an integrated writing session. This pilot provided information on clarity, ease of answering, relation to writing task, and any other potentially problematic issues with the questionnaire. The transcription and analysis of the think aloud resulted in rewording several questions for more specificity. The questionnaire was given to another student to complete after an integrated task, followed by an interview to find out more specifically how the
questionnaire represented her process. This session also led to rewording, clarification, and addition/deletion of several items.

The next step was to pilot both the questionnaire and the task with 46 students at the university in the UAE. To investigate reliability of the questionnaire, Cronbach’s Alpha was used and yielded a reliability coefficient equal to (0.87), which is a relatively high value. Fourteen items did not perform well when investigating item-total correlations and consequently were deleted. The final version of the questionnaire included 41 items (see Appendix B).

**Procedures**

*Testing Session*

Several testing sessions were held during the spring of 2008 in a classroom at the university. Participants were first provided an information sheet describing the study, the procedures for the writing session, and information on their rights as participants. This information was also given verbally by proctors. If participants agreed to continue, they were given the writing task. Verbal instructions also were given with the written prompt. Then they received the process questionnaire, which they were asked to complete after they finished writing their essays. Students were told they would have one hour to complete the task. Each writer was given an identification number; no names were included with their work to assure confidentiality.

*Scoring*

Essays were rated using a holistic scale adopted from the integrated tasks in the TOEFL iBT but revised for this study based on a rating session with the pilot data to incorporate the following changes: (a) clarifying development and organization in the rubric to match the prompt evaluation criteria, (b) providing instructions for raters on scoring essays with plagiarism or with no source use, and (c) removing any references to listening features (the TOEFL rubric is used with essays written based on both reading and listening sources). From the pilot rating, essays were chosen to represent each score as profile and training essays.

Two raters conducted the rating. Both were experienced ESL teachers who had also taught academic writing. They attended a rater training session conducted by one of the researchers. In the session, the study’s purpose was discussed, and they were acquainted with the task and the rubric. Then the raters and the trainer rated one essay together followed by a discussion of scoring. The same process was followed for three more essays. Then raters scored five essays independently followed by discussion and calibration. This procedure was continued until both raters felt comfortable with the rating scale. Each essay was scored by both raters, with an interrater reliability coefficient of $r = 0.75$. No ratings were more than one level apart; on essays where the raters disagreed, a third rating was done by one of the researchers. In addition to providing a score for each essay, raters were asked to note their level of confidence with the score from 1 = no confidence, 2 = confident, to 3 = very confident. For both raters, this measure revealed that the raters ranged from highly confident to confident about the scores assigned as shown in Table 3.
Once the rating was completed, the scores were used to group the writers into three levels of proficiency for the statistical analysis (see Table 4 for scoring and level information).

<table>
<thead>
<tr>
<th>Scores</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>1–2</td>
</tr>
<tr>
<td>Level 2</td>
<td>3</td>
</tr>
<tr>
<td>Level 3</td>
<td>4–5</td>
</tr>
<tr>
<td></td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>21</td>
</tr>
</tbody>
</table>

### Analysis

The essays and questionnaires were analyzed to answer the three research questions. First, discourse features were analyzed in the essays, followed by a comparison across the three levels of writing proficiency. Then, essays were analyzed for source text use and again compared across proficiency levels. Finally, items from the process questionnaire were grouped to consider aspects of source use across the writing levels. Each of these steps will be described in more detail in this section.

### Discourse Features

Writing research uses many different methods to identify discourse features (Hinkel, 2002). In this study, we adopted guidelines by Cumming et al. establishing that indicators for such analysis should meet the following requirements (2005: 8–9):

1. include a range of discourse features including lexical, syntactical, rhetorical, and pragmatic characteristics
2. be applied reliably and meaningfully
3. possibly show differences between compositions scored at different scale levels

We choose several discourse features as a focus for the analysis: lexical sophistication, syntactic complexity, grammatical accuracy, and fluency. However, before analyzing these separate features, some general characteristics of the essays needed tabulation: T-units, the number of sentences, the number of words, and characters. T-units were marked for all essays by two raters, while the other characteristics were calculated using a Microsoft Word feature.

**Lexical Sophistication.** This feature of discourse has been defined as the size of a writer’s productive vocabulary based on the written essay (Wolf-Quintero, Inagaki, & Kim, 1998). Average word length was used to establish lexical sophistication based on prior research (Biber, 1988; Cumming et al., 2005; Engber, 1995; Frase, Faletti, Ginther, & Grant, 1999; Grant & Ginther, 2000). The proposed study adopted the definition used by Cumming
et al. (2005: 9) for average word length, “the number of characters divided by the number of words per composition.” Microsoft Word was employed in this study to calculate the targeted vocabulary measure.

**Syntactic Complexity.** This feature measures “the range of forms that surface in language production and the degree of sophistication of such forms” (Ortega, 2003: 492). This feature particularly has had many different methods employed in previous research. We followed Ortega’s advice (2003) to include several measures for this discourse feature and adapted methods used by Cumming et al. 2005; Henry, 1996; Homburg, 1984; Perkins, 1980; Tedick, 1990:

1. the mean number of T-units per sentences
2. mean length of T-units

**Grammatical Accuracy.** This feature is a standard assessment of language proficiency and has gained increased interest in second language acquisition and L2 writing assessment (Polio, 1997). However, it remains somewhat elusive because of the difficulty in precisely measuring linguistic accuracy in a quantifiable way. Most of the quantitative measures used to judge linguistic accuracy, such as error counts (Fischer, 1984; Zhang, 1987; Carlisle, 1989; Kepner, 1991) and error count with classification (Bardovi-Harlig & Bofman, 1989; Chastain, 1990; Frantzen, 1995; Kobayashi & Rinnert, 1992; Kroll, 1990) have been problematic due to either the unreliability of coding schemes or not accounting for the error severity (Polio, 1997). For these reasons, this study used a holistic rating of grammatical accuracy adopted from Cumming et al. (2005), that was based on a scale used by Hamp-Lyons and Henning (1991):

1. many severe errors, often affecting comprehensibility
2. some errors but comprehensible to a reader
3. few errors, and comprehensibility seldom obscured for a reader

Two raters used this scale to rate all essays for grammatical accuracy. The inter-rater reliability coefficient for these ratings was $r = 0.94$.

**Fluency.** As is common in most studies of writing features, fluency was determined through word count, which has been successful in differentiating proficiency levels in many studies of second language writing (e.g., Cumming et al., 2005; Hirano, 1991; Larsen-Freeman, 1978; Tedick, 1990).

**Source Text Use**

To answer question three, the use of source texts was analyzed in the essays. This variable required defining kinds of source text use and verbatim source use. Initially, a set of ten essays was selected and each T-unit was coded for summary, paraphrase, and direction quotation. A commercial Internet program called Turnitin.com was used to identify verbatim source use across all 131 essays. Based on this analysis, several decisions were made. First, we found that the distinction between summary and paraphrase was difficult for raters and often caused disagreement. As a solution, the two kinds of source use were combined as “indirect” source use and considered in contrast to direct use. There was some concern with the use of Turnitin.com because it was not clear how the program was identifying verbatim
source use, and it searched the entire Internet as well as its essay base to find verbatim source use. As the two source texts were from Internet sources, many matches came up that were not relevant to verbatim source use in the writing sessions of this study. To resolve this problem, the raters coded “direct” source use as “no quotation” or “with quotation” to identify verbatim source use. The raters defined direct use as others have done (Cummings et al., 2005), but counting strings of three or more words from the original as verbatim source use. Following these guidelines, raters retrained on a small set of essays to achieve an inter-rater reliability of coefficient \( r = 0.97 \), then the raters coded the rest of the essays.

**Statistical Analysis of Discourse Features and Source Use**

Since the main purpose of this study was to investigate the relationship between one independent variable with three proficiency levels and a number of dependent variables, discourse features, a one-way analysis of variance (ANOVA) was used to analyze the results. In case of a statistically significant overall ANOVA, follow-up tests using the Tukey HSD formula were carried out to investigate which pairwise comparisons were significant. In addition, the researchers followed this analysis by a check on the homogeneity of variance, which is an assumption of one-way ANOVA. In case of violations of assumption, the Brown-Forsythe procedure and Dunnette’s C follow-up test, which do not assume homogeneity of variance, were used to check the accuracy of results yielded from the ANOVA procedures that require this assumption. Furthermore, the researchers included descriptive statistics for the different discourse features. The following is a list of the dependent variables included in the analysis:

1. average word length
2. the mean number of T-units per sentences
3. the mean number of words per T-unit
4. total number of words
5. grammatical accuracy
6. indirect source use
7. direct source use with quotation marks
8. direct source use no quotation marks
9. total source use

**Processes**

The questionnaire developed included items addressing many areas of interest, but this report will focus on those relating to the initial research question regarding source use in integrated tasks. These 14 questions were grouped into subcategories so that several items addressed each (shown in Table 5). For each item the mean, standard deviation, median and mode were calculated at each writing proficiency level.

The only qualitative data in the study came from the open-ended interview questions. These questions were analyzed using a semi-inductive method (Charmaz, 2004). First, the responses to the questions were divided across the writing proficiency levels. Then, within each level, a line-by-line coding was conducted to find patterns/themes emerging from the responses. Categories were defined by this initial coding and a second coding was conducted for these issues. This second coding was conducted both within levels and across levels. These categories were integrated with the four groups of questionnaire items to synthesize the
quantitative and qualitative results. This process resulted in a more in-depth understanding of process to amplify the quantitative data from the questionnaire.

Table 5. Questionnaire Item Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source text used for ideas</td>
<td>15. I used some of the ideas from the readings in my essay.</td>
</tr>
<tr>
<td></td>
<td>29. I used examples and ideas from the readings to support my argument in my essay.</td>
</tr>
<tr>
<td></td>
<td>32. I reread the readings while I was writing to find ideas to put in my essay.</td>
</tr>
<tr>
<td>Source text used in organizing</td>
<td>31. I used the readings to help organize my essay.</td>
</tr>
<tr>
<td></td>
<td>23. The reading helped me choose an opinion on the issue.</td>
</tr>
<tr>
<td>Integration process</td>
<td>14. I used some words from the readings when I wrote.</td>
</tr>
<tr>
<td></td>
<td>36. I paraphrased the reading in my writing.</td>
</tr>
<tr>
<td></td>
<td>38. I copied phrases and sentences directly from the reading into my essay.</td>
</tr>
<tr>
<td></td>
<td>35. I used the authors’ names in my essays</td>
</tr>
<tr>
<td>Knowledge of integrating reading-writing</td>
<td>39. I have learned how to use reading with my writing in a class.</td>
</tr>
<tr>
<td></td>
<td>40. I have learned about plagiarism.</td>
</tr>
<tr>
<td></td>
<td>41. I feel comfortable using other people’s ideas in my writing</td>
</tr>
<tr>
<td>General reading-writing process (open-ended question)</td>
<td>43. Explain any difficulties that you had with reading or writing in this task.</td>
</tr>
<tr>
<td></td>
<td>44. If you used the readings for your writing, please describe how you used them.</td>
</tr>
</tbody>
</table>

Results

Discourse Features and Proficiency

A one-way ANOVA was used to investigate the relationship between the independent variable, writing proficiency, and discourse features (shown in Table 6). The independent variable included three proficiency levels: levels 1, 2, and 3. This classification, as described in the methodology section, was obtained based on holistic ratings of students’ essays. As for the dependent variables in this study, the researchers identified the following discourse features: grammatical accuracy, fluency, syntactic complexity, and lexical sophistication. In addition, the analysis looked into the relationship between proficiency and another dependent variable, which is the integration style of reading sources in students’ writing.
Table 6. One-way ANOVA Results of Discourse Features across the Three Proficiency Levels

<table>
<thead>
<tr>
<th>Feature</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>373729.48</td>
<td>186864.739</td>
<td>57.385</td>
<td>.000</td>
<td>.473</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>416810.90</td>
<td>3256.335</td>
<td>43.800</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>790540.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammatical Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>7.328</td>
<td>3.664</td>
<td>14.355</td>
<td>.000</td>
<td>.183</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>32.672</td>
<td>.255</td>
<td>14.355</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>40.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntactic Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words per T-unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>34.552</td>
<td>17.276</td>
<td>1.751</td>
<td>.178</td>
<td>.027</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>1263.094</td>
<td>9.868</td>
<td>1.751</td>
<td>.178</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>1297.646</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean # of T-units per sentence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>3.2</td>
<td>1.6</td>
<td>.707</td>
<td>.495</td>
<td>.111</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>289.686</td>
<td>2.263</td>
<td>.707</td>
<td>.495</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>292.887</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical Sophistication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Word Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>.096</td>
<td>.048</td>
<td>.789</td>
<td>.456</td>
<td>.012</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>7.796</td>
<td>.061</td>
<td>.789</td>
<td>.456</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>7.892</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Source Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>2.894</td>
<td>1.447</td>
<td>1.627</td>
<td>.201</td>
<td>.025</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>113.839</td>
<td>.889</td>
<td>1.627</td>
<td>.201</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>116.733</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct Without Quotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>2.180</td>
<td>1.090</td>
<td>.447</td>
<td>.640</td>
<td>.007</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>311.927</td>
<td>2.437</td>
<td>.447</td>
<td>.640</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>314.107</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indirect Source Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>373729.48</td>
<td>43.905</td>
<td>10.094</td>
<td>.000</td>
<td>.136</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>416810.90</td>
<td>4.350</td>
<td>10.094</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>790540.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total # of Source Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Groups</td>
<td>2.00</td>
<td>88.283</td>
<td>44.142</td>
<td>6.315</td>
<td>.002</td>
<td>.09</td>
</tr>
<tr>
<td>Within Groups</td>
<td>128.00</td>
<td>894.648</td>
<td>6.989</td>
<td>6.315</td>
<td>.002</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>130.00</td>
<td>982.931</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fluency

Fluency was estimated based on the text length (the total number of words in each essay). As shown in Table 6 and Figure 1, there were statistically significant differences in...
text length across the three proficiency levels ($F(2, 128) = 57.385, p < .001, \eta^2 = .473$). Since Levene’s Homogeneity of variance test revealed that the population variances across the three groups were not equal and therefore the homogeneity of variance assumption was violated, the researchers used the Brown-Forsythe formula to check the accuracy of the previous results. The Brown-Forsythe formula yielded similar results ($F = 43.80, p < .001$), which confirms the results obtained when homogeneity of variance is assumed. In general, this result showed that L2 writers produce longer essays when their proficiency improves.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>Level 1</th>
<th>Level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>135.8033</td>
<td>40.88105</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>226.1633</td>
<td>63.77119</td>
<td>-90.35999*</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>268.8571</td>
<td>77.88792</td>
<td>-133.05386*</td>
<td>-42.69388*</td>
</tr>
</tbody>
</table>

* statistically significant difference using Tukey HSD ($P = .05$)

Follow-up tests were conducted to assess pairwise differences among the three proficiency levels using the Tukey HSD test. Table 7 provides the means and the standards deviations of the three levels and also the results of the pairwise comparisons. As shown in Table 7, there were statistically significant differences between the mean of Level 1 and both the means of the other two groups. In addition, when the means of both Level 2 and Level 3 groups were compared, significant results were also obtained. Another follow-up test, Dunett’s C, which does not assume homogeneity of variance was employed and yielded similar values to Tukey’s.
**Grammatical Accuracy**

Grammatical accuracy was calculated based on a three-point holistic rubric used by two independent raters. As shown in Table 6, the ANOVA results yielded statistically significant differences across the means of three proficiency levels ($F (2, 128) = 14.355, p < 0.001, \eta^2 = .183$). As shown in Table 8, the mean ratings of grammatical accuracy improved as the proficiency level got higher: (M = 1.7541 for the Level 1, M = 2.1633 for Level 2, and M = 2.333 for Level 3). In order to investigate whether these means are statistically different, the Tukey HSD test was employed. The results illustrated that the mean of the Level 1 group was statistically different from the other two groups. However, Table 8 shows that the mean of the advanced group is higher than that of the intermediate group; the pairwise comparison did not yield any statistically significant differences between the two means. This result showed the grammatical accuracy of Level 1 students improving when they move to the next higher level, while this is not the case in the other two groups.

| Table 8. Pairwise Comparisons of the Grammatical Accuracy Variable |
|-----------------------------|---------------------|------------------|------------------|
|                             | M       | SD     | Level 1 | Level 2 |
| Level 1                     | 1.7541  | .53714 |        |         |
| Level 2                     | 2.1633  | .47201 | -.40917*|         |
| Level 3                     | 2.3333  | .48305 | -.57923*| -.17007 |

* statistically significant difference using Tukey HSD (P = .05)

**Lexical Sophistication**

Average word length was used to measure lexical sophistication of students’ essays based on the reading-to-write task. As shown in Table 6, results of the one-way ANOVA demonstrated no statistically significant differences in lexical sophistication across the three proficiency levels ($F (2,128) = .789, p=.456$). This result is due to the similar values of the means across the three proficiency groups (M = 4.63 for Level 1, M = 4.76 for Level 2, and M = 4.71 for Level 3).

**Syntactic Complexity**

Syntactic complexity was calculated based on the average number of words in each T-unit and the mean number of T-units per sentences. As shown in Table 6, the ANOVA results showed no statistically significant differences in the mean number of words in a T-unit across the different proficiency levels ($F (2, 128) = 1.751, p =.178$). This result was expected given the similar values of means obtained from students at different proficiency levels (M = 13.3256 for Level 1, M = 14.1535 for Level 2, and M = 14.6448 for Level 3). As for the second syntactic complexity measure, the mean number of T-units per sentences, the results found no statistical significant differences across the three proficiency levels, which is similar to the results yielded in the analysis of the mean number of words in each T-unit.

**Source Use in Reading-to-Write Tasks**

In order to investigate the source use patterns in reading-to-write tasks across different proficiency levels, the following indicators were used:

1. direct source use with quotation
2. direct Source use without quotation
3. indirect source use
4. total number of source use in each essay

The first feature addressed in this section is the direct use of the sources using quotations marks. The ANOVA results showed no statistically significant differences in the use of quotations across different proficiency levels \((F(2, 128) = .447, p = .64)\). The second feature, direct source use without quotations, also revealed no statistically significant differences \((F(2, 128) = 1.627, p = .201)\) as shown in Table 6.

Results of the indirect source use yielded statistically significant differences among Level 1, Level 2, and Level 3 students \((F(2, 128) = 10.094, p < .001, \eta^2 = .136)\). Levene’s homogeneity of variance test showed that the population variances were not equal. Given this issue, the Brown-Forsythe formula was used to verify the accuracy of the previous results. Brown-Forsythe formula yielded very similar results \((F = 9.955, p < .001)\) and consequently the null hypothesis was rejected. Following the rejection of the null hypothesis, the Tukey HSD and Dunnett’s C procedures were used to compare the means of the three groups. Results of the pairwise comparisons, in Table 11, indicated that Level 2 and Level 3 students showed similar indirect use while the mean of Level 1 students was statistically different from the means of those two groups.

The last feature addressed in this section is the total amount of source use in students’ essays. This feature was calculated by tallying the number of source use incidents based on data obtained from the last three features: direct source use, direct source use without quotations, and indirect source use. Analysis of variance found statistically significant differences in the total number of source use incidents across the three proficiency levels \((F(2, 128) = 6.315, p = .002, \eta^2 = .09)\) indicating that the proficiency level affected the amount of source use in students’ writing. The follow-up test demonstrated that there were statistically significant differences only in source use between the means of Level 1 and both Level 2 and Level 3 students as shown in Table 11. However, there were no significant differences between the means of the Level 2 and Level 3 groups in the total source use. This result is expected given the fact that both groups had similar means \((M = 4.7143\) for Level 2 and \(M = 4.8095\) for Level 3).

Table 11. Pairwise comparisons of the source use indicators

<table>
<thead>
<tr>
<th>Proficiency Level</th>
<th>M</th>
<th>SD</th>
<th>Level 1</th>
<th>Level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect Source Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>1.9016</td>
<td>1.57803</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>3.5510</td>
<td>2.63044</td>
<td>1.64938*</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>3.5238</td>
<td>1.93956</td>
<td>1.62217*</td>
<td>.02721</td>
</tr>
<tr>
<td>Total Source Use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>3.0984</td>
<td>2.05425</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>4.7143</td>
<td>3.09570</td>
<td>-1.61593*</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>4.8095</td>
<td>2.80391</td>
<td>-1.71116*</td>
<td>-.09524</td>
</tr>
</tbody>
</table>

*statistically significant difference using Tukey HSD \((P = .05)\)
Writers’ Reported Source Use and Proficiency Level

In addition to exploring writers’ differences in discourse features and source text use, a questionnaire was used to delve into how their processes differed. Particularly this study focused on the use of the source texts. In this section, the results from analyzing the questionnaire will be detailed in terms of source text use for ideas and organization; integration processes; knowledge of integrating reading and writing; task difficulties; and integration style.

Source Text used for Ideas and Organizing

Consistent with other studies of integrated task processes, source texts were seen by writers at all levels as a resource for ideas (Plakans, 2008). Table 12 shows that all levels agreed that they used some of the ideas from the readings in their essays, with Level 2 having a slightly higher mean and most writers choosing strongly agree. In the open-ended questionnaire many writers reported using ideas from the texts; for example, a writer at Level 2 said, “I used this article to increase my information and I take some sentences to help me when I wrote essay and I benefit from ideas I read it.”

Table 12. Question 15: “I used some of the ideas from the readings in my essay.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.57</td>
<td>3.80</td>
<td>3.62</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.06</td>
<td>1.17</td>
<td>1.24</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mode</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

Question 29 was similar to question 15, but with a more specific purpose for using the source text ideas: supporting an argument. The task instructions asked writers to use the source texts in supporting their ideas, which encourages this strategy. All writers agreed that to some extent they used the ideas in the source texts as examples, as Table 13 illustrates. Interestingly, many at Level 1 and Level 3 strongly agreed with this, while most writers at Level 2 fell in the middle between agreeing and disagreeing with the statement. For Level 2 this contrasts with their answer to question 15, suggesting that they may be using the source texts to generate their own ideas, but not necessarily connecting author’s ideas to their essays.

Table 13. Question 29: “I used examples and ideas from the readings to support my argument in my essay.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.33</td>
<td>3.45</td>
<td>3.57</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.34</td>
<td>1.28</td>
<td>1.29</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Mode</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
Question 32 addressed whether the writers’ returned to the source texts for ideas while they wrote rather than just reading them in planning. Table 14 shows that across the levels, writers did not report agreement or disagreement with this statement, which suggests that it might have occurred but only occasionally. However, at Level 3 most writers strongly disagreed with this statement. Several explanations could be given for this: they understood the reading more fully after reading one reading or they were less dependent on it for ideas as they wrote.

Table 14. Question 32: “I reread the readings while I was writing to find ideas to put in my essay.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.20</td>
<td>3.00</td>
<td>3.05</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.27</td>
<td>1.35</td>
<td>1.60</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Mode</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Using source texts to organize essays during planning has been found in previous studies of integrated task process (Plakans, 2008). On the questionnaire, writers reported somewhat different uses of this strategy across levels as shown in Table 15. At Level 1 and 2, most writers agreed that they use the readings when organizing, while at Level 3 most strongly disagreed with this statement. The means across groups fell slightly into the disagreement range. These findings might suggest that some writers are using the source texts for more varied purposes than idea generation, however, not to the benefit of their score.

Table 15. Question 31: “I used the readings to help organize my essay.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.95</td>
<td>2.98</td>
<td>2.5</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.33</td>
<td>1.37</td>
<td>1.29</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Mode</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

In contrast to using the readings for organizing, many writers at all levels used the readings to help choose an opinion. This questionnaire item was included in the category of organizing, as it would lead to selecting a topic sentence. However, it may also fall into the “generating ideas” category, which was strong for most writers. Table 16 displays this across writers. Most writers at Level 3 strongly agreed that they used the source texts to develop their opinions. Also, writers at levels 1 and 2 agreed with this statement. Because the readings presented two opinions on the topic, this strategy may be related to the task structure as well as the provision of source texts. One writer mentioned this in response to an open-ended question: “I read first and underline the main ideas that I understood. Then, I built my own opinion about this issue.”
Table 16. Question 23: “The reading helped me choose an opinion on the issue.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.74</td>
<td>3.65</td>
<td>3.52</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.15</td>
<td>1.23</td>
<td>1.33</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mode</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

In summary, writers at all levels reported using the source text to gain ideas and to decide on a position for their argument. The high- and low-level writers reported that they used ideas from the readings for support. The high level marked lower use of rereading the source texts, while writing and using the source text to help them organize.

**Integration Process**

As shown in Table 17, for all levels, there was a moderate level of agreement that they used words from the readings as they wrote. More agreement seemed to come from Level 2, as the mode was 5 and the median 4. In their open-ended responses, writers at levels 1 and 2 mentioned this integration process at the word level with statements like, “I get words from reading” or “I get ideas and some vocab.”

Table 17. Question 14: “I used some words from the readings when I wrote.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.54</td>
<td>3.63</td>
<td>3.3</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.25</td>
<td>1.42</td>
<td>1.22</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>4</td>
<td>3.5</td>
</tr>
<tr>
<td>Mode</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

When asked about paraphrasing from the source texts, shown in Table 18, writers did not report heavy use of this integration style. At all levels the means were on the side of disagreement, with the Level 3 writers slightly but not significantly lower.

Table 18. Question 36: “I paraphrased the reading in my writing.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.97</td>
<td>2.90</td>
<td>2.7</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.21</td>
<td>1.44</td>
<td>1.49</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Mode</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

In terms of direct source use, most writers strongly disagreed that they copied phrases or sentences directly from the source texts. Table 19 shows that this finding occurred at all
levels. This result suggests that writers were not aware of verbatim source use or did not wish to report that they had used this strategy.

Table 19. Question 38: “I copied phrases and sentences directly from the reading into my essay.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.59</td>
<td>2.42</td>
<td>2.38</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.44</td>
<td>1.49</td>
<td>1.56</td>
</tr>
<tr>
<td>Median</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Mode</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

More difference occurred across levels when writers reported on citing authors of the source texts in their essays. The means across levels, displayed in Table 20, were somewhat similar, falling slightly below agreeing that they used this process. However, the mode for Level 3 was strong agreement with the statement regarding use of authors’ names. For this question, there were also overall larger deviations from the mean suggesting a range of responses from the writers.

Table 20. Question 35: “I used the authors’ names in my essays.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.67</td>
<td>2.96</td>
<td>2.85</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.7</td>
<td>1.77</td>
<td>1.93</td>
</tr>
<tr>
<td>Median</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Mode</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

In summary, some differences in integration process occurred across levels. The most striking difference was in citation of the source text authors with Level 3 writers strongly agreeing that they cited. The Level 2 writers reported using the words from the source texts more. All levels were neutral about copying words or phrases from the text.

Knowledge of Integrating Reading-Writing

While not directly addressing the use of source texts in the writing, the items regarding prior knowledge and experience with source text use can inform results regarding source use. These items explored how familiar writers were with the issues around academic reading and writing. Question 39 asked directly about previous learning in classes, showing that all levels reported some agreement with the statement (see Table 21). Level 1 and Level 2 had similar neutral responses to this item, while most writers in Level 3 reported that they strongly agreed with this statement. Their exposure to using reading with writing might have aided them in the task and led to a higher score.
Table 21. Question 39: “I have learned how to use reading with writing in a class.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.36</td>
<td>3.49</td>
<td>3.52</td>
</tr>
<tr>
<td>SD</td>
<td>1.18</td>
<td>1.21</td>
<td>1.36</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Mode</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Question 40 delved more specifically into integration practices by referring to plagiarism. As Table 22 shows, the responses from Level 3 writers were strongly in agreement with this statement, which could explain the results in the written products and their agreement with the statement about citation. Level 1 and 2 writers also slightly agreed but fell mostly in the middle.

Table 22. Question 40: “I have learned about plagiarism.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.10</td>
<td>3.37</td>
<td>4.0</td>
</tr>
<tr>
<td>SD</td>
<td>1.25</td>
<td>1.38</td>
<td>1.30</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Mode</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Writers were also asked about their comfort level with using others’ ideas in their writing. As Table 23 shows, all levels were similar in response to this item, with no strong agreement or disagreement. More writers at Level 2 reported disagreement about comfort than the other levels.

Table 23. Question 41: “I feel comfortable using other people’s ideas in my writing.”

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.20</td>
<td>3.00</td>
<td>3.10</td>
</tr>
<tr>
<td>SD</td>
<td>1.11</td>
<td>1.34</td>
<td>1.26</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Mode</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Task Difficulties

In analyzing the writers’ responses about task difficulties, some differences appeared across levels. Both the Level 1 and Level 2 writers frequently mentioned the vocabulary in the reading being too challenging. Level 3 writers did not remark on this, but instead had a different vocabulary-related issue. They felt that they did not have the scientific terminology to write on the topic. For example, one writer said, “I did not prepare for writing scientific essay. The time was not enough. I forgot some scientific words and how to spell them.” Level 2 writers also responded that spelling was difficult for them, while Level 1 and Level 3 writers did not mention it at all.
Level 1 and Level 2 writers felt that the task was difficult because they did not have enough information about the topic, did not understand the topic, and did not have many ideas to write about it. They also stated that organizing the essay was difficult for them. The Level 3 writers did not mention these difficulties, but instead commented on the time limit and the low interest level of the topic.

Integration Style

Writers’ comments on integrating the readings are included with the summary of the quantitative data; however, some comparison from the qualitative analysis also revealed interesting patterns. The Level 1 writers reported mainly using the source texts for ideas and to “get words.” Level 2 also mentioned gaining ideas and information from the reading, but also commented on using the source text to support their opinion. This use was also found in Level 3 comments. For both Level 2 and Level 3, a small number of writers noted that they did not use the sources in their writing. Level 3 writers answered the question somewhat more specifically by writing that they summarized the texts or used quotations. For example, one writer said, “I summarize the ideas in my own words and also I quote.”

Summary of Key Findings

When writers reported on source use in their composing processes, overall the differences in proficiency levels were not huge. All writers reported using the source texts to generate ideas and opinions. The higher level writers seemed less dependent on the source texts; they reported rereading them less and reported lower rates of using the source texts for organizing their essays. The lower levels reported more difficulty with the vocabulary in the readings. All levels showed writers were neutral about copying directly from the source texts; however, the higher level writers reported more learning and knowledge about integrated reading-writing. These process results show that the writers did not report different processes when asked direct questions about it; however, they revealed some patterns that suggests higher scoring writers had more knowledge and sophistication with integrated reading-writing as well as possibly better comprehension of the readings.

Discussion and Implications

In this study, a number of discourse features were analyzed to investigate how these features behave across different proficiency levels. In addition, the study attempted to examine if certain features could play a greater role in characterizing a certain level. Furthermore, the questionnaire data were used to check whether L2 writers at different proficiency levels use similar or different strategies while working on reading-to-write tasks. The results showed statistically significant overall ANOVA across the three proficiency levels in fluency, grammatical accuracy, indirect source use, and total source use. These results exhibited effect sizes that ranged from medium to large values. However, the analysis of some other features, such as lexical sophistication, syntactic complexity, direct source use, and direct source use without quotations, did not yield significant differences. With regard to overall source use, results showed that students with higher proficiency generally exhibited more source use than those with low proficiency.

Results of the fluency variable analysis demonstrated that L2 writers produce longer essays when proficiency increases. Fluency was the only variable that exhibited significant
differences across the three groups. In addition, it produced the largest effect size among all
the discourse features analyzed in this study. In general, L2 writing studies (e.g., Cumming et
al., 2005; Grant & Ginther, 2000) yielded similar results. Grammatical accuracy also was
another variable that produced significant results. However, pairwise comparisons showed
that there were no significant differences between Level 2 and Level 3. This result may
suggest that grammatical accuracy is a main factor when scoring essays at lower levels. Once
proficiency increases, grammatical accuracy does not take precedence and other discourse
features are at play.

Analysis of syntactic complexity did not yield any significant results. The absence of
significant differences across the three proficiency levels in syntactic complexity is in
agreement with the results of Cummings et al. (2005) who found significant results in the
mean number of words per T-units at different proficiency levels. But, Cumming et al. (2005)
found no significant differences in the amount of subordination measured by the mean
number of clauses per T-unit. The current research used a coordination measure, the mean
number of T-units per sentence (TU/S), not a subordination indicator. However, the current
results are in agreement with a number of studies that showed no significant differences in
syntactic complexity. For example, Hillock (1986) concluded that there is no consistent
association between holistic ratings of writing and syntactic complexity of a written text. In
addition, Ortega (2003) argued that studies establishing proficiency groups based on holistic
ratings usually yield homogenous results across compared levels, which leads to a prevalence
of statistically nonsignificant differences. She adds that it is not appropriate to equate
syntactically complex writing with good writing. Another issue that may interpret the absence
of nonsignificant differences in syntactic complexity is cross-rhetorical transfer. The students
participating in this study speak Arabic as a native language. Consequently, it might be the
case that they transferred a number of rhetorical strategies from Arabic. Cross-rhetorical
influences in L2 writers with Arabic background would be an area of research that needs to be
investigated in future studies.

With regard to lexical sophistication, results yielded no significant differences across
the three proficiency levels. This result is in disagreement with the study of Cumming et al.
(2005) who found significant differences in average word length across different proficiency
levels. However, results of Cumming et al. (2005) yielded a low effect size ($\eta^2 = .03$). One
possible reason for this result may be due to lexical borrowing from the source readings. It is
expected that students would integrate words and phrases appearing in the reading sources in
their writing. Therefore, this practice might have improved vocabulary sophistication across
the three proficiency profiles. This conclusion is in agreement with East (2006) who found
improvement in lexical sophistication when students used dictionaries. In addition, average
word length does not take into account lexical accuracy, a linguistic feature that affects
writing quality. Engber (1995: 194) found that ‘lexical variation without error correlated more
highly with test scores than lexical variation with error.’

The source-use analysis yielded a number of interesting results. For example, students
at higher proficiency levels showed more overall source use compared to low-proficiency
students. This result could be attributed to the complex and demanding nature of source
integration. Cumming et al. (2005, p. 34) argue that effective reading-based writing requires
‘appropriate and meaningful uses of and orientations to source evidence, both conceptually
(in terms of apprehending, synthesizing, and presenting source ideas) and textually (in terms
of stylistic conventions for presenting, citing, and acknowledging sources).’ Their study
found the essays at lower score levels had less verbatim source use with integrated writing tasks. They concluded that lower proficiency writers struggled to comprehend readings in the integrated tasks, which led to their using them less. In the questionnaire, this level mentioned using the readings for ideas, but it also commonly cited difficult vocabulary as a challenge in completing the task, which suggests that reading comprehension may be related to lower source. So, probably there is a proficiency threshold that students should pass before successfully integrating reading sources in their writing. This conclusion was reiterated by a number of low-proficiency students in their responses to the open-ended questionnaire questions. Those students had trouble with numerous vocabulary items and structures in the two passages. Others referred to the overall difficulty of the two reading texts.

When asked about their awareness of plagiarism and source integration issues, the majority of students across the three proficiency levels confirmed their attentiveness to these matters. However, the results demonstrated that those students often used inappropriate textual borrowing. One possible reason behind this behavior could be attributed to the students’ lack of declarative knowledge about plagiarism, although they expressed the opposite in their questionnaire responses. This behavior is expected in the questionnaire methodology when dealing with a thorny issue, such as plagiarism, for social desirability reasons. In addition, the participants were undergraduate students with little experience in academic writing, so they may lack knowledge of textual borrowing strategies. This conclusion is supported by the results of a study conducted by Campbell (1990) who found frequent copying practices among both native and nonnative undergraduate writers. A more politically correct interpretation for this contradiction could be linked to patchwriting (Pecorari, 2003; Howard, 2001; Ange’ lil-Carter, 2000), which refers to the unintentional copying from source. For example, Howard (2001: 1) argues that these practices are caused by:

uneven reading comprehension: the student doesn’t fully understand what she is reading and thus can’t frame alternative ways for talking about its ideas. Or the student understands what she is reading but is new to the discourse. She merges her voice with that of the source to create a pastiche over which she exercises a new-found control.

In this quotation, Howard suggests that the declarative knowledge should not necessarily be the cause of inappropriate textual borrowing, but rather procedural knowledge. This was evident in the current study because students were instructed not to use patchwriting. However, this problem was common in many of the essays across the different proficiency levels.

It is important that inappropriate textual borrowing be interpreted within a culturally situated framework. The educational context in which L2 writers operate is critical in understanding why they use sources improperly. A number of studies (e.g., Pennycook, 1996; Matalene, 1985) referred to these cultural aspects of plagiarism. For example, Pennycook (1996), who worked with Chinese students, concluded that many of those students had a different conceptualization of plagiarism from the Western paradigm. Some of them did not think of unattributed textual borrowing as plagiarism. Another study carried out by Kirkland and Saunders (1991) found that students in countries or cultures that embrace rote learning and memorization tend to use inappropriate textual borrowing. Based on the researchers’ experience, teaching L2 writing at the undergraduate level in the UAE and the Middle East in
general, writing instruction is still heavily focused on independent writing tasks. University students usually work on writing activities that target development of ideas, organization, mechanics, and more importantly grammar. There are a number of reasons behind this trend. First, undergraduate students at the UAE spend the first two years in general English classes where priority is given to basic writing skills. In addition, students are required to take the IELTS test before starting courses in their majors. The IELTS test mainly focuses on independent writing tasks, and that is why writing instructors work hard to prepare students for this task type. Given this context, few activities in these classes, if any, provide students with opportunities to work on reading-to-write tasks. As a result, students are not taught how to integrate reading sources and consequently are not aware of strategies used to integrate sources in their writing. With the current changes in writing assessment, it is hoped that beneficial washback would be achieved in L2 writing classes in the UAE context with more time assigned to source-based writing, and, more importantly, to how to integrate these sources in students’ writing.

These results hold a number of implications for language assessment in terms of construct definition, rating scales, and task development. A construct for writing or for integrated reading-writing has not be decided on in the field (Grabe & Kaplan, 1996; Hirvela, 2004). By investigating discourse features, source use, and process in writing for an integrated task, this study provided some insight into what might be included in an integrated reading-writing construct or, at the very least, what should be considered when interpreting scores from integrated tasks. For the lower level writers, different factors may be at work in scores than for the higher level writers because discourse features such as fluency and grammatical accuracy were more defining and limiting issues for their writing. The Level 1 writers may have been hindered more by general L2 proficiency or L2 reading ability, which impacts the writing features as well as their use of source texts in writing and their reading of source texts. Higher level writers may have been more facile with these features and have had more exposure to integrated writing or issues in academic writing such as plagiarism. These issues suggest that a construct for integrated reading-writing includes ability to control and create language elements of the written product, knowledge and use of integrating sources text, as well as general L2 proficiency and L2 reading ability. The results might suggest that there is a developmental difference in the weighting of the issues for the process and scoring of reading-writing tasks, with more attention on proficiency and discourse features at lower levels and more application of knowledge on reading-writing and source use at higher levels. These implications are speculative and will need more study with different populations and different tasks.

Because construct should shape a rating scale for a writing test, the issues mentioned in the above paragraph might need consideration in scale development. The scale used in this study included issues of source use; however, the scoring was holistic, which causes difficulty in identifying how much source use impacted the final rating. The use of sources, as well as reading proficiency, may be features that raters need to consider in rating integrated tasks. In addition, the finding that discourse features and source use did not differentiate the upper two levels indicates that other features in the rating scale were distinguishing them, particularly organization, coherence, and content. For the lower levels, the writers may not have written enough to create cohesive, organized arguments. Other explanations may be that the raters focused more on language errors and use of source texts at the lower levels because more errors occurred, or the language errors may have prevented the raters from comprehending
aspects of the essay such as organization and coherence. Certainly, these issues reveal the challenges of rating integrated tasks, and decidedly recommend further study of what raters attend to when rating such tasks, as well as the rubrics used in rating. The rating of integrated tasks is an area needing more study.

Because this study only focused on one task, implications for developing integrated tasks are somewhat limited. However, the writers’ questionnaire responses regarding their process suggest that several considerations need to be made by developers. Specifically, the selection of reading texts is no small matter. Writers need to be able to comprehend sources in order to integrate them. Many writers used the source texts for information and generating ideas, which indicates that for fairness across writers, readings should be accessible. If challenging texts are used, then scores need to include reading proficiency in the interpretation. More research is needed about the reading issue and readability levels of texts, as well as the role of reading for different kinds of integrated tasks; for example, summary writing, which might hold even stronger ties to reading ability.

Limitations

Several limitations in this study need to be acknowledged. First of all, the study used one task in order to control task effect. However, this decision creates issues about the topic and task. In the process questionnaire, writers indicated that they had different levels of interest, knowledge, and motivation for writing on this topic, which may have impacted their writing. The use of an argumentative essay is also a limitation, as many different kinds of integrated tasks occur in writing tests and in academic classrooms. Another decision was to focus on a homogenous group of writers to provide clearer results and implications. However, this narrow scope limits implications from the study. Caution should be used in applying the findings to other culture groups or different first-language writers. Lastly, the use of a process questionnaire may have limited the information on writers’ source use. Questionnaire and survey data relying on self-reporting and, in this study, self-reflection on processes may have been hard for writers to access after writing, and thus hindered their ability to accurately or confidently report on their source use when they wrote.

Conclusion

Integrated tasks provide an interesting, innovative, and authentic means to assess academic writing. Writers and test developers see them as a potential solution to a number of dilemmas with impromptu independent tasks. This study investigated several aspects of the products and processes from writers completing integrated tasks in order to learn more about them. The findings show interesting differences between low and high proficiency writers, as well as revealing the integral relationship of the reading texts and writing in such tasks. More questions arose in this study and will require attention for the field to continue using integrated tasks and for test users to understand the resulting scores.
Acknowledgements

This study was made possible by a Spaan Fellowship for Studies in Second or Foreign Language Assessment at the University of Michigan. We would also like to thank the expert scholars who provided input in task and questionnaire development, as well as our raters and participants. The United Arab Emirates University was helpful and receptive in allowing our study to be conducted at their institution.

References


Appendix A

Writing task

Read the question below, then, read the two passages to get more information about the topic. Write an essay on the topic giving your opinion in response to the question. Typically, an effective response will contain a minimum of 300 words. Your writing will be scored based on how well:

- your ideas are explained
- the readings support your argument
- you organize your essay
- you choose words
- you use grammar
- you spell and punctuate

Some people believe that global warming is damaging our planet. Others believe that global warming is not a serious problem. Which point of view do you agree with? Why?

Give reasons and support your writing with examples.

IMPORTANT! PLEASE READ CAREFULLY BEFORE WORKING ON THE TASK:
- The two passages should help you get some ideas about the topic.
- You may go back to the passages to check information while writing.
- You can borrow ideas and examples from the text. However, you should mention the author’s name if you do so.
- Also, if you take exactly the same phrases or sentences mentioned in the passage, put them between two inverted commas (“ ”).
**Reading (1): Scientists Say Global Warming is Undeniable (Adapted from an article by David Smith, Reuters, 2005)**

Scientists have confirmed that climate change is being caused by human activity. A number of studies looking at the oceans and melting ice leave no doubt that it is getting warmer, people are to blame, and the weather is going to suffer.

Tim Barnett, who is a famous global warming researcher, indicates that new computer models that look at ocean temperatures instead of the atmosphere show the clearest signal yet that global warming is well under way. Mr. Barnett said that earlier climate models based on air temperatures were weak because most of the evidence for global warming is not in the air.

Other researchers found clear effects on climate and animals. For example, Ruth Curry, who is from an important oceanographic institute, said changes in the water cycle affects the ocean and, ultimately, climate. She said the changes were already causing droughts in the United States, and Greenland’s ice cap. Sharon Smith of the University of Miami found melting ice was taking with it plants that are an important base of the food supply for many animals. And the disappearing ice meant big Institution found that melting ice was changing the animals such as polar bears and seals were losing their homes.

Given all these serious problems caused by global warming and the way humans have abused the earth, governments must act immediately to save our planet. The future of this planet depends on our actions and any delay would result in serious problems.

**Reading (2): Myths of Global Warming (Adapted from an article by Sterling Burnett, a Senior Fellow for the National Center for Policy Analysis, 2001)**

There is no scientific agreement that global warming is a problem or that humans are its cause. Even if current predictions of global warming are correct, much of the environmental policy now proposed is based on wrong theories.

First, there is a wrong belief that the earth is warming. While ground-level temperature suggests the earth has warmed between 0.3 and 0.6 degrees since 1850, reliable global satellite data show no evidence of warming during the past 18 years. In addition, scientists do not agree that humans affect global climate because the evidence supporting that theory is weak.

Some people also think that the government must act now to stop global warming. However, a 1995 analysis by supporters of global warming theory concluded that the world’s governments can wait up to 25 years to take action with no bad effect on the environment. In short, our policymakers need not act immediately. The government has time to gather more data, and industry has time to develop new ways of reducing its influence.

Supporters of the theory of human-caused global warming also argue that it is causing and will continue to cause all environmental problems. Many famous scientists reject these beliefs. Sea levels are rising around the globe, though not equally. In fact, sea levels have risen more than 300 feet over the last 18,000 years. Contrary to the predictions of global warming theorists, the current rate of increase is slower than the average rate over the 18,000-year period.
Appendix B

ID Number _____________________

Reading-to-write process questionnaire

General Instructions
The writing assignment you just completed is called a reading-to-write task, and we are interested in knowing what you thought about it. This is not a test, so there are no ‘right’ answers, just think back and remember the best you can. Thank you very much for your help.

Opinions about the writing task
Read the statements below and circle your feelings on the scale from 1 to 5. “1” = strongly disagree and “5” strongly agree

I. Writing process

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly disagree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I understood the instructions for the writing task.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>2. I found global warming interesting to write about.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>3. It was clear to me how to complete the task.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>4. I have strong opinions on global warming.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>5. I have written papers on global warming before today</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>6. I made an outline before I started writing.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>7. I wrote a list of my ideas before I started writing.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>8. I did all of my planning before writing.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>9. I planned before and during writing.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>10. I wrote a good essay.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>11. The essay I wrote shows my ability as a writer.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>12. The readings helped me to write better.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>13. I liked reading about the topic before writing.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
<tr>
<td>14. I used some words from the readings when I wrote.</td>
<td>1  2  3  4  5</td>
<td></td>
</tr>
</tbody>
</table>
15. I used some of the ideas from the readings in my essay. 1 2 3 4 5

16. I have worked on a reading and writing task like this one before. 1 2 3 4 5

17. This seemed like a writing assignment that I would have in an university class. 1 2 3 4 5

18. I think my score will be high on this task. 1 2 3 4 5

19. I revised my writing before I finished. 1 2 3 4 5

20. It was easy to support my opinion on this topic. 1 2 3 4 5

21. I didn’t have any trouble in writing this essay. 1 2 3 4 5

II. Reading in the writing task

22. I looked back at the readings often while I was writing. 1 2 3 4 5

23. The reading helped me choose an opinion on the issue. 1 2 3 4 5

24. The readings were interesting to me. 1 2 3 4 5

25. I could understand most of the **words** in the readings. 1 2 3 4 5

26. I could understand the **ideas** in the readings. 1 2 3 4 5

27. I have read articles on global warming in **English** before today. 1 2 3 4 5

28. I have read articles on global warming in **Arabic** before today. 1 2 3 4 5

29. I used examples and ideas from the readings to support my argument in my essay. 1 2 3 4 5

30. I used the readings to help me get ideas on the topic. 1 2 3 4 5

31. I used the readings to help organize my essay. 1 2 3 4 5

32. I reread the readings while I was writing to find ideas to put in my essay. 1 2 3 4 5
33. I could have written this essay **without** the readings. 1 2 3 4 5
34. I reread the readings while writing because I forgot what they were about. 1 2 3 4 5
35. I used the authors’ names in my essays. 1 2 3 4 5
36. I paraphrased the reading in my writing. 1 2 3 4 5
37. I used **only** my own ideas in my writing, nothing from the reading. 1 2 3 4 5
38. I copied phrases and sentences directly from the reading into my essay. 1 2 3 4 5
39. I have learned how to use reading with my writing in a class. 1 2 3 4 5
40. I have learned about plagiarism. 1 2 3 4 5
41. I feel comfortable using other people’s ideas in my writing. 1 2 3 4 5

**IV. Comments**
Please write responses to the following questions.
42. Did you have enough time to complete the task? If not, how much more time would you have liked?
43. Explain any difficulties that you had with reading or writing in this task.
44. If you used the readings for your writing, please describe how you used them.
45. Would you prefer writing with reading on a test to a writing test without reading? Why or why not?
46. What would you tell a teacher who is thinking about using reading-to-write tasks in his/her classes?
47. Is there anything you would like to comment on about this task?

**II Background information**
48. How many years have you been at the university?
49. What is your major?
50. Have you had previous writing courses?
If yes, please list them.

51. What is your level of experience with academic writing (circle best answer)
   No experience  a little  some  a lot  very experienced

52. What is your level of English language skill (circle best answer)
   low  so-so  medium  high  very high

53. If you have taken IELTS, what was your score?

54. If you have taken IELTS, when did you take it last? (circle best answer)
   Less than a year  1 year  2 years  3 years  4 years  More
### Appendix C

**Reading-to-write scoring rubric**

(Adapted from the scoring rubric of the TOEFL iBT integrated task)

<table>
<thead>
<tr>
<th>Score</th>
<th>Task Description</th>
</tr>
</thead>
</table>
| 5     | **A response at this level:**  
- successfully presents their ideas in relation to the relevant information presented in the reading sources.  
- is well organized with well-developed content  
- occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections. |
| 4     | **A response at this level:**  
- is generally good in coherently and accurately presenting their ideas in relation to the relevant information in the reading texts, although may have inaccuracy, vagueness, or imprecision in connection to points made in the readings.  
- has clear organization and logical development.  
- more frequent or noticeable minor language errors; such errors do not result in anything more than an occasional lapse of clarity or in the connection of ideas. |
| 3     | **A response at this level**  
- conveys some relevant connection to the reading, but only vague, global, unclear, or somewhat imprecise connection to points made in the reading.  
- development is somewhat limited, but some specific support for their argument is provided.  
- occasionally lacks cohesion but has a basic organizational structure.  
- includes errors of usage and/or grammar that are more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections. |
| 2     | **A response at this level**  
- contains some relevant information from the readings, but is marked by significant language difficulties or by significant omission or inaccuracy of important ideas from the readings  
- lacks logical organizational coherence and development. Ideas are very general and lack specific details in support.  
- contains language errors or expressions that largely obscure connections or meaning at key junctures, or that would likely obscure understanding of key ideas for a reader not already familiar with the topic. |
| 1     | **A response at this level**  
- provides little or no meaningful or relevant coherent content from the readings and does not follow an organization pattern or develop content.  
- includes language that is so low and it is difficult to derive meaning. |
<table>
<thead>
<tr>
<th>0</th>
<th><strong>A response at this level</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• either merely copies sentences from the reading, rejects the topic, not connected to the topic, is written in a foreign language, or is blank.</td>
<td></td>
</tr>
</tbody>
</table>

Notes:
-- If only one phrase or sentence is copied from the reading, do not assign a “0” but base the rating on the rest of the essay.
-- If language use & development are at a certain level but the readings have not been included (directly or indirectly), add “NS” to the score
-- If ideas from the readings are used but not cited, do not rate lower unless the writer is copying directly (use NC marking).
Investigating Different Item Response Models in Equating the Examination for the Certificate of Proficiency in English (ECPE)

Tian Song
Michigan State University

ABSTRACT When item response models are applied in equating, the assumption of local independence is required. Polytomous item response theory (IRT) models can be considered as alternatives to dichotomous models if the assumption is violated. This study compares the performance of the dichotomous IRT model and a combination of dichotomous and polytomous IRT models in equating two forms of the Examination for the Certificate of Proficiency in English (ECPE). Traditional equating methods are used as a baseline for comparison. The results reveal that a combination of the three-parameter logistic model and the generalized partial credit model yield results similar to the traditional equating functions for the listening section, and the three-parameter logistic model performs better in the GCVR section.

In high-stakes testing programs, there is a concern that different forms might differ in difficulty, and scores on the forms are not comparable. Equating is “a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably” (Kolen & Brennan, 2004, p. 2). Therefore, test equating is a requirement for fairness to examinees.

In the literature, a wide array of equating procedures have been developed, which can be categorized into two major methods, namely traditional methods (mean, linear, and equipercentile equating), and item response theory (IRT) methods (true-score and observed-score equating). When item response models are applied in equating, the assumption of local independence is required. The local independence assumption states that after taking into account examinee ability, examinee responses to the items are statistically independent. In other words, for a given examinee, the responses to different items are not related. For example, one item does not provide clues to the correct answer to another item. However, for those items based on a common stimulus, such as reading passages or charts, local independence likely would not hold (Yen, 1993; Wainer & Thissen, 1996). In this situation, the use of dichotomous IRT models to equate tests might cause a problem. To address this problem, items associated with a common stimulus could be scored as a testlet, with summed scores of the items producing a total score for that testlet. The testlet could then be treated as a single polytomous item.
A previous study by Lee, Kolen, Frisbie, and Ankenmann (2001) has demonstrated that for tests composed of testlets only, equating based on polytomous IRT models produces results that more closely agree with the results of traditional methods than they do with dichotomous models, where the violation of the local independence assumption is severe. The present study is closely related to Lee et al.’s study, but differs in an important aspect, in that I extend the comparison of dichotomous and polytomous item response models in equating tests composed of testlets only to mixed-format tests. This study provides new evidence on the performance of different IRT models in equating tests. A mixed-format test is a test containing a mixture of different item formats (e.g., a mixture of multiple-choice and constructed response items), and is more widely used in classroom and large-scale assessments. As Baker and Kim (2004) indicated, there are many combinations of dichotomous and polytomous models that can be used to analyze data from mixed-format tests, such as a combination of the three-parameter logistic (3PL) model and graded response (GR) model, and a combination of the 3PL and generalized partial credit (GPC) model.

The primary purpose of this study is to compare equating results based on dichotomous and a combination of dichotomous and polytomous IRT models. Because traditional equating methods, such as mean, linear, and equipercentile equating, use total test scores and are not affected by the violation of local independence, they are considered as baselines for comparison. To be specific, the research question addressed is: In equating mixed-format tests, which IRT model produces the results that more closely agree with the results of traditional methods?

Competing IRT Models

Under item response theory, the interactions of a person with test items can be adequately represented by a probabilistic expression. That is, the probability of correct response to a given item is a function of both the characteristics of person and items. Over the last few decades, the use of IRT models, such as three-parameter logistic (3PL) model (Birnbaum, 1968), the generalized partial credit (GPC) model (Muraki, 1992), the graded response model (Samejima, 1969), and the nominal response model (Bock, 1972), has grown considerably in practical testing programs. In this study, the 3PL model is used for dichotomously scored items and the GPC model is for polytomously scored items.

Three-Parameter Logistic (3PL) Model

The 3PL model is the most general model for scoring dichotomous items. Under the 3PL model, the probability of examinee $i$ giving a correct response for item $j$ is

$$P(U_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1-c_i) \frac{e^{B_{ai}(\theta_j - b_i)}}{1 + e^{B_{ai}(\theta_j - b_i)}}$$

where $U_{ij}$ represents the person $i$’s score on the test item $j$,

- $a_i$ is the discrimination parameter,
- $b_i$ is the difficulty parameter,
- $c_i$ is the pseudo-guessing parameter,
\( \theta_j \) is the examinee’s ability, and
D is the scaling constant (typically 1.7).

**Generalized Partial Credit (GPC) Model**

For those items based on a common passage, scores are summed for the items to produce a total score for that passage, and then the polytomous IRT model is applied to it. For example, a five dichotomous item reading passage could be treated as a polytomous item ranging from 0 to 5. In this study, the generalized partial credit model is used. The mathematical expression for the generalized partial credit model is given below

\[
P(u_{ij} = k | \theta_j) = \frac{e^{\sum_{u=1}^{k} d_{au}(\theta_j - b_j + d_{u})}}{\sum_{v=1}^{m_i} e^{\sum_{u=1}^{k} d_{av}(\theta_j - b_j + d_{u})}},
\]

where \( k \) is the score on the item,
\( m_i \) is the total number of score categories for the item,
\( d_{au} \) is the threshold parameter for the threshold between scores \( u \) and \( u-1 \),
\( a_i \) is the overall discrimination of the item, and
\( b_i \) is the overall difficulty of the item.

Here \( a_i \) is assumed to be the same at all thresholds, but may differ across items. The threshold parameter \( d_{il} \) is defined as 0, and \( d_{ik} \) indicates where the probability of responses changes from being greater for score category \( k - 1 \) to being greater for score category \( k \). Usually the sum of the \( d_{ik} \) parameters is constrained to 0 for estimation purposes (Reckase, personal communication, February 2008).

**EQUATING METHODS**

**IRT Equating**

Equating with item response theory is simply to put item parameter and ability estimates from two forms on a common scale. There are two major ways to develop a common scale. It can be constructed by simultaneous estimation of item parameters on a combined dataset from two forms (concurrent calibration), or an alternative way is to estimate item parameters for the two forms using two separate runs of the software, then apply linking methods (mean/mean, mean/sigma, Haebra, or Stocking and Lord method) to put them on a common scale. However, Kolen, and Brennan (2004) point out that reporting IRT ability estimates has a few disadvantages in practical testing programs. First, it is difficult to explain to examinees why the same number-correct score may receive different ability estimates. Second, examinees located at the lower and upper end of the distribution often have relatively greater amount of measurement errors. Therefore, it is better to convert estimated IRT abilities to number-correct (NC) scores and develop a relationship between NC scores on two forms.
True and observed score equating are the two methods currently available for conducting IRT equating. In true score equating, for a given ability value of theta, the number-correct true scores associated with this theta on two forms are considered to be equivalent. In IRT, the number-correct true score that is equivalent to \( \theta_j \) is defined as

\[
\tau(\theta_j) = \sum_{l} p_{lj}(\theta_j; a, b, c)
\]

where \( p_{lj} \) is the probability of examinee \( i \) giving a correct answer for item \( j \), and summation \( j \) is over items. This equation is also referred to as the test characteristics curve, which relates IRT ability to number-correct true score. Basically, there are three main steps in true scoring equating:

1. Specify a true score in one form.
2. Find the \( \theta \) that corresponds to that true score.
3. Find the true score on another form that corresponds to that same \( \theta \). The true score in the last step will be considered as the equated score.

Observed score equating is conducted by estimating the frequency distributions of number-correct observed scores using item parameter and \( \theta \) estimates for each form, and then using conventional equipercentile equating method to approximately equate these estimated observed scores. For example, in dichotomous IRT observed score equating:

1. For one form, use the compound binomial distribution (Lord and Wingersky, 1984) to generate the distribution of observed number-correct scores for examinees with a given \( \theta \), which is denoted as \( f(x|\theta) \).
2. Accumulate the observed-score distribution for examinees at each \( \theta \), and get the observed-score distribution for examinees of various abilities using

\[
f(x) = \int_{\theta} f(x|\theta) \varphi(\theta)d\theta
\]

with \( \varphi(\theta) \) is the distribution of \( \theta \)
3. Follow a similar procedure and get the observed-score distribution for the other form.
4. Apply equipercentile methods to equate scores from two forms.

For polytomous IRT models, the procedure is similar except that a generalization of compound binomial distribution—compound multinomial distribution—is used to model the observed number correct score distribution (Thissen, Pommerich, Billeaud, & Williams, 1995).
Traditional Equating

Two traditional equating methods used in this study are linear and equipercentile equating. In linear equating, the equating function is developed by setting the standardized deviation scores on the two forms to be equal. The equated scores deriving from this method have the same mean and standard deviation as the original scores. Equipercentile equating is to identify scores on one form that have the same percentile ranks as scores on another form. Both methods use total test scores, and are not affected by IRT assumptions. Therefore it is reasonable to consider them as baseline methods for comparing the performance of IRT models in equating.

Method

Data

The data are from the Examination for the Certificate of Proficiency in English (ECPE), which is an English as a second or foreign language test battery designed for individuals who have advanced-level language proficiency (English Language Institute, 2006, 2008). The ECPE is developed by the English Language Institute at the University of Michigan (ELI-UM), and is administered annually at approximately 125 authorized test centers in approximately 20 countries. There are four sections in each test: speaking, writing, listening, and grammar/cloze/vocabulary/reading (GCVR). The four sections are individually scored and examinees are awarded a Certificate of Proficiency based on their aggregated scores of these four sections.

Only the listening and GCVR sections are investigated in this study. All test items in these two sections are multiple-choice items. The listening section has 50 items: the first 35 items are individual items, each based on one short conversation or question; the last 15 items are based on three long dialogues, each dialogue having 5 questions. The GCVR section has 100 scored items: 30 individual items for grammar, 30 individual items for vocabulary, 20 cloze items sharing one passage, and 4 reading passages with 5 questions for each. For those items based on common reading passages or dialogues, the local independence assumption likely would be violated, and polytomous IRT models might be considered as an alternative to dichotomous models.

The two ECPE forms from year 2004–05 and 2006–07 are equated. Across the two forms there are 10 common items in the listening section, and 20 common items in the GCVR section: 10 grammar and 10 vocabulary items. Table 1 displays the descriptive statistics of raw scores on these two forms.

Table 1. Descriptive Statistics for ECPE Forms

<table>
<thead>
<tr>
<th>Test</th>
<th>Sample Size</th>
<th>No. Items</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form 2004–05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening</td>
<td>33027</td>
<td>50</td>
<td>38.76</td>
<td>6.06</td>
<td>-0.643</td>
<td>3.209</td>
</tr>
<tr>
<td>GCVR</td>
<td>33027</td>
<td>100</td>
<td>68.03</td>
<td>11.48</td>
<td>-0.208</td>
<td>3.010</td>
</tr>
<tr>
<td>Form 2006–07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening</td>
<td>35074</td>
<td>50</td>
<td>34.03</td>
<td>6.80</td>
<td>-0.348</td>
<td>2.794</td>
</tr>
<tr>
<td>GCVR</td>
<td>35074</td>
<td>100</td>
<td>66.79</td>
<td>11.66</td>
<td>-0.164</td>
<td>2.980</td>
</tr>
</tbody>
</table>
Analysis

Calibration. There are two choices of IRT models to analyze the dataset. If each item is considered as a unit of analysis and the test as one composed of dichotomous items only, the 3PL model is used and item parameters are estimated using BILOG-MG (Zimowski et al., 2003) with default options. If those listening, cloze, and reading items sharing a common stimulus are scored as blocks, they are treated as polytomous items for analysis. To be specific, three polytomous items were created for the listening section, four polytomous items for the cloze,\(^1\) and four for the reading passages in the GCVR section. In this case, the test was treated as a mixed-format test composed of both dichotomous and polytomous items, and a combination of the three-parameter logistic model and the generalized partial credit model is used. Item parameters are estimated using PARSCALE (Muraki & Bock, 1991).

IRT Equating. After item parameters and \(\theta\) estimates had been obtained for each form, equating was performed and the scores on Form 2006–07 were transformed to those on Form 2004–05. In this study, a random-group equating design is considered. Two groups taking the tests in 2004 and 2006 are assumed to be equivalent. It is not necessary to place item parameters of the two forms on a common scale for IRT equating. This step could be skipped, but performing it will reduce estimation errors (Hanson & Beguin, 2002). Therefore, using common items’ parameter estimates from two forms, a linear transformation is estimated by Stocking and Lord’s (1983) characteristic curve method. STUIRT (Kim & Kolen, 2004) is applied, and it handles both the dichotomous scored and the mixed-format test. After the item parameters and ability estimates are rescaled to a common metric, IRT true score and observed equating are conducted using the computer program POLYEQUATE (Kolen & Cui, 2004) for the mixed-format test and the computer program PIE (Hanson, Zeng & Cui, 2004) for the dichotomous scored test.

Traditional Equating. For comparison purposes, linear and equipercentile equating are conducted for each section using the computer program RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui & Chien, 2004).

Evaluation Criteria. The descriptive statistics of equated score distributions (mean, standard deviation, skewness and kurtosis) for each equating method are calculated and compared. Following Lee, Kolen, Frisbie, and Ankenmann (2001), the overall level of discrepancy between each IRT equating method and traditional equating methods can be evaluated by unweighted root mean square (URMS) and weighted root mean square (WRMS). The smaller these two indices are, the more closely the IRT equating results agree with traditional equating results. The formulas for URMS and WRMS are

\[
\text{URMS} = \sqrt{\frac{1}{n} \sum (a_i - b_i)^2},
\]

where

- \(a_i\) is equated score from IRT equating,
- \(b_i\) is equated score from linear or equipercentile equating,
- \(n\) is number of items, and
- \(i\) represents each number correct score.

\(^1\) 20 cloze items were collapsed into one polytomous item, but the parameter estimation did not converge. Therefore, 4 6-category (0 to 5) polytomous items were used here.
Investigating Different Item Response Models in Equating the Examination for the Certificate of Proficiency in English (ECPE)

\[ WRMS = \sqrt{\frac{1}{\sum f_i} \sum f_i (a_i - b_i)^2}, \]

where \( f_i \) is the frequency distribution of the number-correct score for equated form. This index takes into account the frequency distribution of equated scores. Therefore, the scores that a large proportion of examinees receive would have greater effect on this index, and the scores that no, or a small proportion of, examinees receive would have little or no effect.

**Results**

**Descriptive Statistics of Equated Scores**

In this study, the scores on Form 2006–07 are transformed to the Form 2004–05 scale. Table 2 presents the moments of equated scores for each method and the absolute value of the difference between the equated score moments and the Form 2004–05 moments (|DIFF|). For the listening section, the equating method using a combination of 3PL and GPC models yields more similar means to those of the target form 2004–05. The mean of the equated scores using 3PL&GPC-TS (true score) and 3PL&GPC-OS (observed score) were 38.68 and 38.76, respectively. The differences from form 2004–05 means are 0.08 and 0.00, which are much smaller than those of the 3PL-TS and 3PL-OS methods (0.74 and 0.65). However, the 3PL&GPC method produces much larger differences in standard deviation (SD).

Table 2. Moments for Equating Form 2006–07 to Form 2004–05 for Linear and Equipercentile Methods and IRT Methods

| Test         | Mean | |DIFF| |SD | |DIFF| | Skewness | |DIFF| | Kurtosis | |DIFF|
|--------------|------|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Listening**|      |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Form 2004–05 | 38.76| 6.06| -0.643| 3.209|
| Form 2006–07 | 34.03| 6.80| -0.348| 2.794|
| Linear       | 38.76| 6.06| 0.00| 2.794|
| Equipercentile| 38.76| 6.07| 0.01| 0.295|
| 3PL-TS       | 38.02| 6.06| 0.00| 0.415|
| 3PL-OS       | 38.11| 6.02| 0.04| 0.164|
| 3PL&GPC-TS   | 38.68| 5.82| 0.24| 0.328|
| 3PL&GPC-OS   | 38.76| 5.86| 0.20| 0.324|
| **GCVR**     |      |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Form 2004–05 | 68.03| 11.48| -0.208| 3.010|
| Form 2006–07 | 66.79| 11.66| -0.164| 2.980|
| Linear       | 68.03| 11.48| 0.00| 2.980|
| Equipercentile| 68.03| 11.48| 0.00| 0.030|
| 3PL-TS       | 67.95| 11.51| 0.03| 0.071|
| 3PL-OS       | 67.95| 11.47| 0.01| 0.082|
| 3PL&GPC-TS   | 67.78| 11.99| 0.51| 0.054|
| 3PL&GPC-OS   | 67.78| 11.84| 0.36| 0.272|
For the GCVR section, the 3PL equating method provides more similar moments to those of the target form than the 3PL&GPC method. In terms of 3PL true scoring equating, the differences in mean, SD and kurtosis from the target are 0.08, 0.03, and 0.082 respectively, which are much smaller than those of the 3PL&GPC method. This was also true for the observed scoring equating. However, the 3PL&GPC equating method produces more similar skewness values.

**Equateing Conditional on NC Scores**

Conditional on each number correct (NC) score, the differences between the equated score of IRT equating methods and the equated score of the baseline equating methods (linear and equipercentile equating) are calculated. The smaller the absolute value of the difference is, the more similar the equating function is to the baseline equating function. Figures 1 and 2 display the difference scores for the listening and GCVR sections, respectively, with the difference score on the vertical axis and NC score on the horizontal axis.

*Listening Section.* In Figure 1, in terms of true score equating, the 3PL&GPC equating function was more similar to the linear equating function than the 3PL method in the score range 15–25. For scores above 25, both the 3PL&GPC and 3PL methods produce equivalents similar to those of linear equating. For scores below 15, the differences between both methods and linear equating are very large (the differences are around -8 for the scores below 10). The cause might be that few examinees scored below 15, which yielded a large amount of equating errors. The patterns are similar for observed score equating.

In Figure 2, in terms of true score equating, the 3PL&GPC equating function is more similar to the baseline equipercentile equating function than the 3PL method for most scores between 10 and 40. For scores above 40, the equated score of the two methods are similar to those from equipercentile equating. Moreover, observed score equating performs better than true score equating, which provides more similar equivalents to those of equipercentile equating.

![Figure 1. Comparsion of IRT Models using Linear Equating as Baseline for Listening Section](image)
Investigating Different Item Response Models in Equating the Examination for the Certificate of Proficiency in English (ECPE)

Figure 2. Comparison of IRT Models using Equipercentile Equating as Baseline for Listening Section

**GCVR Section.** In both figures 3 and 4, the true score and observed score equating provides similar equating relationships except that the 3PL&GPC-TS performs differently below the score 30. The number of examinees who scored below 30 was very small; therefore no reliable pattern would be expected due to the small sample size and large equating errors. In terms of IRT models, the 3PL equating method yields more similar equivalents to those of linear and equipercentile equating than does the 3PL&GPC method for most NC scores.

Figure 3. Comparison of IRT Models using Linear Equating as Baseline for GCVR Section
The different findings for the listening and GCVR sections might be due to the different degree of violation of the local independence assumption. If the assumption is severely violated, the use of polytomous models would eliminate the effect of dependence among items and likely to improve the equating. In this study, it seems that the IRT assumption is violated less with the GCVR section than with the listening section.

**Weighted Root Mean Square and Unweighted Root Mean Square**

Table 3 shows the Weighted Root Mean Square (WRMS) and Unweighted Root Mean Square (URMS) between equated scores and the baselines. These two indices represent the overall level of discrepancy between each IRT equating method and the traditional equating methods.

<table>
<thead>
<tr>
<th>Test</th>
<th>WRMS</th>
<th>URMS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>Equipercentile</td>
</tr>
<tr>
<td>Listening</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3PL-TS</td>
<td>0.985</td>
<td>0.761</td>
</tr>
<tr>
<td>3PL-OS</td>
<td>0.795</td>
<td>0.659</td>
</tr>
<tr>
<td>3PL&amp;GPC-TS</td>
<td>0.798</td>
<td>0.570</td>
</tr>
<tr>
<td>3PL&amp;GPC-OS</td>
<td>0.615</td>
<td>0.461</td>
</tr>
<tr>
<td>GCVR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3PL-TS</td>
<td>0.185</td>
<td>0.236</td>
</tr>
<tr>
<td>3PL-OS</td>
<td>0.140</td>
<td>0.210</td>
</tr>
<tr>
<td>3PL&amp;GPC-TS</td>
<td>0.662</td>
<td>0.669</td>
</tr>
<tr>
<td>3PL&amp;GPC-OS</td>
<td>0.528</td>
<td>0.542</td>
</tr>
</tbody>
</table>
For the listening section, in terms of WRMS, 3PL&GPC equating methods provide more similar equating relationships to the baseline methods than 3PL methods. For example, using equipercen tile equating as the baseline, the WRMS of 3PL&GPC-TS and 3PL&GPC-OS are 0.570 and 0.461, respectively, which are smaller than those of 3PL methods (0.761 and 0.659, respectively). However, using URMS as the criterion, two models perform similarly and observed score equating yields more consistent results with traditional equating methods than with true score equating.

For the GCVR section, 3PL equating methods yield smaller WRMS and URMS than 3PL&GPC methods. This indicates that 3PL provides more similar equating relationships to the baseline methods. This is true for either true score or observed score equating.

Summary and Discussion

As a large-scale certification test with high stakes, the ECPE needs to ensure fairness and consistency in each testing situation. The problem of comparability among test scores using different test forms must be addressed. When equating is conducted under item response theory, failing to take into account the effect of local dependence among items might distort the equated scores, and disadvantage individual test takers.

In this study, two ECPE forms were equated using different IRT models and compared to traditional equating methods. The results reveal that a combination of 3PL and GPC models performed better than the 3PL model for the listening section, especially for low and medium scores (ranging from 15 to 25). However, for the GCVR section, the 3PL model yielded a more similar equating function to the traditional equating function. The dissimilarity between the two sections might be due to the different degree of violation of the local independence assumption.

The choices of dichotomous and polytomous models in this study were restricted. Only the three-parameter logistic model and the generalized partial credit model were investigated, thus the results from this study may not generalize to other models. More IRT models should be studied in the future, such as the graded response model and the nominal model.

In addition, only real data were analyzed in this study and the results might be limited to this data. Further research using simulation techniques needs to be pursued. Using simulation, different factors could be manipulated, such as the percentage of polytomous items in mixed-format tests, the length of tests, and the choice of IRT models. Therefore, the effects of different factors on equating relationships in mixed-format tests could be better evaluated.

Acknowledgments

The author thanks the University of Michigan English Language Institute (ELI-UM) for providing funding for this research, and also thanks Dr. Jeffrey Johnson for editing and reviewing.
References


ABSTRACT This paper forms part of a broader project in which we have attempted to “track” a marking guide for open-ended (or “limited production”) listening test items from its inception to its eventual use among assessors. One key issue that has arisen from this ongoing project is assessor decision making while using and applying a marking guide for open-ended test items. In contexts where open-ended items are used, a marking guide is intended to stipulate precisely what kind of response should be accepted as evidence of the ability under test. However, there often remains scope for assessors to apply their own interpretations of the construct in judging responses that fall outside the necessarily limited information provided in a marking guide. This paper explores this issue in relation to the listening component of the Occupational English Test (OET)—a language test for overseas-qualified health professionals wishing to work in Australia.

The study presented in this paper forms part of a larger research project being conducted at the Language Testing Research Centre that aims to chart the development and eventual use of a marking guide for the listening component of the Occupational English Test (OET)—a language test for overseas-qualified health professionals wishing to work in Australia—with a view to informing innovations in the design of future marking guides. Marking guides are regularly used in tests with open-ended response tasks “in order to avoid introducing inconsistencies in the way scorers apply criteria of correctness” (Bachman & Palmer, 1996, p. 207). This is often achieved through several iterations and trials during which time the test developer creates a guide (often called a “key” or “scheme”) that contains a list of acceptable answers. However, while a marking guide is intended to stipulate precisely what kind of response test designers accept as evidence of the ability under test, there remains scope for assessors to apply their own interpretations of the construct in judging responses that fall outside the limited information provided in the marking guide. As key stakeholders in the implementation of a test, it is important that assessors’ interpretations be taken into account.

1 We would like to acknowledge the cooperation of the OET Centre, and in particular the involvement of John Pill, who was the joint author of a paper related to this study which was presented at the 2008 Language Testing Research Colloquium in Hangzhou, China.
account, not only for the purpose of reliability, but also to ensure that the test does indeed reflect the ability that it aims to measure.

With these concerns in mind, the project has been guided by three overarching research questions:

1. What decisions do markers make while working independently with a marking guide to score answer papers?
2. Do these decisions accord with the test designers’ interpretation of the construct?
3. How sufficiently are the needs and perceptions of the users of the marking guide addressed in the design of the marking guide by test developers?

This paper concentrates on the first of these questions.

Marking the OET Listening Subtest

The case under study in this project is the Occupational English Test (OET). The OET is used to evaluate the English-language competence of qualified medical and health professionals who wish to practice in an English-language context. The test was originally developed by McNamara (see 1990; 1996), and has been used throughout Australia since the early 1990s. It is now also used in New Zealand and Singapore. The test seeks to ensure that candidates are prepared, in language terms, for the world of work in the professions of dentistry, medicine, nursing, pharmacy, physiotherapy, dietetics, occupational therapy, optometry, podiatry, radiography, speech pathology, and veterinary science. The OET is made up of four subtests, defined simply as the four macro skills: speaking, writing, reading, and listening.

The focus of this research, the OET listening subtest, consists of two parts. In Part A, which usually takes around 25 minutes, candidates listen to a consultation between a health professional and a patient/client (e.g., a general practitioner and his/her patient). Candidates are required to take notes under a series of headings, provided to simulate taking a case history and notes during a consultation. In Part B, also lasting approximately 25 minutes, candidates listen to a monologic talk or seminar presentation on a health-related topic and complete a mixture of fixed-choice and open-ended task-types including: sentence completion, short answer questions, lecture note completion, chart/diagram completion, and other similar tasks. These broad specifications are shown in Table 1 below.

Table 1. An Overview of the OET Listening Subtest

<table>
<thead>
<tr>
<th>Part A</th>
<th>Part B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Consultation between patient and health professional (approx. 25 minutes)</td>
</tr>
<tr>
<td>Task(s)</td>
<td>Notetaking under headings</td>
</tr>
</tbody>
</table>
Both parts of the test are marked by a group of 8–10 experienced assessors who are specially trained for each administration (this is because a different set of test materials is used for each administration of the test). Scoring is achieved with reference to a detailed marking guide developed at the Language Testing Research Centre.\(^2\)

In order to illustrate what assessors typically work with while marking, an example task from Part B (which will be the focus of this paper) and its related marking guide entry are illustrated in the figures below. Figure 1 shows a diagram completion task in which the candidate is asked to note down a range of risk factors or predisposing factors that may lead to skin cancer. The corresponding information in the marking guide is shown in Figure 2.

---

\(^2\) Both parts of the test work consistently well, and regularly achieve reliability estimates above 0.9. Another point of interest is that candidates around the cut score are routinely double marked.
The notation used in the marking guide is such that a forward slash “/” indicates that either adjacent word is acceptable (i.e., “trauma to skin” or “damage to skin” are both acceptable responses for the same mark); the bolded uppercase “OR” indicates that either of two distinct terms is acceptable (e.g., “excessive sun exposure” or “sunburn” are both acceptable responses for the same mark); and the bolded uppercase “AND” indicates that both pieces of information are required for a correct response, so that “chlorinated pools” on its own is not sufficient without “sun.” Thus, according to the marking guide, the responses shown in Figure 1 would have gained the candidate four marks out of a possible six, with no mark given for “chemicals” or “in ships.”

The Challenge of Marking Open-Ended Responses

In designing the marking guide, the key considerations are that the entries be succinct, but that a full range of potential answers be included to the greatest extent possible. This range of answers is derived partly through paneling sessions where items are discussed among test developers, and partly from the range of responses provided by a trial population. However, while the marking guide is designed to be as accurate and comprehensive as possible, its contents do not comprise a list of the only acceptable answers. This is primarily because the communicative and authentic approach of the OET allows a certain level of flexibility in judging the correctness of a response to an open-ended item. Specifically, the assessors for the OET are instructed to accept spelling variations and misspellings; to accept abbreviations (which are used extensively within the health professions); to disregard grammatical errors that do not affect meaning; and to consider synonyms or alternate phrasings of answers on their merits. To this end, the guidelines that assessors read on their marking guides instruct:

The essential point to keep in mind is whether an answer indicates an appropriate response to the question not whether it follows the suggested answer verbatim.

While these broad parameters allow for greater flexibility, and arguably a more valid scoring procedure for the assessment of the construct of listening, they also pose challenges both for test developers and for assessors in that candidates may respond in ways that have not been anticipated. This challenge for marking open-ended task types in a variety of language assessment contexts has been noted in several texts (e.g., Weir, 1993; Bachman & Palmer, 1996; Alderson, 2000). It is generally recommended that, to ensure objectivity, test developers should take steps to make their marking guide as comprehensive as possible. However, while trialing marking guides can help to achieve this level of breadth, it does not entirely remove the potential for unanticipated responses to arise in test takers’ scripts. When a response does fall outside the necessarily limited information given in the marking guide, the onus falls upon the assessor to interpret the acceptability of the response in relation to the content of the text and the requirements of the task (see Buck, 2001). This need for interpretation introduces an element of subjectivity into the marking process.

One of the ways in which this scope for interpretation is “managed” within the assessment procedures for the OET is through a routine markers’ meeting. The flowchart shown in Figure 3 illustrates that once the marking guide has been developed and refined after trial, it is then handed over to a test provider. The assessors then take part in a training session
during which they familiarize themselves with the marking guide. At this training session, assessors have the opportunity to query the contents of the marking guide directly with test developers and with the assessment manager. The way they do this is by completing the full test themselves and then using the marking guide to mark their own papers. The experience of comparing their own responses with those in the marking guide always raises issues for discussion and these are worked through, as a group, item by item. Once this process is completed, assessors all mark the same sample script and compare how they awarded marks for each item as a standardization check. Again, issues with the interpretation of responses are raised and consensus is reached among the group. In the days following this meeting, assessors begin to work independently with the marking guide in their routine marking.

![Diagram](image)

**Figure 3. The Marking Guide: From Designers to Users**

The process illustrated above serves to limit the decisions that markers need to make during the final stage, and to ensure greater consistency among assessors when decisions do need to be made. However, following previous administrations, the assessors’ log (a common notebook kept for markers to record any difficulties they experienced during routine marking) has consistently shown that assessors continue to make detailed notes while they are marking about particular items where they have needed to make a decision about a candidate’s response that does not match the information given in the marking guide.

As a preliminary exploration of these sites of decision making, the study reported here presents data collected after the final stage in which a group of assessors reflected on their experiences working independently with the marking guide. This study was intended to address one of the overarching research questions articulated at the beginning of the paper:

What decisions do markers make while working independently with a marking guide to score answer papers?
We broke this overarching question down into two more specific points of enquiry for the current study:

1. What are common sites of decision making for assessors while marking Part B of the OET listening subtest?
2. On what bases do assessors make decisions?

Methods

To investigate these research questions, we conducted a stimulated recall group discussion (Gass & Mackey, 2000) with three OET assessors following their independent use of a marking guide in a routine administration of the OET.

Participants

The three assessors who took part in the focus group discussion were Laura, Emma, and Penny. All were experienced ESL and literacy teachers, and two (Laura and Emma) had been marking OET scripts for many years. The discussion was moderated by the two authors of the paper, who may be conceived of as participant observers, given that they are also developers of the OET listening subtest. The OET assessment manager was also peripherally involved in the discussion, but took on more of an observer role.

Process of Data Collection

Three assessors were asked to volunteer to take part in the research project during a routine markers’ meeting. After Laura, Emma, and Penny volunteered, they were instructed that they would be asked to join in a focus group discussion after they had worked independently with the marking guide after the administration of a new test. They were asked to keep any notes they made on the marking guide with them, and to be mindful of any particular difficulties they had in applying the marking guide, or any points at which they felt they had to make a decision on a response that fell outside the parameters of the marking guide.

Following routine marking (which is conducted over weeks), the focus group was convened. The discussion lasted around two hours (including an interval break) and was video-recorded. The rationale for video-recording the discussion was so that each speaker could be easily identified at the transcription stage.

The assessors were asked three general questions at the beginning of the discussion in order to elicit some background information. The discussion then followed a structure in which we (the researchers/developers) worked through each item of the test in order, and asked the assessors to comment at any points that they recalled having difficulty. In this way, the marking guide itself provided a stimulus for recall of decision making. This item-by-item stimulated recall was supplemented by the more detailed notes each assessor had kept on their own sites of decision making.

---

3 These are pseudonyms.
Data Analysis

The focus group discussion was transcribed and analysed for content by the two researchers to identify macro categories of “sites of decision making” (see research question 1). This process involved each researcher coding the transcript separately, and then collaborating to compare how they had categorised the data and reach consensus. The macro categories that emerged are discussed below, together with representative examples. Within each category, several “bases for assessor decisions” were also identified (see research question 2). These are also discussed throughout the next section.

Common Sites of Decision Making

From the outset, it should be stated that the data indicated that assessors were generally very happy with the marking guide and felt that in most cases it provided clear guidance for them to mark test papers with confidence. The data did, however, yield comments on several sites of decision making, and after analysing the data, these have been divided into three broad categories:

1. Decisions regarding spelling;
2. Decisions regarding the correctness of an overelaborate response;
3. Decisions regarding the adequacy of response.

In the sections below, each of these sites will be explained in more detail, and the bases upon which assessors made decisions at these sites will be illustrated and discussed.

Spelling

This first category—decisions regarding spelling—appeared within the data to be the most common site of decision making for assessors during the marking process. Excerpt 1 below illustrates one typical instance in the data where an assessor reported difficulty in dealing with variant spellings of the word “tough”:

Excerpt 1

Laura: So, can you take “t-a-f”? ‘Cause when you say it, /tʌf/, it sounds like tough. There’s no difference. And then when they attempt to remember what the awkward spelling is for this word, they were coming up with things like “t-a-u-g-h”. So, I tended to give them… both.

Despite the prevalence of misspellings, assessors generally reported being able to make swift decisions by applying a “rule of thumb” that had been agreed on at a previous markers’ meeting. This rule holds that a misspelling be deemed acceptable provided it is a reasonably close phonemic match of the target word, and the meaning of the response remains clear. Thus, as one of the assessors stated with regard to “tough”: 
Excerpt 2

Penny: If it sounds like tough when you say it, then I gave it.

Also according to the spelling rule, if a word is misspelt and thereby becomes another word with a different meaning that could plausibly be part of the answer (though it would be wrong), it is not allowed. In this sense, a misspelling is distinguished from a mishearing. This second aspect of the “rule of thumb,” however, caused some difficulty for assessors, as they recalled specific instances where distinguishing between a misspelling and a mishearing was not necessarily clear-cut. For example, Excerpt 3 is drawn from discussion over an item where the desired response was “asymmetry,” but where several candidates had written “eye-symmetry”:

Excerpt 3

Emma: Well, if it was “e-y-e-symmetry,” I didn’t give it. But if it sounded like one word, “a-s-s-e-m,” sort of thing, you know…

Researcher A: Or “i-s”…?

Emma: Yeah.

Laura: But if it was “e-y-e-symmetry,” as one word. “E-y-e-s-y-m”… as one word… “eyesymmetry”?

In the example above, it is clear that Laura was not sure whether “eyesymmetry,” as an unhyphenated term, should be considered a mishearing in the same way as “eye-symmetry.” In this case, there was no clear basis for making a decision either way.

Underlying some of these difficulties was the notion that the assessors appeared sympathetic towards the candidates in dealing with the somewhat broad, Australian English pronunciation of this particular lexical item in the text. Indeed, applying the rule of thumb in its strictest form did not always sit comfortably with the assessors’ intuitions about the ability of a candidate they were marking, to the extent that the assessors viewed common mishearings to be symptomatic of problems with the text rather than perceptual errors. This is also demonstrated in Excerpt 4, which concerned an item where candidates were required to note down the term “normal skin.”

Excerpt 4

Laura: I was fascinated by the number who had done very well, or were doing very well, but wrote down “noble skin”.

Emma: Yes.

Laura: And I just wondered if it’s possible to listen to it and could it have sounded like a /b/ rather than an /m/?

Nevertheless, even though the assessors were troubled by the application of the rule of thumb in certain circumstances, the three assessors generally agreed with each other with regard to the acceptability of the spelling variations that were discussed.
Correctness of an Overelaborate Response

In contrast to spelling, decisions regarding the correctness of an overelaborate response was only reported once in the data. The category here refers to instances in which candidates would write very long responses to short answer questions that included an unnecessary level of extra information (probably based on their general knowledge of the subject) which the assessors needed to judge for correctness. According to the marking guidelines, even if a candidate notes down the correct information, they might still “talk themselves out” of a correct response if they negate the desired response in any additional information they provide. In the excerpt below, Laure reports having struggled with knowing how to mark responses such as these:

Excerpt 5

Laura: A lot of them put in information that they would know, like, ah, “can’t use this for such and such because”... and they were elaborate answers... and I’m sure they were right...

Researcher A: But they sometimes didn’t match what was in the marking guide?

Laura: I would say that I wasn’t absolutely sure if it matched what was in the marking guide...

Because this was the only example of this type of answer in the data, there was no clear indication of how assessors dealt with an overelaborate response. The excerpt above seems to suggest a level of trust from Laura that the candidate—who is a health professional—must know more about the topic than she does “and I’m sure they were right...,” indicating that this might be a default approach in such cases.

Adequacy of Response

The third category—decisions regarding the adequacy of a response—arose when an answer did not match the marking guide, such as when answers were paraphrased, or when synonyms were used. This site of decision making was further divided into two subcategories:

i. Semantic distinction

Making a decision about whether a particular alternate word or phrase in a response demonstrated understanding of what the speaker had said, or whether it showed understanding of a different concept

ii. Sufficiency of an answer

Making a decision about whether enough information was included in an alternate answer to sufficiently match the idea represented by the answer in the marking guide

An example of where the issue of “adequacy of response” was resolved without great problem may be seen in the excerpt below, which shows a discussion between the three assessors revolving around the semantic distinction between “petrol chemical” and “petrochemical.” “Petrochemical” was the desired answer in the marking guide, and the
assessors are recalling their decision over whether they would accept “petrol chemical”—which was a common response in candidates’ papers—as a reasonable alternative.

**Excerpt 6**

Emma: Ah yes “petrol chemical” rather than *petrochemical*

Laura: I gave “petrol chemical” right from the start

Emma: Well I think we decided to do that

Penny: See that’s what happens also, we discuss it a few days in and then come up with a consensus but what happens at the very beginning is sometimes a bit…

In this example, the strategy recalled by the assessors in dealing with a problem of semantic distinction was to discuss the meaning of a common alternative answer and to reach a consensus. This strategy echoes Weir’s (1993) recommendation for utilising a marking guide that “marking be done in the same room initially until these variations in acceptable answers are finalised” (p. 112). However, for the OET assessors, the consensus approach during the marking process was only possible when assessors were marking in the same location at the same time (which is not always the case), and where there were large numbers of candidates giving exactly the same type of alternative response (to the extent that it is a common enough variation to be remarked upon).

**Points of Divergence in Judging Adequacy of Response**

The data concerning other instances of “adequacy of response” were noteworthy because, in describing the challenges they faced in marking responses that fell into this category, a number of issues emerged that illustrate the ways in which different assessors may perceive the construct in different ways. When answers deviated from the marking guide, and when there had been no scope for a consensus discussion, assessors needed to “fall back” on their own conceptualization of the abilities under test in order to make decisions. In the case of the OET, this means their conceptualization of listening ability in a health context. In illustration of this, Excerpt 7 below shows a situation where assessors demonstrated mindfulness of the construct when making a decision. In this example the assessors recalled their decisions in dealing with an item where the desired response was “excision biopsy,” and a small number of candidates were writing “incision biopsy”:

**Excerpt 7**

Laura: And if they wrote “incision biopsy” instead of *excision*?

Emma: Yes, *excision*.

Laura: Should we have marked them right or wrong? Is it listening… or is it understanding? And is there a diff… a real difference between “incision” and *excision biopsy*.

Emma: It’s hard to do an excision without an incision… but it has a particular meaning.
When faced with these sorts of semantic issues, the challenge for the assessors was to gauge whether the candidate had actually understood what the speaker said and had paraphrased the information (indicating a deep understanding), in which case to penalize that candidate would be to reduce the construct to one of “listening” (Laura’s term), which might be understood here as a construct of basic word recognition. However, balanced against this mindfulness that the use of a synonym might indicate “real” comprehension is the notion that certain terms must be used in precise ways in health contexts. This notion is clearly articulated in Emma’s coda in Excerpt 7: “but it has a particular meaning.”

A further example of a discussion revolving around a semantic distinction shows clearly that the issue of precision formed a point divergence in how assessors judged the acceptability of another response. Excerpt 8 shows the first part of conversation in which assessors recalled an item where candidates were required to note down “liquid nitrogen,” and where several candidates wrote down “liquid nitrate”:

**Excerpt 8**

Emma: “Liquid nitrate”…

Researcher A: And you accepted it as “nitrate” for nitrogen?

Emma: Yes I did, because I didn’t know the difference but then I went home and found out

Penny: I, if I got it, I’m pretty sure I wouldn’t have accepted it…

This excerpt shows very clearly two different decisions about the same response from two assessors—Emma and Penny. In Emma’s case, her initial decision was attributed to a lack of knowledge about the difference between the terms “nitrate” and “nitrogen.” However, even though Emma reported that she did find out the difference, in Excerpt 9 she defends her initial decision, while Penny argues for the need for precision on this item:

**Excerpt 9**

Penny: Liquid nitrogen is a particular item so…

Emma: But you have to know that

Penny: But these people are health professionals

Emma: If they are a speech pathologist why would they know that?

David⁴: Speech pathologists know a lot

Penny’s argument—“but these people are health professionals”—appears to be based on the notion that the OET is a test of communicative competence for specific purposes, and as such, that the distinction between nitrate and nitrogen is information that one should expect health professionals to know. Yet, Emma’s counter argument is premised on the idea that this type of knowledge might not be expected of all professions who take the test; in effect, an argument for fairness.

---

⁴ Pseudonym for the Assessment Manager
It is evident in the two examples above concerning “excision biopsy” and “liquid nitrogen” that, in each case, decisions over a single semantic distinction appear to have been informed not simply by the assessors’ judgments of whether one word was a plausible synonym for another, but by much broader notions that encompass the assessors’ understandings of the construct; their conceptualization of the overall purpose of the test; and their desire to maintain fairness in evaluating responses from candidates’ with a range of professional backgrounds.

Assessor Knowledge

The discussion presented in Excerpt 9 was also noteworthy because it signaled that the bases for assessor’s decisions in instances where the marking guide was not sufficient might depend, to some extent, on assessors’ own level of knowledge in, sometimes, particularly technical areas of medicine. Particularly, Penny appeared to have a somewhat different orientation compared with Emma and Laura, and seemed to expect a higher level of “health literacy” from the candidates. Interestingly, when we reviewed the background data each assessor provided during the discussion, it revealed that Penny has quite a different background from the other two assessors who were involved in this study.

Penny was currently working as an ESL literacy teacher, but she also had a strong research background in health science. In the interview, she said of this background, “I don’t know if it makes me a better or worse marker… it complicates things sometimes.” The other two assessors came from a primarily literacy/ESL teaching background. Emma had a particularly strong marking background, having worked on a number of assessment programs, whereas Laura also had a background as a consumer representative for cancer services, and said of her own health literacy, “I’d certainly be above average for knowledge in the medical field.”

A final example—concerning a decision regarding the sufficiency of a response—demonstrates clearly how these sorts of knowledge bases were brought to bear on decisions throughout the marking process. Excerpt 10 below shows the discussion of an item where candidates were asked “What is one of the confirmations a histopathologist can make with a malignant skin cancer sample?” and where they were required to write down “margin of safety around the tumour.” The discussion begins with Penny expressing a divergent view from the other two assessors, which leads to a debate over whether “safety around the tumour” is acceptable on its own, or whether the word “margin” is necessary for a sufficient answer.

Excerpt 10

Penny: You took “safety around the tumour”?
Laura: Yes, I thought it indicated an area of safety
Penny: Ah, ah “safety around the tumour,” I don’t think I gave it
Emma: Oh no no “safety around the tumour” for instance I can see that that it’s not really wrong
Laura: If it confirms a safety, if it confirms what the safety…
Penny: I understood it differently as I didn’t give it, I think “safety around the tumour” can be understood in lots of ways. I didn’t think it was
enough information, but I am willing to accept that I am too harsh sometimes.

Laura: Well what does “safety around the tumour” mean to you?

Emma: Or *margin of safety?*

Penny: “Safety around the tumour” could mean how virulent it is or something like that.

The reasoning presented for both “sides” is similar to that presented in Excerpt 9. Emma’s statement, “safety around the tumour… I can see that it’s not really wrong,” indicates a degree of tolerance; perhaps falling back onto the position identified earlier of trying to be fair, and not assuming that all test takers would recognize the need to include the specific expression “margin of safety.” Laura also concurs: “yes if it confirms the safety…..” However, Penny’s turn, which begins, “I understood it differently… I think safety around the tumour can be understood in lots of ways. I don’t think it was enough information,” is made clearer in light of the knowledge that she has direct experience in the health field.

In Penny’s case, her background knowledge forms another external factor that affects her decisions at this micro level. Penny was, nevertheless, aware of this tension—"I am willing to accept that I am too harsh sometimes”—and elaborated on these issues thus:

**Excerpt 11**

Penny: I tried trying to stick to the marking guide but then when I look at an answer that has that diverge from the marking guide, I try and work out whether the person understood what they were supposed to understand. I mightn’t be right here about not accepting… “safety around the tumour”. That was my understanding though. I was based… because I think my understanding about a lot of this is fairly good.

Penny’s reflection that “I try to work out whether the person understood what they were supposed to understand” highlights one of the fundamental challenges for test developers training assessors to use marking guides: chiefly, that for each item, assessors must develop similar conceptualizations of what a candidate is supposed to understand, and also agree on whether or not a candidate’s response shows adequate evidence of this understanding. In other words, in this specific purposes context, for assessors to make consistent decisions about items that fall outside the parameters of the marking guide, they must share a deep knowledge of the construct, a similar level of knowledge of the content of test input, and similar notions of how language might be used minimally in a written response to express understanding.
Summary and Implications

This study raises several implications for the broader project of making innovations in the design of useful marking guides for open-ended listening test items. Firstly, as noted earlier in this paper, feedback from assessors indicated that, overall, the marking guide works well and is generally applied without problem in marking the vast majority of items on any given test in which assessors have been involved. The instances detailed above were viewed as exceptions, and, as test developers, this is encouraging. However, through this small-scale study we were able to identify three common sites of decision making for assessors when candidate responses deviated from the marking guide, and then to explore some of the strategies that assessors use to deal with those types of responses, and, as an extension, some of the external factors that may influence the decisions made by individual assessors.

The findings are noteworthy in that it seems that assessors were highly conscious of wanting to achieve consistency both in their own marking and with the marking practices of other assessors. It was also interesting to find—as the analysis above has shown—that assessors are deeply mindful of much broader issues concerning the purpose of the OET and the construct under test in their general approach to marking listening papers. However, this study also raises questions about how to deal with the fact that differences in individual knowledge bases, or in the orientations of these assessors toward the test, may lead to slightly different positions on what assessors accept as demonstrations of understanding in those instances when test takers’ responses deviate from the marking guide—particularly with regard to decisions of the “adequacy of response” type.

One method that might be useful in following up this issue would be a quantitative analysis of assessor consistency and severity using multifaceted Rasch analysis (see McNamara, 1996). Using estimates of assessor behaviour generated from the analysis, we would be able to identify any assessors who were marking very differently from others, and use this information to give individualised feedback in preparation for the next round of marking (see Wigglesworth, 1993; Elder, Knoch, Barkhuizen, & von Randow, 2005). However, this process cannot remove these sites of decision making in the OET listening test; and, theoretically, every test administration presents a new set of entirely unique challenges for assessors that may not be rectified through a process of giving feedback on their past marking behaviour.

A different course of action would be to focus more energy on ensuring that assessors share as much knowledge as possible either prior to marking, or in the very early stages of marking. Firstly, it would be very useful to encourage a greater level of dialogue between test developers and assessors in order to communicate the construct and rationale for decisions about what is in the marking guide more rigorously and more collaboratively at the training stage. During the group discussion it became clear that assessors seek reassurance, and are particularly keen to talk with the test developers to achieve some sort of shared understanding. In this vein, a longer and more thorough training stage might be useful where assessors (together with the developers and the assessment manager) concentrate on a number of scripts chosen carefully as being “problematic.”

Secondly, it appears the greater collaboration between assessors themselves at the early stages of the marking process would be beneficial. We have seen from one of the examples above (“petrochemical” versus “petrol chemical”) that this can be useful; assessors can learn from, and be assured through, dialogue with each other. Having assessors marking
in the same room together for at least the first day of routine marking would create the opportunity for further discussion of matters that commonly arise early.

Thirdly, during the course of the interview, assessors raised the possibility of a “stage one” check—in the form of a second markers’ meeting—after an initial set of papers have been marked. This would be advantageous because more sites of decision making will have been identified in the early stages of marking, and these can be worked through as a group a second time, with the test developers involved. Again this could provide reassurance for markers, and give them structures to fall back on. Ideally, this might occur after assessors had marked 10–20% of their papers.

Finally, this research highlights the importance of routinely gathering assessors’ perspectives on a marking guide in order to help developers predict problematic items. This is because, as we found through engaging in this process, the assessors—having marked over one hundred papers themselves at each administration—have some very interesting things to say about which items are “working” as well as which items appear to be testing deep understanding, and which do not. Our future projects will seek to involve the assessors as important sources of knowledge not only with respect to the marking guide, but also with respect to the test itself, and in this sense, to receive feedback “from the coal face.”

References
