Evaluating Test Consequences Based on ESL Students' Perceptions: An Appraisal Analysis

Elizabeth Lee[1]
Iowa State University

ABSTRACT

Ensuring that test-score use brings about socially positive consequences for test-takers is an important aspect of test validation. While many studies use an inductive approach to evaluate test consequences, few studies have implemented Appraisal analysis. To that end, this case study investigated the test consequences of an English reading placement test that was administered at a large American university. In this study, English as a second language (ESL) students (n=8) who took the placement test and an ESL reading course were interviewed; an Appraisal analysis was conducted to identify the students' positive and/or negative perceptions toward the placement test and the ESL reading course. Using an argument-based approach to validity framework, the findings were treated as evidence to evaluate the test consequences of the placement test. The results showed that, while taking the ESL course helped students gain some valuable academic reading skills, students felt that test anxiety, fatigue, and verbally demanding questions hindered their test performances. Understanding what students experienced while taking the test can help test-developers devise solutions that will improve the test-taking situation for future test-takers. This study also illustrates how an Appraisal analysis of test-takers' discourses can provide a systematic and fine-grained approach to evaluating positive and/or negative test consequences.

Key words: *appraisal, placement test, validity*

INTRODUCTION

For second language (L2) learners in higher education, the ability to comprehend academic materials is critical to their academic success (Neumann et al., 2019). Previous studies show that advanced English as a second language (ESL) students are better able to comprehend academic materials than lower-level ESL students (Mokhtari & Sheorey, 1994). Despite variations in students' understanding, college-level ESL students of all levels generally dedicate greater amounts of time and effort to their academic reading than native speakers of English, as a result

[1] Elizabeth Lee received her Ph.D. in Applied Linguistics and Technology at Iowa State University. Her research interests include second language testing and assessment and second language reading and writing. Correspondence should be sent to Elizabeth Lee, E-mail: orangexpants@gmail.com

of English being their second language (Sheorey et al., 1995). While it is expected that processing academic texts would involve more effort for non-native speakers of English, this may raise some concerns in higher education settings where instructors do not differentiate readings or assignments based on students' language proficiency, particularly for lower-level ESL students.

In order to support these learners' academic reading needs, many English-medium universities develop and use placement tests to determine which students would benefit from receiving additional ESL language support (Green, 2012). As with any high-stakes tests, it is important to establish strong validity with placement tests, because failing to do so can negatively affect not only placement decision-making but also practical matters such as tuition, access to scholarships and other funding opportunities, and time to degree completion. The stakes attached to placement-test-score use are high, and therefore, ensuring that score use brings about socially positive consequences for test-takers is important.

To demonstrate how test consequences are investigated using the argument-based approach to validation, the degree to which placement-test-score use brings about socially positive consequences for test-takers was measured using an Appraisal analysis. Martin and White's (2005) Appraisal theory was adopted because the framework provides a nuanced linguistic approach to understanding speakers' positive and/or negative attitudes. The present study examined ESL students' perceptions of a placement test and an ESL reading course based on their positive/negative evaluative language choices. Prior research used inductive approaches to analyzing stakeholders' perceptions of test consequences (e.g., Cheng et al., 2007; Han & Cheng, 2011; O'Laughlin, 2011). While these studies show that negative test consequences inevitably hurt stakeholders, the discussions focus on the drawbacks of test-score use rather than evenly considering both its benefits and drawbacks. If we are to make a fair evaluation of test consequences, however, it is important that we ask stakeholders to consider outcomes that are both positive and negative.

An argument-based approach to validity (Kane, 2006) enables analysts to systematically and empirically investigate whether and to what extent positive and negative test consequences exist. This is done by first establishing a network of inferences in which one's arguments about the test-score interpretation and use are clearly stated in the form of if-then statements. We subsequently examine the validity of each argument by gathering relevant empirical evidence that would either support or reject this argument. If we claim that a test-score use brings about positive test consequences, and evidence supports this, then we could argue that there exists some degree of positive test consequences. On the other hand, if the gathered evidence contradicts our assertion, then this would likely weaken the validity of our claim (see the Literature Review section for more details).

In this study, an Appraisal analysis is used to quantitatively and qualitatively examine the extent to which test consequences are socially beneficial for ESL test-takers. As such studies are lacking to date, the present study demonstrates how findings from an Appraisal analysis can inform the validity of one's claims about test consequences.

LITERATURE REVIEW

An Argument-Based Approach to Validity

While today it is viewed as a system of claims expressing score interpretation and use, the concept of validity has undergone a number of changes over the decades. Before the early 1950s, validity was treated as a property of a test and was examined mainly by correlating the test score of concern to other assessment measures (Shaw & Crisp, 2011). Starting in the mid-1950s, however, the definition of validity expanded to a trinitarian model which includes content validity, criterion-related validity, and construct validity (Shaw & Crisp, 2011). Briefly, content validity is an estimate of how much the targeted content domain is assessed by the test; criterion-related validity is an estimate of how well the test predicts future performances; and construct validity is an estimate of how well the test relates to theoretical constructs (Shaw & Crisp, 2011). Moreover, beginning with the trinitarian model of validity, researchers maintained that validation is an interpretation of the test rather than the test itself (Cronbach, 1971). Over time, this interpretation-based validation was expanded into an argument-based approach to validity. A prominent figure in this development was Messick (1989), who constructed a unified validity concept based on a matrix of test interpretation and use on the one hand and on evidential basis and consequential basis on the other. A number of models subsequently emerged, such as Bachman and Palmer's (2010) assessment use argument framework and Kane's (2006) interpretive argument framework, the latter of which is adopted in the current study.

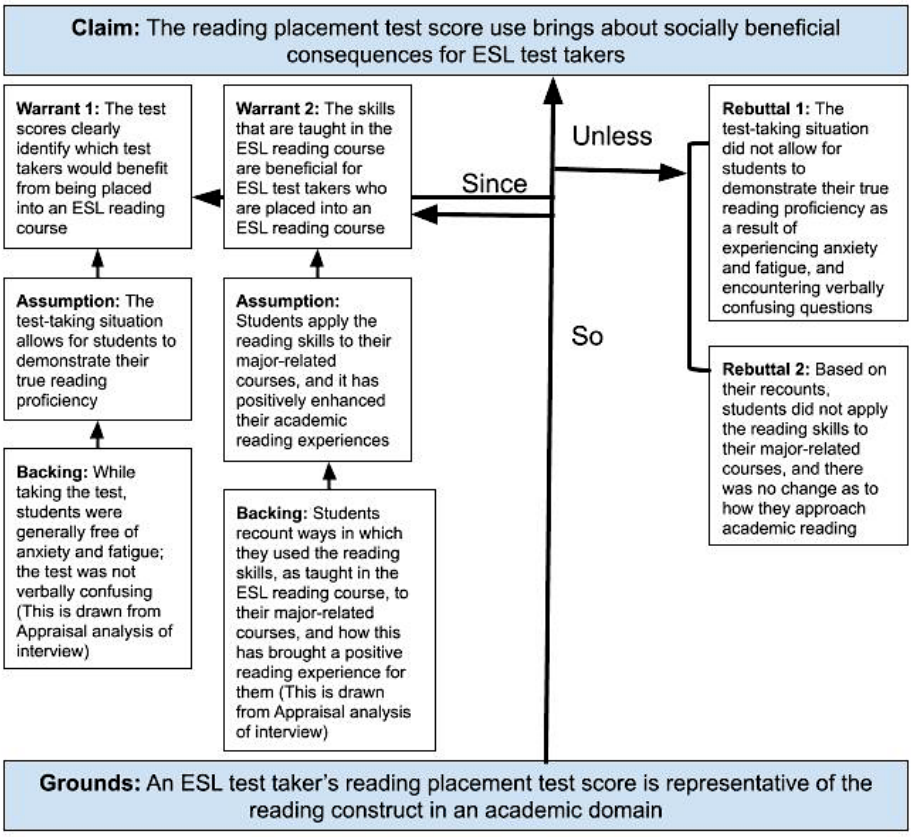
In an argument-based approach to validity, test-score interpretations and uses are posited and established within a network of inferences, which are then backed by scientific evidence (Kane, 2006). In other words, using empirical evidence, an argument-based approach to validity examines the degree to which one's argument about a test-score interpretation and use is justified. A test-score *interpretation* relates to a construct definition or what the test-score intends to measure about a learner's language ability. A test-score *use* relates to intended decisions and test consequences. While empirical evidence is required to justify interpretations and uses, evidence alone does not guarantee that one's argument is valid. Ultimately, our arguments must be able to convince our stakeholders that the interpretations and uses of a test score are appropriate and justified (Kane, 2006; Messick, 1989).

To convince our stakeholders, arguments must be clear, coherent, and plausible. This is achieved by tightly laying out our claims, warrants, and assumptions in a network of inferences (Kane, 2006). The mapping of a single inference, as can be seen in Figure 1, follows Toulmin's (1958) model of argument. A *claim*, or conclusions (e.g., a decision for the decision inference, Kane, 2006), moves from the *grounds*, which are data or sources of previously collected evidence. The *claim* is conditional on one's stated *warrants* (i.e., rules that allow one to move from the grounds to the claim) and *assumptions* (i.e., statements that unpack the warrant and that are evaluated through research), and these require *backing*, or empirical proof, as assumptions and warrants are not self-evident. The more backing we collect, the more confident we are in the soundness of our assumptions and warrants, which in turn support the validity of our claim. A claim may also include exceptions or *rebuttals*, which are conditions that mitigate the strength of one's claim. If the evidence supports any of these rebuttals, the strength of the claim is weakened.

In an argument-based approach to validity, there exist several inferences that enable one's argument to move from observed test performances to score-based interpretations and uses. However, for the purpose of the present study, we will limit our discussion to the inference that specifically deals with test consequences, that is, the *decision inference* (Kane, 2006). The decision inference is the last type of inference that appears in an argument-based approach to validity framework. The claims stated within this inference are related to decisions and

consequences that are derived from test-score use, as seen in Figure 1. Findings from the Appraisal analysis were used to evaluate the assumptions associated with the warrants and the claim within the decision inference. The assumptions are specifically examined because, as Kane (2006) states, "[d]uring the appraisal stage, studies of the most questionable assumptions in the interpretive argument are likely to be most informative, but it is also prudent to check any inferences and assumptions that are easy to check" (p. 26). In other words, an argument-based approach to validity requires that analysts carefully examine the assumptions within an interpretive argument. The following section discusses how test consequences are examined.

FIGURE 1
The Decision Inference of the Reading Placement Test



Measuring Test Consequences

Evidence for evaluating test consequences should come from sources that reflect both intended (positive) and unintended (negative) consequences (Kane, 2006; Lane, 2014). One approach to identifying sources of test consequences is construct irrelevant variance (CIV) (Haladyna & Downing, 2004; Messick, 1989). CIV is a concept in which undesired variables interfere with test-takers' performances, and it is considered a threat to validity because the test score reflects not only the construct but also unexpected errors. For instance, potential CIV sources include overly complex wording of stems and distractors, test anxiety, and fatigue (Haladyna & Downing, 2004). While the last two are not always within the test administrator's control, test anxiety and fatigue are nevertheless problematic as they can negatively affect test-

takers' performances.

In addition to CIV, impact is another potential source for evaluating test consequences which is usually observed after placement decisions are completed. For example, students' self-reported experiences of taking remedial ESL courses and how these affected their language learning can inform the impact of test-score use (e.g., Cheng et al., 2007; Han & Cheng, 2011).

According to previous validation studies, high stakes test-score use can bring about unintended consequences. For instance, Cheng and Sun (2015) evaluated the consequences of the uses of the Ontario Secondary School Literacy Test (OSSLT) by drawing on previous interviews and survey studies (e.g., Cheng et al., 2007; Han & Cheng, 2011). The OSSLT is a literacy (i.e., reading and writing) test that is administered to all 10th grade high school students in the Ontario school region. Its purpose is to determine whether students have met the literacy standards set by the Education Quality and Accountability Office in Ontario, Canada. While students may retake the test in 12th grade, students who fail to pass the test, many of which are ESL students, are required to stay an additional year in high school to take ESL courses and improve their grammar knowledge.

According Cheng et al.'s (2007) and Han and Cheng's (2011) research, the OSSLT test posed several problems: For one, the test measured a high degree of vocabulary knowledge rather than purely literacy; and L2 students who lacked knowledge of the test genre were found to be at a disadvantage. Consequently, the authors concluded that score use of the OSSLT brought about more negative than positive test consequences for L2 students, especially as they were required to delay their time to graduation in order to take additional ESL courses. While some students reported that taking these courses helped improve their overall grammar, they continued to struggle in their content-area courses as well as experiencing isolation from the mainstream community.

In another study conducted by O'Laughlin (2011), practices of using IELTS scores at one Australian university were found to be somewhat problematic. At the university where the study was conducted, the policy was that admitted students who scored the minimum band score (6.5) were required to take additional ESL courses. While most first-year undergraduate students who met the minimum IELTS band-score requirement fared well in their academic courses, the author still maintained that failure to use the IELTS scores for diagnostic and tracking purposes did not benefit students and faculty, as such information was crucial for facilitating English language learning and teaching. In other words, in the absence of tracking and diagnostic information, courses were likely to be designed and taught without having a clear understanding of students' language needs, and the amount of progress students made from one ESL course to the next was assumed rather than recorded and shared with all stakeholders. Moreover, the interpretations and uses of the IELTS test score were not made clear to all admissions staff, and as a result, inconsistent decision-making was also observed. O'Laughlin's (2011) research reveals that the consequences of test-score use can yield more negative than positive consequences when score users do not carefully consider the interpretations and uses of a test score.

In summary, as seen with Cheng and Sun's (2015) and O'Laughlin's (2011) studies, it is important to critically investigate whether and to what extent test-score use brings positive and/or negative test consequences. To this end, in the current study, the assumption associated with the first warrant is evaluated by identifying the presence of CIV, and the assumption associated with the second warrant is assessed by measuring the impact of the placement decision (see Figure 1 above). Together, these would partially inform the extent to which the claim within the decision inference is warranted.

Appraisal

From a systemic functional perspective, Appraisal is a kind of interpersonal system that we use to express our feelings, attitudes, and stances toward other people's behaviors, things, and ideas (Martin & White, 2005). Based on this functional understanding of Appraisal, Appraisal analysis refers to an analytical technique in which the Appraisal system is adopted as a framework for purposes of observing how speakers/writers express their feelings, attitudes, and stances. For example, within the field of applied linguistics, applications of Appraisal analysis have been adopted to investigate English learners' use of evaluative language (e.g., Liu & McCabe, 2018); to understand the changing reviewer positioning between screencast video and text feedback in ESL writing (e.g., Cunningham, 2018); and to examine learners' attitudes toward using an automated writing evaluation tool (e.g., Huffman, 2015). Appraisal consists of three interrelated sub-systems known as, *Attitude*, *Engagement*, and *Graduation*. Each sub-system is made up of a set of interpersonal resources, and together, they allow speakers/writers to express their positionings.

For the purpose of the present study, the discussion will focus on Attitude (see Figure 2). According to Martin and White (2005), this sub-system is concerned with the ways in which people express positive and negative emotions, judgments about people's behaviors, and evaluations of things, ideas, and processes. In Appraisal terms, attitudes related to emotions are known as *affect*; attitudes related to judgments about people's behaviors are referred to as *judgment*; and attitudes associated with evaluations of things, ideas, and processes are called *appreciation* (Martin & White, 2005). It should be noted that speakers can express affect, judgment, and appreciation in terms of nouns, verbs, adjectives, or adverbs.

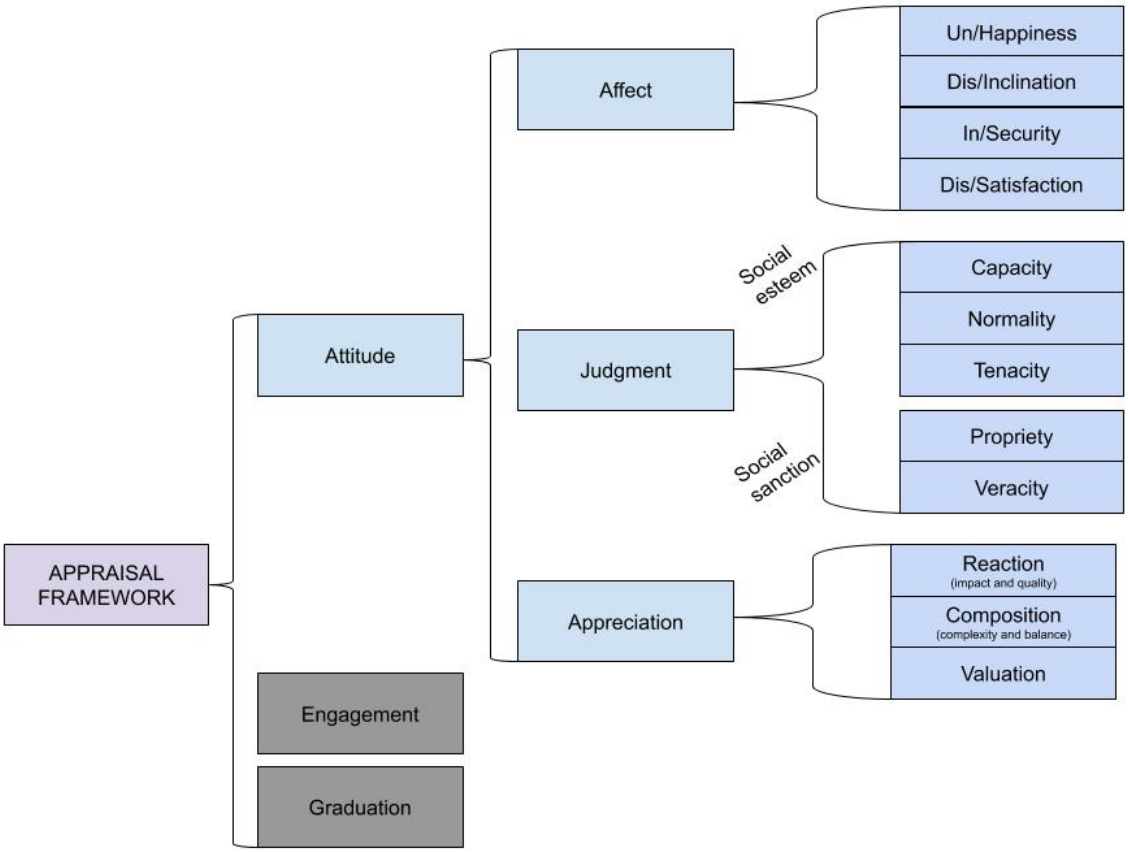
Affect, judgment, and appreciation can be further expanded into subclasses, which are useful for fine-grained linguistic analysis. The subclasses of affect are *un/happiness* (e.g., love/hate), *dis/inclination* (e.g., desire/fear), *in/security* (e.g., confident/anxious), and *dis/satisfaction* (e.g., interesting/boring). The subclasses of judgment are *social esteem* (i.e., judgments about a person not related to legality/morality) and *social sanction* (i.e., judgments about an individual related to legality/morality). Each of these subclasses can be further expanded: Social esteem can be further distinguished into three subcategories, *capacity* (e.g., capable/incapable), *normality* (e.g., normal/unusual), and *tenacity* (e.g., responsible/unreliable). On the other hand, social sanction can be characterized into two subcategories, *propriety* (e.g., good/evil) and *veracity* (e.g., honest/deceptive).

Finally, there are three subclasses for appreciation, namely, *reaction*, *composition*, and *valuation*. With the exception of valuation, reaction and composition can be further expanded into subcategories as seen with judgment. Reaction is concerned with impact (e.g., amazing/terrible) and quality (e.g., good/bad), and composition is concerned with balance (e.g., orderly/disorderly) and complexity (e.g., easy/difficult) of things, ideas, or processes. It should be noted that impact associated with Reaction is not the same as impact related to testing consequences, though speakers may express an impact-related attitude to express the impact of test use. Valuation, on the other hand, encompasses attitudes that would signal importance (e.g., significant/insignificant). A more comprehensive discussion of Attitude can be found in Martin and White (2005). Using these categories of Attitude, it is possible to identify a wide range of affect, judgment, and appreciation resources that are expressed by writers/speakers.

Because the presence of positive and/or negative affect, judgment, and appreciation markers can be partially affected by the orientation of the reader, it is important to clearly state

one's reading position before starting the analysis (Martin & White, 2005). Generally, one could take on one of three positionings: compliant, resistant, or tactical (Martin & White, 2005). In brief, a compliant reader sympathizes with the speaker/writer's point of view; a resistant reader objects to the speaker/writer's perspectives; and a tactical reader re-interprets the speaker/writer's opinions for an entirely different reading purpose. The author chose a compliant reader positioning (see section on Author's Positioning for more detail below).

FIGURE 2
The Appraisal Framework (Martin & White, 2005)



Using Appraisal for Evaluative Purposes

While Appraisal has not been widely adopted in language testing, it is adopted as an analytical framework in a wide variety of language-related research (see Hang & Bednarek, 2020, for a bibliography of Appraisal research). This is because the framework offers a systematic yet adaptable way in which any spoken or written text can be analyzed—all the while taking into account the context in which the discourse occurred—in order to identify attitudes and/or positionings that are both explicit and implicit. In other words, because people's feelings and opinions may not always be "spelled out" and are often affected by the situation and the audience, Appraisal is a powerful model for identifying the subtle ways in which speakers express themselves. Huffman (2015), for instance, investigated graduate students' attitudes toward an automated writing evaluation tool. It was found that positive and/or negative attitudes,

specifically affect and appreciation, were associated with the user's level of confidence and perceived degree of control. A user who felt confident and in control over the automated writing tool expressed greater positive affect and appreciation. Although Huffman did not frame her investigation within an argument-based approach to validity, through Appraisal, she was able to conclude that positive benefits existed with the automated writing tool. As shown in Huffman's study, Appraisal is a method of analysis that is helpful in examining perceptions of socially beneficial consequences.

While acknowledging that self-reports alone do not fully account for the actual impact of the assessment, this case study nevertheless sheds some new light on how students experienced and reacted to one reading placement test—an issue not previously investigated through this type of analysis. To that end, the current study aims to evaluate the test consequences of placement-test-score use by analyzing students' assessments of a placement test and ESL reading course in an Appraisal framework. The study was part of a larger validity investigation that evaluates whether the placement-test score is overall appropriate for placement decision-making. It is hoped that the findings and discussion from this study encourage future researchers to draw on students' perceptions and adopt Appraisal analysis as a means to evaluating the validity of test-score use. To evaluate the test consequence, the present study investigates the following research question: What attitudes do students express toward the placement test and the ESL reading course?

METHODOLOGY

Context of Study

The study was undertaken at a large American university where international students whose TOEFL scores were below 100 took the university's placement test, consisting of a reading, writing, and listening section, at an on-campus testing site a couple of weeks before the start of the semester. Students who failed to pass the reading section of this placement test subsequently registered for an ESL reading course and took a second placement test, labeled as a *diagnostic test*, on the first day of class. Although nominally called a diagnostic test, the purpose of this in-class test was to identify students who would actually benefit from taking the eight-week ESL reading course; students did not receive any diagnostic feedback other than a pass/fail notice. While no more than one or two students typically pass the test, the intention of administering this second test was to ensure that placement decisions were, to the extent possible, made in students' best interest. For students who passed the test, the requirement to take the course was waived, while students who failed the test remained in the course. The current study focuses on evaluating the test consequences of this follow-up, in-class reading test which, for the sake of brevity, will be referred to as *placement test*.

Although an official blueprint stating the test's purpose could not be accessed, based on the learning objectives of the course in which it was implemented, it can be inferred that the test was intended for placing students who do not retain strong academic reading skills (such as identifying main ideas or causes and effects) into the ESL reading course. Based on the test results, 57 students failed the in-class test and therefore took the eight-week-long ESL reading course, which taught a range of reading skills including summarizing, finding main ideas, annotating and highlighting, and academic vocabulary.

Participants

After obtaining IRB approval, consent forms were distributed in all six course sections to recruit students who would share their test-score information, grade reports, and coursework, and would participate in an oral interview at the end of the semester. Of the 57 total ESL students who took the placement test and were placed into the reading course, 30 students agreed to participate in the study. The majority of the students majored in a STEM or business-related field, with none pursuing arts or humanities. Over half of the sampled students spoke Mandarin, while other first languages, such as Farsi and Tamil, were represented in significantly smaller percentages. In terms of gender and educational level, there were about four times as many male students as female students and about twice as many undergraduate as graduate students. Out of a total score of 100, the average score was 54 and the median was 56. The test scores ranged from 31 to 68 (IQR: 48–63, s.d.: 10.51), indicating that there was a very wide range of performances among the 30 test-takers. However, most students eventually passed the ESL course, earning final grades in the B range. The 30 students' test-score information, grade reports, and coursework were used to investigate the scoring inference, generalization inference, and the extrapolation inference; however, as the purpose of the current study is to examine the decision inference, these will not be presented here.

Out of the 30 students who agreed to participate in the study, eight agreed to participate in the interview. Although the rate of interview participation was low, the sample represents the central tendency, where over half of the test-takers scored between 48 and 63. As can be seen in Table 1, all eight students pursued a STEM or a business-related major. Their placement-test scores ranged between 49 and 67, and final grades were mostly in the B range. Along with Mandarin, other first languages (L1s) such as Farsi, Tamil, Malay, and Kazakh were represented among these participants. The ratio of males to females, and undergraduates (U) to graduates (G), were both four to four.

TABLE 1
Characteristics of Interviewees

Student	Gender	Education level	L1	Major	Score on the placement test (0 –100)	Final grade in the course (F–A)
1	F	U	Farsi	Computer engineering	61	B-
2	M	U	Mandarin	Software engineering	67	D
3	M	G	Kazakh	Statistics	60	B
4	M	G	Mandarin	Civil engineering	63	B
5	F	U	Malay	Business	60	B+
6	F	G	Tamil	Computer science	49	B+
7	M	G	Farsi	Computer	63	C+

8	F	U	Mandarin	science Agronomy	63	B+

Instrument

The placement test measured students' academic reading skills and strategies. There were two reading passages, each having a word length of 700. Using the Flesch-Kincaid reading indices, the first passage scored 31.7, which is equivalent to a college-level reading passage; the second passage scored 44.1, which is equivalent to an 11–12th grade level passage. The first reading passage discussed globalization, and the second discussed sociolinguistic rules; both were argumentative text types in which the author argued for/against an issue. For the first passage, there were 15 items in total; the first five were matching items where students had to select an appropriate summary heading for each of the five paragraphs. The following six items were multiple choice items that prompted students to find details or make inferences about the text. The next five items tested students' vocabulary knowledge, requiring students to select the best synonyms for words that appeared in the passage. For the second passage, there were 12 items in total; the first three were true/false items where students had to make inferences about the text; the next four items were multiple choice questions asking about details or inferences; and the last five items were vocabulary knowledge questions, similar in form to the first reading passage.

Individual semi-structured interviews were conducted to draw students' perceptions of the test and the ESL course. There were 21 interview questions in total, half of which inquired about students' perceptions of the test, and the other half of which inquired about the ESL reading course (see Table 2). Follow-up questions were asked where more explanation was needed. On average, the word count of each transcribed interview (excluding the researcher's questions and comments) was approximately 3870. The total word count that was analyzed for the purpose of this study was 30958.

TABLE 2
Interview Questions

Theme	Questions
Perceptions of the placement test	1. How was your experience with taking the placement test?
	2. What parts of the test did you like? Why?
	3. What parts of the test did you not like? Why?
	4. Comparing the readings that were done in the course, how were the reading passages that appeared on the placement test?
	5. What reading skills were focused on the placement test?
	6. What reading skills were not focused on the placement test?
	7. What skills and strategies did you use to complete the test?

	8. Why do you think you received the score that you got on your placement test? 9. Do you think the test was an appropriate test for making decisions about who should take the ESL reading course? Why or why not? 10. What would you include or take away from the placement test to make it better? Why? 11. If you retook the placement test, do you think you would have passed it? Why or why not?
Perceptions of the ESL reading course	1. How was your experience with taking the ESL reading course? Why? 2. What parts of the course did you like? Why? 3. What parts of the course did you not like? Why? 4. How were the readings that were done in the course? 5. What reading skills were taught in the ESL reading course? 6. What reading skills were not taught in the ESL reading course? 7. What skills and strategies did you use to complete the reading questions? 8. Why do you think you received the grades that you got in your ESL reading course? 9. Do you think the course taught you the necessary reading skills that you need in order to do well in your other courses? Why or why not? 10. What would you include or take away from the reading course to make it better? Why?

Procedure for Data Collection and Analysis

Students were contacted by email after the eight-week course was completed in March and interviews were conducted within the month after the completion of the course. All interviews were conducted one-on-one in a quiet office room. At the interview, students' placement tests and course materials and assignments were shared to aid their memories. Each interview lasted between 45 minutes to one hour and was audio-recorded. The interviews were transcribed manually in NVivo 12. Transcripts were subsequently shared with the interviewees for member-checking. Any identifiable information about the student was removed during the transcription process and replaced with a non-identifiable three-digit code for analysis. The transcribed data was read multiple times and memoed within the same NVivo 12 software. The data were coded according to Martin and White's (2005) Appraisal framework. Instances were categorized into the main categories of Appraisal analysis—affect, judgment, and appreciation—and then into positive/negative subclasses. Consultations were received from a faculty and a Ph.D. colleague experienced with Appraisal analysis. Ten percent of the data were coded independently by a trained second coder for inter-coder reliability. Using Cohen's kappa, the intercoder reliability was 0.87 with a 95% confidence interval of (0.82, 0.91), which is

considered a strong agreement beyond chance between two coders (McHugh, 2012).

Author's Positioning

Martin and White (2005) maintain that it is important for readers to be aware of their subjective positioning, and to account for this in their studies. The author of this study maintains a compliant reader positioning: As a native speaker of English and an ESL instructor who has taught and administered language assessments for seven years, she is sympathetic to students' concerns with taking language assessments and ESL courses. While the attitudes expressed by the ESL students are analyzed from this subjective positioning, to investigate both positive and/or negative aspects of test-score use, the interview questions were framed such that students would consider expressing both positive and negative affect, judgment, and appreciation about the test and the ESL reading course.

FINDINGS

Frequencies and Proportions

Although students expressed both positive and negative attitudes toward the test and the ESL reading course, these attitudes were expressed in various degrees (see Table 3). Of the 1009 identified attitudes, 312 were related to the test and 697 were associated with the course. This variability was due to differences in the size of the evaluated materials and times of exposure: For the test, students evaluated the content and tasks of a single placement test that was taken only once, whereas for the course, students assessed several course reading passages and assignments that were completed over the eight-week period. Second, when we compare the raw counts, students indeed expressed more attitudes toward the course than the test; however, when we separate attitudes by test and course, the attitudinal instances captured from this single placement test were relatively substantial compared to the course materials.

For the test, there was almost an even balance of positive (52.88%) and negative attitudes (47.12%). Of the positive attitudes associated with the test, appreciation, particularly +composition, was expressed almost four times more frequently than either affect or judgment. Composition is a type of Attitude marker that is generally used to express how well a thing, idea, or process is constructed, and as previously mentioned (see Appraisal section above), it can relate to balance or complexity. In this study, positive/negative composition was specifically attributed to the balance or complexity of the test and the course materials. For example, if a course material was perceived as easy, it would be considered a +composition. On the other hand, if a test was perceived to be difficult, it was treated as a -composition. Similar to +composition, of all the expressed negative attitudes toward the test, -composition was observed almost three times more regularly than categories of affect or judgment. Because the interview questions prompted students to provide assessments of the test and the test-taking situation, it was expected that appreciation would appear in relatively great numbers.

When it came to evaluating the course, there were twice as many positive attitudes (66.86%) as there were negative attitudes (33.14%). Similar to the test, appreciation appeared in greater frequency than affect or judgment. +/-Composition resources were still the highest-occurring attitudes expressed by the ESL students, and relatively high percentages of +/-capacity

and +inclination were also expressed. Overall, the figures show that students expressed more positive than negative attitudes toward the course, whereas they expressed nearly equal proportions of positive and negative attitudes toward the test.

In the next three sections, findings from the qualitative analysis are presented and discussed. These sections are followed by a discussion of the Appraisal analysis findings with respect to the test consequences.

TABLE 3
Frequencies and Percentages of Attitudes

Category (n)	Polarity (n, %)	Resource (n)	Sub-resource (n)
Test (312)	Positive (165, 52.88%)	+*appreciation* (111)	**+comp (73)** +react (20) +val (18)
		+*affect* (29)	+hap (1) **+incl (20)** +sat (1) +sec (7)
		+*judgment* (25)	**+cap (24)** +norm (1) +prop (0) +ten (0) +ver (0)
	Negative (147, 47.12%)	-*appreciation* (82)	**-comp (71)** -react (8) -val (3)
		-*affect* (29)	-hap (2) -incl (0) -sat (4) **-sec (23)**
		-*judgment* (36)	**-cap (34)** -norm (0) -prop (1) -ten (1) -ver (0)
Course (697)	Positive (466, 66.86%)	+*appreciation* (321)	**+comp (111)** +react (71) **+val (139)**
		+*affect* (89)	+hap (7) **+incl (65)** +sat (7) +sec (10)
		+*judgment* (56)	**+cap (52)** +norm (0) +prop (0) +ten (4) +ver (0)
	Negative (231, 33.14%)	-*appreciation* (147)	**-comp (111)** -react (15) -val (21)

	-affect (23)	-hap (4) -incl (6) -sat (5) -sec (8)
	-judgment (61)	**-cap (55)** -norm (1) -prop (1) -ten (4) -ver (0)
Total	1009	

Students' Appreciation Toward the Test and the Course

All eight students expressed a high degree of +/-composition when it came to evaluating the test and the test-taking situation. According to Martin and White (2005), composition is related to an evaluator's perception or the way in which they see "balance and complexity" (p. 56) in objects and processes. +Composition was attributed to the format of the test. Students recounted having taken similar reading tests in their home countries and while preparing for the TOEFL exam. The vocabulary section was found to be especially easy for three out of four graduate students as they had learned many of the academic words in their undergraduate studies. For instance, selecting the best definitions for *advocates, integration, perceived, ongoing,* and *empirical* were considered easy because these were encountered in their undergraduate studies. Although scores on the vocabulary section did not vary much between the four graduate and four undergraduate students, graduate students expressed more confidence in their vocabulary knowledge than undergraduate students.

On the other hand, being familiar with a test structure (i.e., reading a passage and then answering multiple-choice questions) was not the same as perceiving that the test was manageable. Consequently, an equally substantial number of -composition items were observed during the interview. In various degrees, all eight interviewees felt that completing the entire test (which consisted of two 700-word reading passages and 27 reading questions) within 40 minutes was "too much," especially on the first day of class in which the test was unannounced. The level of difficulty expressed by students, however, did not match with the students' placement test scores (see Table 1). It may be that some students had greater difficulty recalling their precise experience (e.g., Student 2 and 7) or that some students were naturally more expressive in detailing their situation compared to others (e.g., Student 6 and 8). Although test scores did not exactly match with the level of difficulty expressed, it is important to note that, in general, students encountered difficulty in part due to having to take a reading test for which they had not prepared.

Another shared concern was that the reading passages of the test were particularly long and challenging:

> "Because it's **longer[-comp]**. So I know that the diagnostic test must be **harder[-comp]** than the level of the ESL reading course because it's going to indicate whether you're enrolled in this class, right or not? So it's supposed to be **harder[-comp]** than our level, but at the same time, I think it's **just too much[-comp]**. . ." (Student 1)

> "Yeah because I think if you want to read this one, for my, the reading speed is **not enough[-comp]** in 40 minutes. When I'm in China, I do some reading, I will circle the whole construction. I don't have that time in this 40 minutes so it's **hard[-comp]** for me to understand this article." (Student 2)

While the test questions and reading passages were considered difficult by six out of eight participants, the students held somewhat different opinions about the ESL course materials, where they expressed nearly twice as many +composition as -composition. -Composition was mostly associated with skills and tasks that students had never encountered prior to enrolling in the ESL reading course. All four undergraduate students, for example, found outlining and summarizing challenging because they had never been exposed to these tasks in their home countries.

However, students expressed more +composition than -composition regarding the ESL program. Graduate students, for instance, found outlining and summarizing relatively easy and straightforward because they had practiced these skills in their undergraduate studies. In addition, +composition was attributed to textbook passages and reading questions, which were perceived to be short and comprehensible, containing few technical words. If technical words were present, the definitions of these words were defined in the margins, which made processing texts much more accessible. Below, Student 2 shared that the reading passages from the textbook aided his comprehension:

> "I can **easy[+comp]** to know what evidence exactly and what the idea is." (Student 2)

Students expressed not only a high degree of +composition but also relatively high degrees of +valuation and +reaction. When course tasks were found to be important, useful, and relevant to their learning, students frequently used valuation, or attitudes expressing importance (Martin & White, 2005). Many agreed that taking the ESL reading course was a valuable educational experience, helping them increase their reading skills and vocabulary knowledge, which they were able to apply to their major-related courses:

> "I'm grateful I take this class because **it is very useful[+val]** with my other classes. Especially when, for my business level, I need to read textbooks all the time, where there are longer passages and stuff and hard vocabulary. Since I took this class, **it really helped[+val]** with my business level class." (Student 8)

+Reaction was used to refer to content and tasks which students considered interesting and those which created a "good" impression. Reaction is triggered by the evaluator's affection to matters and it is strongly connected to impact and desire (Martin & White, 2005). Course tasks, such as annotating and highlighting and completing article assignments, were expressed as +reaction. Self-selected articles and reading passages from the textbook were also found to be interesting because they related to students' personal interests:

> "We can choose our own article instead of the teacher choosing for us. Because if she chooses it for us, maybe, I don't like it but since I can choose my own and relate to my major, I think I **like[+react]** it. And it's easy to understand too." (Student 8)

+Reaction, however, was not limited to course tasks and reading passages. Student 5, for example, recounted that taking the ESL reading program was a great introduction to adjusting to an American university as a first-year international student:

"The course has to match at our levels and **that's why I like it[+react]**. That's the beginning. What I mean is, you have to be the leader, to open the English door, to know, to take some very first steps, to learn how to read an English article." (Student 5)

In summary, the present findings associated with appreciation are similar to what was found by Cheng et al. (2007) and Han and Cheng (2011) when they interviewed students about their perceptions of the OSSLT examination. In these studies, L2 students expressed that the test tasks were hard and difficult and that the reading passages of the OSSLT were complex. Although students in Han and Cheng's (2011) study expressed many more negative than positive views toward both the test and the ESL course, these differences may be due to different testing situations and variances in the ESL programs. In the case of Han and Cheng's (2011) study, ESL students were high school students who had lived in Canada for most of their lives. The OSSLT exam was a well-known region-wide high-stakes exam and the consequences for failing the test were made apparent to the entire school: L2 learners would take additional ESL remedial courses prior to earning their diplomas. Being labeled an English language learner had also isolated them from fully participating in their mainstream community (Han & Cheng, 2011). Conversely, the stakes for failing to pass the reading placement test considered in the current study would result in taking one additional ESL reading course that was only half-a-semester long. Another explanation for the difference between this and prior studies may be that, as the interviews in Han and Cheng (2011) were conducted immediately after the test and informally throughout the school year, the authors found greater observations of negative perceptions from their L2 students compared to a one-time formal interview in this study. Nonetheless, the fact that some high school L2 students found parts of the course, such as learning grammar, good and helpful, is similar to what was expressed by the eight ESL students toward the ESL reading course in this study.

Students' Affect Toward the Test and the Course

For both the course and the test, the most frequently occurring affect was +inclination, as expressed by all eight students. Inclination is an irrealis affect or an emotion that responds to something that has not yet happened, and it would involve expressing a desire or fear (Martin & White, 2005). +Inclination was highly expressed when students were asked to respond to question 10 in the interview: "What would you include or take away from the placement test/course to make it better? Why?" In response, the majority of students wished that they had been provided an opportunity early on to prepare for the test. For example, Student 4 suggested the idea of providing practice test questions on the test webpage so that students could practice ahead of time:

> "**I think we need more[+incl, implicit]** practice exams on the website. Before having this class, we didn't know what we were taking. I mean, what kind of questions we're going to answer." (Student 4)

Here, Student 4 suggested that providing practice exam questions on the university website would have enabled students to become better prepared for the test. Other suggestions included increasing test time, limiting the number of questions, providing a warm-up activity, and informing students about the placement test long before the first day of class. Students also expressed +inclination with the ESL reading course. They felt that, as the course was only eight weeks long, providing reviews and checking in with individual students could have further enhanced their learning experiences:

> "I think [for] one unit lecture, **you can ask[+incl, implicit]** the students, after we do the annotating and highlight, **you can ask[+incl, implicit]** him or her to read aloud their annotating and highlight." (Student 7)

Student 7 recommended greater student participation in class as he felt that the course did not exercise much interaction among students. If there had been more student involvement, then students might have better understood how their peers practiced various reading skills, rather than having the instructor share all the solutions. -Security was another type of affect that was communicated about the test. Many students exhibited anxiety and a lack of confidence while taking the in-class placement test. When students were informed that, should they fail the test, they would be placed into an ESL reading course, examinees felt an enormous sense of fear and pressure to perform their best. As a result, as seen with Students 1 and 6, some students encountered insecurities about their test-taking abilities:

> "It happens for me in exams, especially reading, because I'm just like, **if I'm thinking that's right, it's definitely not[-sec, implicit]**, or some other feeling like that. That's what I really felt in the first day." (Student 1)

> " And in those days, I was kind of **surprised, shocked[-sec]** and I took this test." (Student 6)

Student 2 experienced -security in the form of self-doubt, constantly second-guessing her choices, whereas Student 6 experienced it as an unwelcome surprise. Another common insecurity was related to cognitive fatigue. Because students were expected to complete two 700-word reading passages and 27 multiple-choice questions in less than an hour, fatigue had inevitably affected some test-takers who were still adjusting to a new country and school system. As seen in the following two examples, Student 3 expressed exhaustion after completing the placement test as a recently-arrived student, whereas Student 6 experienced fatigue as a result of having to complete a test all the while worrying over unsettled residency issues.

"I was **exhausted[-sec]** by the end of the day because at that time I came to Iowa so that was the thing." (Student 3)

"Yeah. I have to adjust to a new environment and I have to rent a house. I'm new in America. **What is this? My mind is somewhere else[-sec]**." (Student 6)

Similar to what was observed with -composition, students' final test scores did not align exactly with the amount of -security that they expressed. However, as students' test scores fell

within a narrow range, it may be that differences would not be apparent in this case study. Despite this limitation, a major concern that was raised by the eight students is that students had encountered fatigue which would have negatively affected their overall test performances.

As shown above, test-takers' test anxiety and fatigue would be considered potential sources of CIV. Although the testing situations are different, the findings also accord with the test anxieties and tiredness observed among L2 students in Cheng et al.'s (2007) and Han and Cheng's (2011) research. According to these studies, a major factor that induced test anxieties and fatigue was that students had to complete a series of reading passages, multiple-choice questions, and open-response tasks in one 75-minute-long sitting at the start of the new semester. While the reading placement test considered in this study took only 40 minutes, participants still reported fatigue. Admittedly, the OSSLT is a more cognitively demanding test with greater stakes involved; however, both high-stakes testing situations, combined with time pressure, can inadvertently raise anxiety and fatigue in test-takers.

Students' Judgment Related to the Test and the Course

+/-Capacity far exceeded any other judgment resources, and this was observed across all eight students. This is because some of the interview questions elicited students' judgments about their own skills and strategy use, and about their test and course performances. Capacity reflects how capable an individual is deemed by an assessor (Martin & White, 2005); it can be expressed as either admiration or criticism of an individual's social esteem. Students expressed more -capacity than +capacity when it came to evaluating their test performances. They claimed that their failure to pass the exam was due to their poor reading and test-taking skills:

"I think, maybe, **I was not good at skimming and scanning[-cap]**. **I didn't develop my skills[-cap]** before taking the ESL reading course." (Student 4)

On the other hand, interviewees felt that their level of reading comprehension developed as a result of taking the ESL reading course. For example, Student 7 perceived that he was reading in English with greater ease and grasping at the gist of each body paragraph, though he still considered summarizing a challenge:

"Now I'm reading my geology book and **I am doing the summary[+cap]**. I think the speed improved for me. **I can read it very quickly and know the meaning of the paragraph[+cap]**, but for the summary, you need to get the point of this paragraph and **I think I need to take time to do it[-cap, implicit]**." (Student 7)

As can be seen, students used both +/-capacity to self-assess their current reading levels, highlighting both their strengths and areas for improvement. Students generally appreciated learning summarization and annotation skills because these were not taught in their home countries. Furthermore, acquiring these skills enabled them to apply their reading skills to their major-related studies. Student 6, for instance, noted how he continued to use the reading skills that were taught in the ESL reading course for his own research:

"**I personally use annotation and outlining[+cap, implicit]** because I always have

articles that I read." (Student 6)

The findings confirm previous research which showed that understanding and applying academic reading skills and strategies can be incredibly useful for university-level L2 students (Neumann et al., 2019) and can help build students' self-concept and confidence (Mokhtari & Sheorey, 1994; Sheorey et al., 1995). On the other hand, the results from the current study somewhat deviated from Cheng et al.'s (2007) and Han and Cheng's (2011) findings, particularly with regards to the reasons for expressed -capacity that were shared by L2 students. Unlike the high school ESL students who expressed lack of confidence in their academic courses in part due to having been isolated from the mainstream community, the undergraduate and graduate L2 students in this study attributed their lack of mastery of academic reading skills to their limited L2 reading training in their home countries. However, once they became acquainted with academic reading skills, students reported increased confidence in reading materials for their major-related courses. The findings show that attitudes expressing -capacity may be inevitable when a student has failed to pass a high stakes test. However, the upside is that negative self-perception can turn into a +capacity over time when students continue to receive proper ESL support and are encouraged to apply their learning to other areas of study.

Consequences of Placement-Test-Score Use

As seen in the previous sections, interviewees used a wide range of positive and/or negative Attitude resources to communicate their evaluations of the placement test and the ESL reading course. These evaluative choices can, in turn, inform the consequences of the placement-test-score use. Messick (1989) and Haladyna and Downing (2004) maintain that test consequences are considered problematic if they are found to be tied to CIV, such as overly complicated wording of questions, test anxiety, and fatigue. It should be noted that while difficulty of wording could be perceived as construct relevant on a reading test (i.e., the perception of the wording as difficult may be related to a test-taker's language proficiency), if the aim of the test is to determine whether students can demonstrate fundamental reading skills (e.g., identifying main ideas or causes-and-effects), it is important that verbal demands (e.g., technical vocabulary words, negatively worded statements, qualifiers) are kept to a reasonable number such that test-takers are able to demonstrate reading skills to the best of their ability, rather than having their performance be hindered by complex linguistic features (Abedi, 2006; Haladyna & Rodriguez, 2013). Although this is not to say that the test itself is biased, text anxiety, fatigue, and difficult wording may have negatively contributed to certain students' performances on the reading test—as seen with the interviewees' responses.

In this study, it was found that the wording of questions was perceived to be confusing, and the test-taking situation triggered fatigue and test anxiety in students. Although the test was only 40-minutes long, as the test was administrated on the first day of class, unannounced, the anxiety and fatigue experienced by certain incoming students who had recently arrived to the United States were inevitable. Furthermore, unlike the TOEFL, for which students have access to plenty of information and resources and time for preparation, students had very few of these before taking the placement test considered in this study. As such, the evidence does not fully support the assumption that students were generally free of anxiety and fatigue while taking the test or that the test was free of overly high verbal demands. Therefore, our first warrant, that the

test scores clearly identify which students would benefit from taking the ESL reading course, remains inconclusive.

Yet, the placement decision was obviously beneficial for these same students, who were able to gain valuable academic reading skills in the end. This confirms our second assumption that students recounted ways in which they would use reading skills for their major-related courses, and that this would bring a positive reading experience for learners. Subsequently, this supports our second warrant that the skills taught in the ESL reading course would be beneficial for test-takers who were placed into the ESL reading course.

In conclusion, the findings from the Appraisal analysis partially support the claim that the placement-test-score use brings about socially beneficial consequences for ESL test-takers. While students experienced a positive increase in their reading ability, there were also problems with the test-taking situation. Certainly, more test-consequence studies, with a larger sample of students, are needed to enhance our understanding of the extent to which the test-taking situation caused hindrance. Nevertheless, this study reveals that the perspectives drawn from these student interviewees provide a rich and nuanced picture about the test consequences at issue: Consequences were neither absolutely positive nor negative; rather, some aspects of the test were found to be more socially beneficial than other aspects of the test.

CONCLUSION

The present study investigated consequences of placement-test-score use by analyzing the evaluative language choices that ESL students used to discuss one test-taking situation and ESL-reading-course experiences. Overall, with regards to the question of what attitudes students express toward the placement test and the ESL reading course, students found that taking the ESL reading course was useful for their major-related studies, though many disagreed that the test-taking situation allowed them to demonstrate their best test performance. Based on this, we can only partially support the claim that the placement-test-score use brought about socially beneficial consequences for these ESL students.

Previous research used inductive approaches to arrive at conclusions about test consequences, concluding that score uses were more negative than positive (Cheng et al., 2007; Han & Cheng, 2011; O'Laughlin, 2011). In this study, Appraisal analysis, situated within an argument-based approach to validity, was used because (1) it provided a more systematic and nuanced approach to measuring positive/negative attitudes expressed by speakers, and (2) it allowed the researcher to account for both positive and negative test consequences. In other words, the study attempted to impartially measure consequences of test use, as is in line with current recommendations for measuring test consequences (Kane, 2006; Lane, 2014).

This article demonstrates the usefulness of applying Appraisal analysis in examining test consequences. Appraisal allows us to fine-grain our analyses on a positive/negative spectrum, giving due attention to all sides of test use. As the number of interviewed participants in this study were extremely small, future research should be aimed at collecting interview responses from a larger sample of ESL students. Furthermore, as this study primarily investigated students' perceptions, in order to strengthen the validity of the stated assumptions, it would be useful to compare students' perceptions to teachers' viewpoints as well as to analyze in detail students' individual performances on the test. Nevertheless, the study shows that Appraisal analysis, within an argument-based approach to validity framework, is a valuable alternative approach to

an inductive analysis of ESL students' perceptions. By asking questions that prompt students to consider both the intended and unintended consequences of test use, we were able to understand the subtle variations in which students expressed polarity toward the placement test and the ESL course. It is hoped that future validation research considers incorporating Appraisal analysis for language test validation purposes.

REFERENCES

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Lawrence Erlbaum Associates Publishers.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford University Press.

Cheng, L., Fox, J., & Zheng, Y. (2007). Students' accounts of the Ontario Secondary School Literacy Test: A case for validation. *The Canadian Modern Language Review, 64*(1), 69–98.

Cheng, L., & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework. *Language Assessment Quarterly, 12*(1), 50–66.

Cronbach, L. J., (1971). *Test validation*. In R.L. Thorndike (Ed.), *Educational measurement* (pp. 443–507). Washington D.C.: American Council on Education.

Cunningham., K. J. (2018). Appraisal as a framework for understanding multimodal electronic feedback: positioning and purpose in screencast video and text feedback in ESL writing. *Writing & Pedagogy, 9*(3), 457–485.

Green, A. (2012). Placement testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 164–170). Cambridge University Press.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items.* Routledge.

Han, H., & Cheng, L. (2011). Tracking the success of English language learners within the context of the Ontario Secondary School Literacy Test. *Canadian and International Education, 40*(1), 76–96.

Hang, S., & Bednarek, M. (2020). Bibliography of appraisal, stance, and evaluation. Available at: https://www.researchgate.net/publication/338541405_Bibliography_of_appraisal_2019.

Huffman, S. R. (2015). *Exploring learner perceptions of and interaction behaviors using the Research Writing Tutor for research article introduction section draft analysis* (14418). [Doctoral dissertation, Iowa State University]. Iowa State University Digital Repository.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 17–64). American Council on Education.

Lane, S. (2014). Validity evidence based on test consequences. *Psicothema, 26*(1), 127–135.

Liu, X., & McCabe, A. (2018). *Attitudinal evaluation in Chinese university students' English writing: A contrastive perspective*. Springer Singapore.

Martin, J. R., & White, P. R. R. (2005). *Language of evaluation: Appraisal in English*. Palgrave Macmillan.

McHugh, M. L. (2012). Interrater reliability: Kappa statistic. *Biochem Med, 22*(3), 276–282.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). Macmillan Publishing Co, Inc.

Mokhtari, K., & Sheorey, R. (1994). Reading habits of university ESL students at different levels of English proficiency and education. *Journal of Research in Reading, 17*(1), 46–61.

Neumann, H., Padden, N., & McDonough, K. (2019). Beyond English language proficiency scores: Understanding the academic performance of international undergraduate students during the first year of study. *Higher Education Research & Development, 38*(2), 324–338.

O'Laughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly, 8*(2), 146–160.

Shaw, S., & Crisp, V. (2011). Tracing the evolution of validity in educational measurement: past issues and contemporary challenges. *Research Matters: A Cambridge Assessment publication, 11*, 14–19.

Sheorey, R., Mokhtari, K., & Livingston, G. (1995). A comparison of native and nonnative English speaking students as college students. *The Canadian Modern Language Review, 31*(4), 661–677.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge University Press.