International Journal of Engineering Education Vol. 34, No. 4, pp. 1347–1362, 2018
Printed in Great Britain

0949-149X/91 $3.00+0.00

The Critical-Thinking Engineering Information Literacy Test (CELT): A Validation Study for Fair Use Among Diverse Students*

KERRIE A. DOUGLAS, TODD FERNANDEZ and ŞENAY PURZER
Purdue University, School of Engineering Education, Seng Liang Wang Hall, 516 Northwestern Ave, West Lafayette, IN 47907, USA. E-mail: douglask@purdue.edu, tfernan@purdue.edu, purzer@purdue.edu

MICHAEL FOSMIRE
Purdue University, Physical Sciences, Engineering and Technology Division, Libraries, Wilmeth Active Learning Center, West Lafayette, IN 47907, USA. E-mail: fosmire@purdue.edu

AMY VAN EPPS
Harvard University, Sciences and Engineering Services, Cabot Science Library, 1 Oxford St., Cambridge, MA, USA. E-mail: amy_vanepps@harvard.edu

Information literacy and lifelong learning are essential for engineers as they constantly renew and expand their knowledge and skills to keep abreast with the development of new technologies. However, the lack of validated information literacy assessments relevant for engineering students makes it difficult to determine how well those students are acquiring needed information literacy skills. We describe validity evidence for the Critical-Thinking Engineering Information Literacy Test (CELT), an instrument designed to assess students' information literacy associated with critical thinking in an engineering context. By examining psychometric properties of CELT through Rasch modeling applications, we present evidence of appropriate and fair use of CELT among first-year engineering student populations. From our analysis, we find that CELT is appropriate for use in the classroom to assess information skills associated with critical thinking among first-year engineering students, when students' experience with English language is part of their score interpretation. We discuss specific recommendations for use with students who have little experience learning in an English language environment.

Keywords: Assessment; information literacy; instrument development; Rasch model

1. Introduction

Today's graduates in engineering must constantly renew and expand their skills to meet the demands of a rapidly changing knowledge-based society. As part of that growth, they must be adept in gathering, evaluating, and using information to make evidence-based decisions. In ABET's 2016–2017 engineering program accreditation criteria, student outcome 3.i indicates that engineering graduates should be able to recognize, "the need for, and an ability to engage in life-long learning" [1]. Yet, studies examining lifelong learning among engineering students has shown little improvement in these skills over the years [2]. The proposed changes to ABET criteria highlight a specific aspect of life-long learning, "An ability to recognize the ongoing need for additional knowledge and locate, evaluate, integrate, and apply this knowledge appropriately" [3]. Both versions of ABET demand the development of engineering students as curious, persistent, life-long learners who are fluent in information gathering tools and methods. These students are also able to distinguish between robust and weak information, integrate new knowledge when tackling complex challenges, and generate knowledge through this process. Together these abilities align with information literacy (locate, evaluate, integrate, and apply information) [4] and critical thinking (analyze, synthesize, use engineering judgment, evaluate information) principles.

Despite the importance of information literacy in engineering, only a handful of studies have specifically examined engineering students' information literacy skills [5–8]. Likewise, while there are rigorously developed tests for assessing information literacy in undergraduate students such as *Standardized Assessment of Information Literacy Skills* [9], *iSkills* [10], and *Information Literacy Test* [11], there is a lack of instruments that targets information literacy in an engineering specific context. To address this need, we developed the Critical-Thinking Engineering Information Literacy Test, or CELT, which is designed to assess information literacy skills in contextualized problems to better understand how students use information when faced with an engineering-related issue or topic [8].

CELT is a scenario-based, multiple-choice instrument designed to measure competencies at the

* Accepted 17 February 2018.

intersection of information literacy and critical thinking, which we label critical-thinking information literacy. CELT includes two short technical narratives that include a combination of appropriate and inappropriate uses of information to support claims. Each scenario is followed by multiple-choice questions that probe understanding of the quality and appropriate use of information. CELT is a relatively short instrument (17 multiple-choice items) designed to be classroom friendly, while still possessing adequate technical properties to justify use.

For an assessment instrument in engineering education to be practically useful, potential users (e.g., administrators, educators, researchers, and students) should be able to readily see the relevance of the instrument to engineering degrees and professions [12]. Hence, CELT is designed to be potentially useful for program accreditation related ABET outcomes and inform future lines of large-scale research considering the role of information literacy in preparing future engineers.

1.1 Purpose of study and research questions

The purpose of this work is to study the validity of using CELT as a measure of first-year engineering students' information literacy in a technical context. Our approach to the development of CELT [13] and ongoing validation studies derive from the conceptualization of validity as an argument for interpretation and use of an instrument, based on evidence and reasoning [14–16]. In addition, the Standards for Educational and Psychological Testing [17] state that the three "cornerstones" of quality in assessment rest on evidence of reliability, validity, and fairness. The contemporary view of validity urges the need to explicitly state assumptions including issues of fairness, such as whether students understand items similarly and where the test exhibits no gender bias [18]. Similarly, Jorion, Gane, James, DiBello, and Pellegrino [19] argue for the need to explicitly identify types of evidence to substantiate usage claims for test results specific to concept inventories. While CELT is not a concept inventory, we have explicitly based our claim for CELT's use on evidence of reliability, validity, and fairness and designed validation studies to evaluate the degree of support for our usage claim. Previous CELT studies informed the design of CELT and iterative revision (see Section 2.3) [13, 20, 21]. We focus the current work on studying a version of CELT informed by previous studies, but not previously tested. Specifically, we examine evidence to substantiate the following claim: Students' scores on CELT can be used as an indication of first-year engineering students' skills in information literacy associated with critical thinking in an engineering context. To evaluate the appropriate uses of CELT, we investigated the following research questions:

- What is the appropriate scoring structure for CELT?
- To what extent do item characteristics of difficulty and discrimination vary across CELT items?
- How reliable are scores (measured by person-item fit) on CELT for indicating first-year engineering students' information literacy?
- To what extent do item characteristics vary across majority (male, experienced with English language instruction) and minority (not male, new to English language instruction) populations?

2. Literature review

2.1 Information literacy and engineering students

All students need to develop competency in information literacy. The Association of College and Research Libraries' Information Literacy Competency Standards for Higher Education [4] include the ability to recognize a need for information, locate the needed information, evaluate the quality of the information and source, and then effectively apply it. Engineering educators have argued that similar behaviors are associated with life-long learning among engineers: being able to demonstrate reading, writing, listening, and speaking skills; an awareness of what needs to be learned; following a learning plan; identifying, retrieving, and organizing information; demonstrating critical thinking skills; and reflecting on their understanding [22]. In an information-intensive discipline like engineering, continuous learning is a necessity, and effective information gathering is critical for continued success [23].

Within the information literacy standards, the components related to evaluation and application closely align with aspects of critical thinking theory. Glaser [24] states that critical thinking "requires [the] ability to recognize problems . . . gather and marshal pertinent information, to recognize unstated assumptions and values . . . to appraise evidence and evaluate arguments . . . to draw warranted conclusions and generalizations . . . and to render accurate judgments about specific things and qualities in everyday life" (p. 6). Paul and Elder [25] similarly describe that a "well-cultivated critical thinker: raises vital questions and problems, formulating them clearly and precisely; gathers and assesses relevant information . . . effectively comes to well-reasoned conclusions and solutions, testing them against relevant criteria and standards . . ." (p. 4).

The importance of information literacy is especially evident in design practices. The design skills

associated with gathering and using information for decision-making distinguish experts from novices [23, 26, 27]. Fosmire and Radcliffe [28] highlight the role of information literacy in engineering design through their Information-Rich Engineering Design (I-RED) model as they state, ''design is a learning activity whereby existing information is consumed and new information is created'' (p. 3).

A handful of studies have been conducted to study engineering students' information literacy skills [29–31]. Denick, Bhatt, and Layton [32] performed a citation analysis of engineering students' design assignments. They found a heavy reliance on the use of non-technical websites, incorrect or incomplete citations, and a lack of use of reference materials to find technical information. Wertz et al. [21] found similar results with regards to evaluation and documentation of information resources but also added that students were fairly reasonable in drawing conclusions or inferences from the information they consulted. Similarly, Atman et al. [27] studied engineering students' information-gathering skills when completing design tasks. Their study showed differences in students' information-gathering behaviors as compared to experts in the field. Younker and McKenna [33] found similar results in their investigation of how engineering students used information to support their design decisions by examining student reports. Their study showed that students mostly relied on self-knowledge and assumptions rather than information obtained through external sources.

More recent studies targeted the development of models and protocols for assessing information literacy in engineering. One such protocol, Info-SEAD: Seeking, Evaluation, Application, and Documentation [7, 34], provides operational definitions for each sub-component of the model. Among these four components Seeking and Documenting are skills that are more procedural and involve knowledge of appropriate search and documentation practices. The evaluation and application components, however, significantly overlap with critical thinking. Evaluation refers to students' ability to critically assess information and information sources using appropriate criteria and determine whether they provide trustworthy information. Application refers to students' ability to extract information and apply it appropriately to a problem they are facing. Application includes reading a document critically to determine what information is relevant to the problem at hand, understanding the meaning of the information, and utilizing that information to build an argument advocating for a solution to their problem. Vitally, application also involves the identification of information gaps, i.e., what is still unknown about the problem.

2.2 Assessment of information literacy

There are a number of assessment instruments designed to measure aspects of information literacy, although none of them are adequate for classroom use targeting information literacy in a context familiar to engineering students. Wertz et al. [7] conducted a comprehensive and still up-to-date review of published instruments designed to assess different aspects of information literacy. The majority of these instruments assess general information literacy skills [9, 11]. Other instruments target self-directed learning readiness, [35–37], which are designed to assess attitudes and confidence toward independent learning.

There is still a need for instruments that can effectively and efficiently measure information literacy in a context that is applicable to engineering students. As described above, CELT is designed to meet this gap, as a test that is easy to administer and score, allowing more timely feedback, that will directly address information literacy skills critical for engineering students.

2.3 CELT construction and revisions

The Critical-Thinking Engineering Literacy Test (CELT)'s development started in 2010 and involved numerous studies following both classical test theory and modern test theory [13]. In total, we have created six versions reflecting the long and iterative process of assessment instrument development [38]. We used each study to formatively evaluate CELT's ability to assess critical-thinking components of engineering students' information literacy and continuously refined the test in an effort to increase the quality of the tool. At each revision, our team used both qualitative and quantitative information based on recommendations from Devellis [39] and Haladyna & Rodriguez [40].

Early versions of CELT conceptualized information literacy broadly, with learning objectives related to seeking, evaluating, applying, and documenting information. Poor internal consistency, KR-20 $\alpha = 0.39$ [13] suggested that we needed to be more narrow in scope or substantially increase the length of CELT. We considered the construct and scope of CELT's assessment based on rationale, theory and empirical evidence. Our intention was for CELT to be readily useful for classroom assessment. Therefore, a decision was made to narrow the scope of CELT to focus only on evaluating and applying information. Our decision is in alignment with Boone and Scantlebury [41], who suggest a tight focus when determining sub-constructs for newly developed instruments. Hence, the framing and operational definition of critical-thinking information literacy was a crucial step in the develop-

ment of CELT to clearly identify the construct to be measured and the extent to which it would be measured [42].

We conceptualized the critical-thinking component of engineering information literacy as a higher order construct consisting of evaluation and application concepts. We determined that the scope of CELT would span these critical-thinking components of information literacy using technical content appropriate to first-year engineering students. At different stages during CELT's development, we gathered external feedback from experts in information literacy and engineering education. During development, we also included open-ended items asking students to provide reasoning on their selected responses and then evaluated whether the items appeared to be measuring the learning objectives [13].

To evaluate the argument that CELT measures skills related to critical thinking, we tested [43] the hypothesis that a measure of informational literacy would be significantly correlated to measures of critical thinking [44] by examining convergent aspects of validity through correlation with the Critical Thinking Assessment Test (CAT).

We revised and tested items and distractors to improve internal consistency and present in this paper the blueprint for the instrument in its current form [45].

The current version of CELT addresses nine learning objectives. These objectives are shown in Table 1, along with the corresponding items in CELT. Experts in information literacy and critical thinking (i.e., librarians, researchers and educators) and assessment (i.e., psychometricians) reviewed the learning objectives and associated test items for face validity.

The final version of CELT, tested in this study, is composed of 17 items distributed between two scenarios. The first scenario contains a memo written to a University Residence Hall Director by students who want to improve energy efficiency and sustainability of their dorm buildings. The second scenario consists of a letter to the editor, from a group concerned about the safety of consuming genetically engineered food. Each scenario is followed by several multiple-choice items. Each scenario presented data, arguments, and a set of recommendations.

3. Methods

In the following sections, we outline our approach for validating CELT by describing the participants, data collection and data preparation process. We then discuss two assumptions of Rasch analysis as well as tests we carried out to evaluate any violations of those assumptions prior to our primary analysis. Finally, we detail the methods of the Rasch model and measurement invariance across groups.

3.1 Participants and data collection

We administered CELT to engineering students enrolled in a first-year design course in the fall semester of 2014. All instructors assigned CELT as part of a course activity on information literacy and instructed students that they would award grade points based only on completion of CELT, rather than performance on the assessment. Students completed the instrument using an online survey tool. In total, 1225 students (19% female, 74% male, and 7% who did not identify a gender) accessed CELT. Of these respondents, 77% indicated English as the language of instruction at their previous educational institution.

3.2 Item scoring and data cleaning

We followed a series of steps to clean the data and prepare the data set for analysis. First, we converted students' test item responses from raw multiple-choice responses (i.e., A, B, C, or D) to scored dichotomous values (i.e., 0 or 1) indicating an incorrect or a correct response to each item. Additionally, we deleted cases of students who did not answer all questions in CELT (34) resulting in 1191 complete responses.

Second, we reviewed the 1191 completed responses for patterns of 'careless' response behavior to remove a source of construct irrelevant variance [46, 47] and ensure a more appropriate

Table 1. CELT Blueprint

CELT Learning Objectives	Test Items
1. When provided with a passage, student identifies information missing but needed to complete an argument.	1, 10
2. When given a passage, student identifies assertions that refute an argument.	2, 11
3. When given a brief passage with multiple facets, student summarizes information.	3, 12
4. When given a chart or table, student interprets quantities.	4, 13
5. When provided with an argument statement, student identifies the information that is relevant to the argument.	5, 14
6. When given a selection of statements, student identifies the ones that are supported with credible citations.	6, 15
7. When given a document, student identifies its purpose (persuade, inform, entertain).	7, 16
8. Student determines the anticipated audience for which a provided document was written.	8, 17
9. When given a selection of sources, student identifies credible ones.	9

sample for further analysis [48]. We were particularly interested in whether students completed CELT in a plausible length of time, especially given that their incentive was for completion of the instrument rather than the accuracy of answers.

An examination of the responses identified three patterns of careless behavior: (1) Overall response time (e.g., a response time below the interquartile range of response times), (2) the maximum count of single answers (e.g., a student's response to all 17 items was A), and (3) patterns in the raw answers (e.g., a repeating zigzag pattern). Fig. 1, using a 50 point moving average, shows that the relationship between the length of time students took to complete CELT and the number of correct responses is notable at shorter completion times. The left side of Fig. 1 illustrates patterns for careless answers or guessing for those who completed the test in less than eight minutes and answered very few items correctly.

In total, 183 individual students matched the first pattern, 114 matched the second pattern, and 175 matched the third pattern. We removed responses from the sample that triggered two or more of the patterns prior to analysis of item behavior following Meade and Craig's [46] suggestion to not rely on a single test for identifying careless responses. The data cleaning process led to the exclusion of 125 of the 1191 complete responses. The remaining final sample used for further analysis included 1066 students.

3.3 Validating underlying assumptions of Rasch analysis

Before moving forward with the Rasch analysis, we first checked whether the two assumptions of dimensionality and independence of items are met by CELT. Rasch analysis tests the dimensionality through the fit of the data to a 'proper' test model [49]. Failures of either assumption (dimensionality or independence of items) manifest as changes in the fit values. Therefore, separating the impact of these assumptions from other influences on item fit is useful to ascertain the appropriateness of our measurement approach. The issue of dimensionality is important not only for determining whether to use a one-dimensional or multi-dimensional Rasch model, but of practical importance scoring and interpreting results. A one-dimensional model would support scoring all CELT items together in one final score. A two-dimensional model would

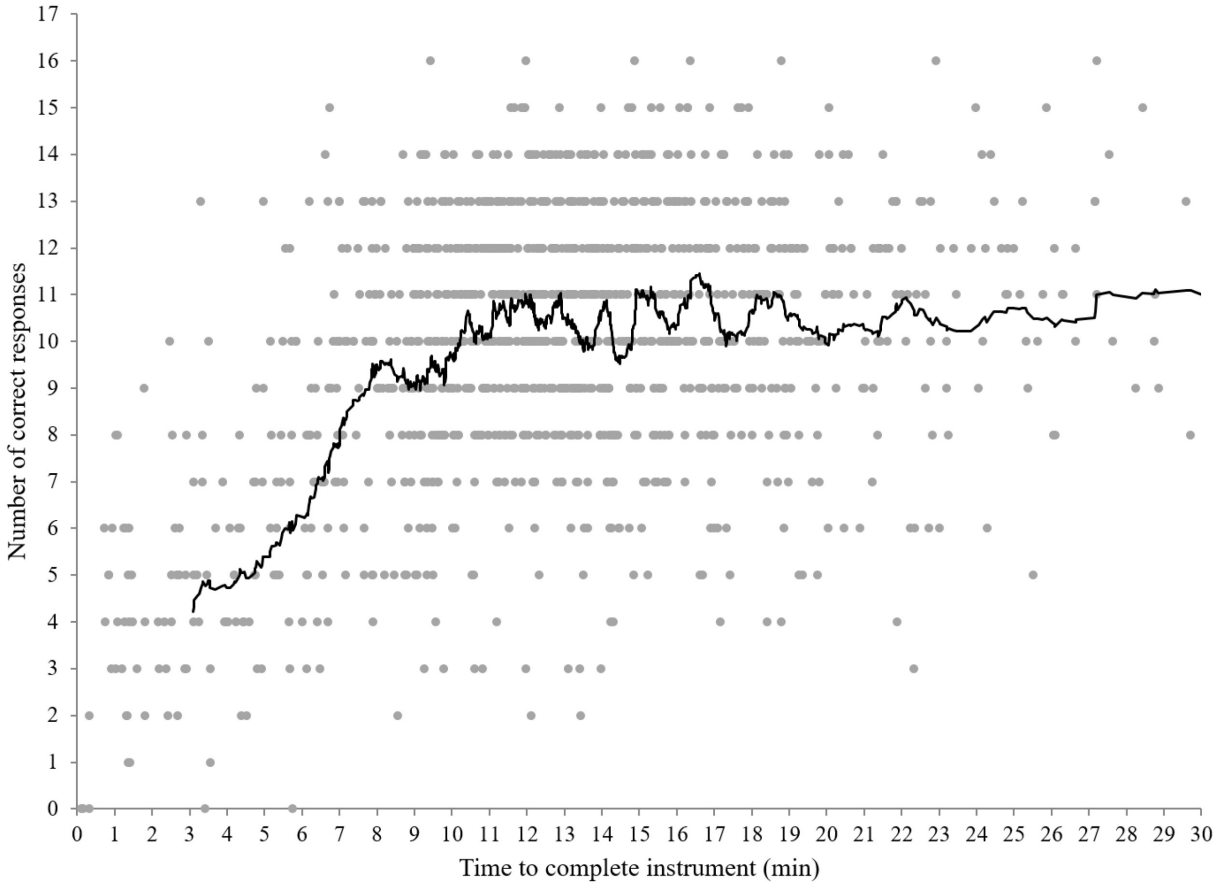

Fig. 1. Comparison of trend between time to complete CELT and number of correct responses using a 50 point (0.5%) moving average.

support scoring evaluation items and application items separately.

We designed CELT to measure evaluation and application aspects of information literacy; however, other assessments treat information literacy as a one-dimensional construct [50–52].

3.3.1 Dimensionality of CELT

Since evaluation and appropriate use of information might not be fully independent constructs, in this study we posited a series of confirmatory factor analysis models to test that hypothesis. Based on the test blue-print, we posited three potential models for the relationship between proposed constructs. Model 1 contains a single-factor construct, Information Literacy, Model 2 consists of two independent factors, Evaluation and Application, while Model 3 consists of three factors, independent Evaluation and Application factors with a higher-level Information Literacy factor.

We evaluated the models using the Lavaan package in R [53], (see Table 2) which indicates that all models are good fits for our data. The values of RMSEA and CFI were generally consistent across the models, varying only at the third decimal place. This indicates that the models are functionally equivalent descriptions of the data.

The dimensionality results indicate that treating CELT as a one factor model is the most justified scoring structure. Although the two-factor model produces similar fit indices as the one-factor model, the correlation of evaluate and apply factors is 0.90, further indicating that the two factors are not fully separate.

3.3.2 Item independence

The item independence assumption posits that a respondent's interaction with one item does not influence their responses to other items [54]. Item inter-dependence comes from many sources, including asking multiple questions about the same passage as we do in CELT. Mathematically, independence appears as a *lack* of correlation between the error residuals of items in a model. Practically, Baghaei [55] suggests that meaningful failures of the independence assumption will appear as Guttman cases or significant overfits in Rasch analysis.

To check item independence, we calculated the correlation matrix of the error residuals from the CFA check. The maximum correlation between error residuals was between items 12 and 16, which was both non-significant and small ($r(1064) = 0.056$, $p > 0.05$), suggesting independence. Following the guidance from Baghaei [55], the item fit measures in the results section further examine the independence assumption as they do with dimensionality.

3.4 Rasch analysis

We used a Rasch model to evaluate item difficulty and appropriateness in relationship to student ability [41]. The Rasch model creates a probabilistic relationship between the difficulty of a test item and the test takers' ability [49].

We performed the Rasch analysis using the Extended Rasch Modeling (eRm) package within the R statistical software package [56]. We also used eRm to create an item-person visual map, a variation of a Wright map [57], which gives a visual comparison of test item difficulty and their relationship to person ability.

Our Rasch Analysis consists of five components, described in the following subsections: measure of fit, difficulty, discrimination, reliability, and differential functioning. Each component examines a different aspect of the functioning of CELT and provides evidence for appropriate use.

3.4.1 Measures of fit in the Rasch model

We examined the infit and outfit mean square statistics to check how closely our data matches the Rasch model's expectations for variance and whether students' actual responses fit the model's expectations of their performance [49]. The outfit statistic indicates how accurate the prediction model is at the extremes of ability (i.e., does the model correctly predict a person's behavior on items that are much easier or much harder than their ability), while the infit statistic is more sensitive to

Table 2. Results of CFA dimensionality study

				RMSEA			
Model	χ^2	**df**	**p**	**Value**	**95% Conf.**	**CFI**	**BIC**
1	172	119	0.001	0.020	0.013–0.027	0.894	21496
2[1]	170	118	0.001	0.020	0.013–0.027	0.896	21501
3	170	117	0.001	0.021	0.013–0.027	0.894	21508

Note. Root Mean Error Approximation (RMSEA) Confirmatory Fit Index (CFI), and Bayesian Information Criteria (BIC) are widely used measures of fit in CFA analysis.
[1] Evaluate and apply factors have a correlation value of 0.90.

variance when person ability and item difficulty are approximately aligned [57].

Both fit criteria have an expected value of one, denoting a solution with the exact amount of randomness expected by the Rasch model. Datasets with less randomness than expected by the Rasch model result in fit statistics less than one, called an 'overfit,' while results with more randomness result in fit statistics greater than one, called an 'underfit.' Acceptable mean square fit values for regular testing are 0.7 to 1.3, with acceptable values for high stakes testing between 0.8 and 1.2 [58]. In addition to value, we examined the fit statistics for trends or correlation between item difficulty and fit statistics. Such a correlation would indicate a potential outside source of variance on the items and scale. We also compared the infit and outfit statistics for each item. The relationships between the infit and outfit values for an item indicate whether construct ability or other sources of variance, such as guessing, contribute to students' answers [49].

3.4.2 Difficulty in the Rasch model

An item-person map can aid in identifying whether the test was a comparatively easy or difficult measure of ability in respondents [49]. The item-person map uses two histograms to compare the distribution of item difficulty with the distribution of person ability [57]. The primary interest is to ensure the instrument has sufficient difficulty range to discriminate based on student ability and not so easy that a ceiling effect occurs or so difficult that few students perform well (floor effect). The criteria used to avoid a ceiling or floor effect is that less than 2% of participants exceed the floor or ceiling [59]. In addition to the extrema criteria, the difference between mean item difficulty and the mean of person ability is a useful metric for difficulty targeting. We assess this metric by standardizing the mean person ability using the standard error. Targeting is generally reported in 'standard errors' with scales that have less than one standard error between person ability and item difficulty meeting accepted criteria for 'good'.

3.4.3 Discrimination in the Rasch model

While the Rasch model sets the discrimination *parameter* of the logistic model to one for all items within the derivation [49], other methods allow for analysis of effective person-ability discrimination using a Rasch framework. At a basic level, the fit statistics indicate effective discrimination of person ability within a scale [58, 60]. Lower mean square fit values indicate an item or scale with a higher ability to discriminate (i.e., a higher likelihood that item difficulties will accurately predict correct or incorrect response patterns).

Additional evidence of CELT's discrimination ability is available from the calculated item difficulties. A wide range of difficulty values indicates CELT's ability to measure a range of person abilities. The range bounds the ability limits of test-takers that can be assessed (i.e., without a ceiling or floor effect) [49], [61]. Finally, difficulty values that are spaced approximately equally allow a more accurate sorting of person abilities within the instrument's difficulty scale [62]. Large gaps between item difficulties leave areas of the person-ability spectrum that are effectively unmeasured. Inversely, closely spaced item difficulties perform effectively redundant measurement functions and provide limited new information.

3.4.4 Person-item fit reliability in the Rasch model

While Classical Test Theory (e.g., Cronbach's alpha), measures reliability as the repeatability of raw scores or the internal consistency of items scored [63], Rasch analysis estimates a person separation reliability parameter for unidimensionality and reliability testing [64]. The formulation of person separation reliability focuses on assessing the ability of a scale as a repeatable measure of the same ability of a person on a logistic scale [65]. Appropriate values for person separation reliability are minimally 0.70 to 0.80. However, they are not unit tests and are considered holistically within the overall picture from all measures of fit [64].

3.4.5 Differential item function analysis

After analysis of the entire population, we employed a differential item function (DIF) analysis to investigate measurement bias in the individual items and the overall test. DIF analysis compares the probability that test takers from two different subgroups will correctly respond to an individual item [17]. In order to make appropriate comparison of groups, we randomly selected from the majority group the same number of respondents as are in the minority group. Our DIF analysis compared student performance subdivided by a student's reported gender and by an indicator of their experience with the English language (i.e., whether or not their previous institution primarily used English for instruction).

We used the Rasch model approach to DIF analysis as defined by Wright and Masters [57]. This approach involves three main steps: (1) Calculating the item difficulties separately for each subgroup using the Rasch model, (2) plotting subgroup difficulties against each other using a scatterplot, and (3) comparing the subgroups via linear regression. We make the comparison using two criteria that can show a difference in performance between the two subgroups, indicating a potential for construct irrelevant variance to exist. The first criterion

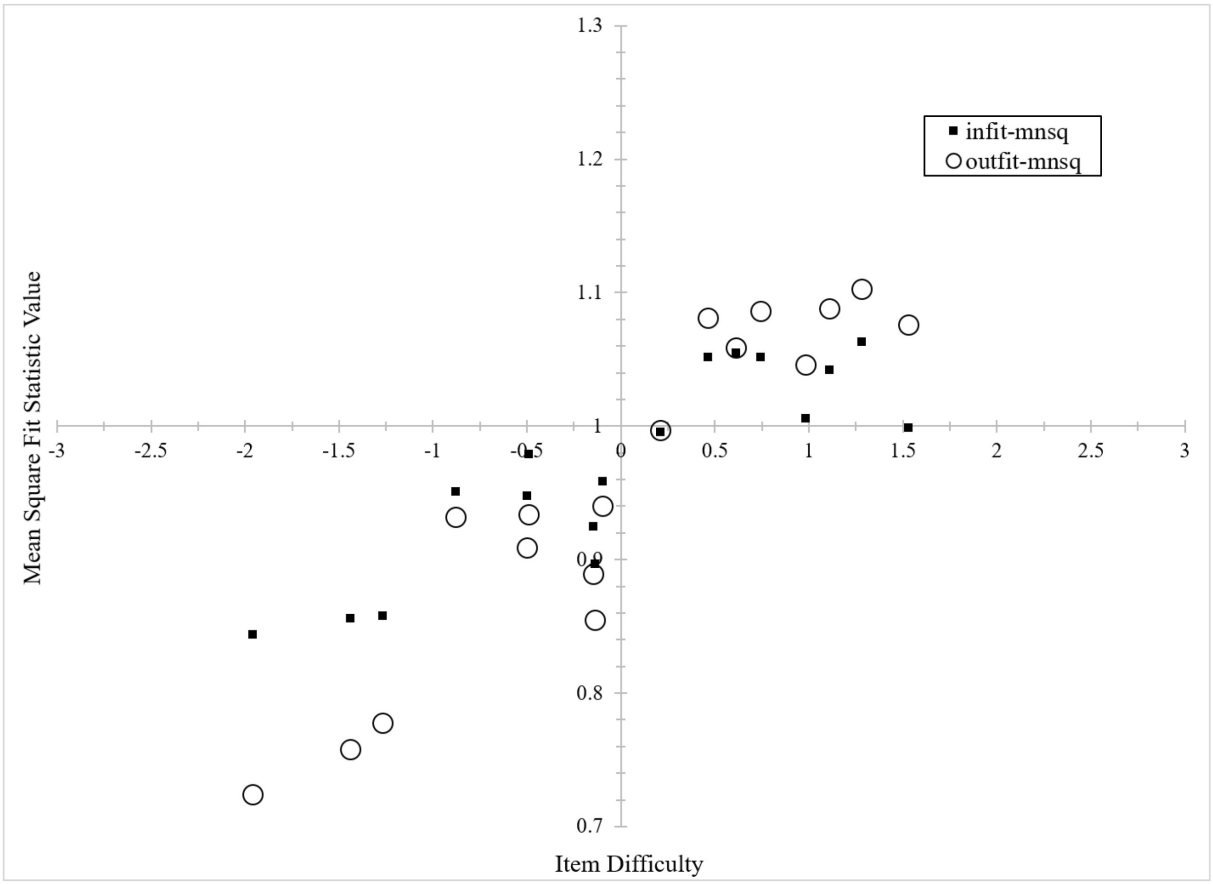

Fig. 2. Mean square fit values (y-axis) by item difficulty (x-axis) in logits.

we used examined the linear regression directly. A perfect no-DIF would result in a regression line with a slope of one, indicating that the populations share an equivalent estimation of increasing person ability as well as a similar measure of minimum respondent ability on the overall scale [49]. Second, we analyzed the performance of individual items by adding control limits to the plot. The control limits use a ±0.50 logit intercept offset with a slope identical to the initial regression. Items that fall outside of these control limits suggest potential bias, and researchers should further evaluate these items as candidates for revision or adjustment [66].

4. Results

4.1 Evaluation of Rasch model fit for CELT Scores

We evaluated CELT using a Rasch approach to difficulty, discrimination, and reliability. Table 3 presents the fit statistics and difficulty values. Item difficulty estimates are calculated in logits, where a 0 is the mean of item difficulty estimates [49]. The lower the value, the easier the test item. Conversely, higher item difficulty estimates indicate more difficult items. CELT item difficulty values range from –1.96 to 1.53, indicating a range of easier and more difficult items. However, the main concern is whether the variation between item difficulties effectively discriminates or determines students' evaluation and application skills.

All infit and 14 of the 17 outfit values are within the fit range suggested for 'high-stakes' measurement (0.80 to 1.2). The infit values range from 0.84

Table 3. Rasch results for item difficulty, outfit and infit

		Mean Square Fit	
Item #	**Item difficulty**	**Outfit**	**Infit**
1	–0.88	0.93	0.95
2	0.45	1.08	1.05
3	–0.10	0.94	0.96
4	0.74	1.09	1.05
5	–1.99	0.72	0.84
6	1.28	1.10	1.06
7	–1.43	0.76	0.86
8	0.21	1.00	1.00
9	–0.49	0.93	0.98
10	1.54	1.08	1.00
11	0.61	1.06	1.06
12	–0.15	0.89	0.93
13	–0.50	0.91	0.96
14	–1.26	0.78	0.86
15	0.98	1.05	1.01
16	–0.15	0.86	0.90
17	1.11	1.09	1.04

to 1.06 with an average of 0.97 and the outfit values from 0.72 to 1.10 with an average of 0.96. The three 'low' outfit values are still within the suggested value range for 'good' measurement. Because the values are below 1.00 rather than above, the fit suggests that there is less randomness in the responses than would be expected in a normally distributed ability set [58]. The low fit values, which occur on the three easiest items, simply suggests students are highly likely to complete these items and that they may not provide much information for discriminating student ability for most students, not that the items function poorly. Because the average fit values are very close to the target of 1.00, the model shows approximately the expected amount of variance.

The x-axis in Fig. 2 shows the item difficulty in logits. The y-axis shows the mean square fit values referenced against the expected value of one. Both the infit and outfit values are plotted. The final indicator of fit is the strong similarity between the outfit and infit statistics, also apparent in Fig. 2. The similarity of fit statistics indicates that CELT items perform similarly whether closer to or further from a student's ability. From this analysis, we have one source of evidence that CELT items are appropriately difficult and discriminate between first year engineering students with a range of information skills.

The mean ability of the participants, 0.60 logits, serves as an additional measure of appropriate difficulty. The average standard error of the person-ability calculation for all participants is 0.57; meaning that the person ability is 1.05 'standard errors' above the mean item difficulty. This result is slightly beyond the 'good' criteria suggested by Fisher [59], and indicates that CELT may be somewhat easy for some students.

We used three measures to assess the discrimination ability of CELT. First, the infit and outfit values in Table 3 also indicate that CELT has a strong initial capability for discrimination based on the fit of the data to the model [58, 60]. Second, the average gap between item difficulties is 0.22 logits, which is reasonable. However, the item-person visual shows the gaps in difficulty between items are not uniformly distributed (Fig. 3).

Two groups of items have very similar difficulties, which limits the efficiency of discretization. Items 9 (difficulty of –0.49) and 13 (difficulty of –0.50) as well as items 12 (difficulty of –0.14) and 16 (difficulty of –0.15) are 0.01 logits apart. These items will likely serve a redundant function rather than providing increased information for the determination of students' ability [62]. Conversely, the maximum difficulty gap between two items of sequential difficulty, 0.52 between items 5 and 7, is quite large. This large gap, wherein no items exist to further sort or discriminate person ability, occurs between the two easiest items on CELT. Whether this significantly affects the ability of CELT to effectively discriminate depends on how the difficulty of the instrument aligns with the difficulty of the population.

This is further shown in the item-person visual [57], Fig. 3, where student ability and item difficulty are aligned on the same logit scale. Fig. 3 allows for a visual comparison of the distribution of student ability to the distribution of item difficulty. Students are charted in a histogram according to ability on the x-axis. Next, items are placed on the chart in order from easiest to difficult. The distribution of student ability is slightly negatively skewed, with peaks around 0.25 and 1.00 logits. We found the range of item difficulties in CELT to be –1.53 to 1.96 logits giving the test a range of 3.49 logits. This compares to a person ability range of 5.51 logits. In addition, there are two bins of students that have more ability than any item is difficult. The result of this difference in range is that 93 participants (9%) have a measured ability greater than the ability of the most difficulty item (item 10), indicating students have an ability above CELT's discrimination range. Conversely, only 2 participants (less than 1%) measured ability is below the difficulty of the easiest item (item 5). In addition, the skewness indicates that there may be a potential biasing within the items, which we assess with the DIF analysis [17].

4.2 Person-item fit and reliability

We found the person separation reliability for the data to be 0.47 (comparable to the Cronbach's alpha of 0.46). This is less than the criteria we targeted (>0.7) [64]. However, in the holistic view of fit in Rasch, the other measures of fit should be taken into consideration. The results of the infit and outfit indicate that the individual item logistic models are appropriately effective at predicting performance on items. Additionally, the difficulty results noted that the mean ability is higher than the mean difficulty (i.e., an easier test). Together, those values indicate that the person separation reliability functions as validation of the previous results and an indication that separation reliability will likely decrease at higher ability levels.

4.3 Gender differential item functioning in CELT

The next step in evaluating CELT was a DIF analysis exploring correlations between students' self-reported gender and measured person ability. Both the male (n = 816) and female (n = 217) subgroups were of sufficient size using the guidelines from Paek and Wilson [67]. The analysis did not

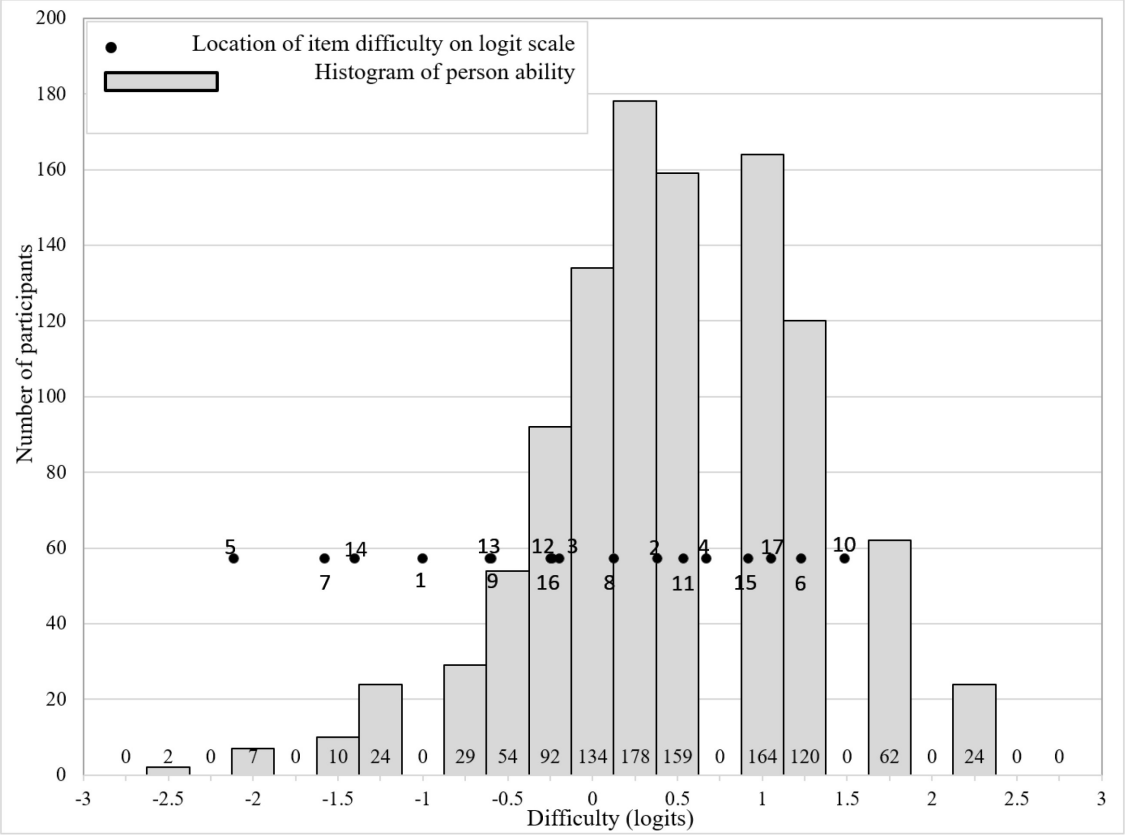

Fig. 3. Item-person visual map showing distribution of person ability and location of item difficulty.

include participants who failed to respond to the gender question or who chose the 'prefer not to answer' for the gender item ($n_{excluded}$ = 33) because the group is below the population size necessary to estimate item difficulty.

Three items, 8, 10, and 14, had difficulties that put them adjacent to the item-difficulty control limits used for identification of DIF, as shown in Fig. 4. However, given the general spread of items, we do not believe any of these items represent outliers that present a concern about differential function. The slope of the regression line, 0.98, indicates that the test was effectively equal in overall difficulty between men and women. Further, the female subpopulation item difficulty showed a slightly wider range of item difficulty than for males, (4.04 and 3.39 logits, respectively). For purposes of comparison, we checked the no-DIF result using a t-test to compare the raw scores from males (M = 10.49, SD = 2.46) and females (M = 10.41, SD = 2.28), which also showed no significant difference ($t(1031) = 0.44$, $p = 0.66$), in performance.

4.4 Differential item function based on English language knowledge in CELT

The second DIF analysis used a dichotomous categorization of participants' facility with the English language. Specifically, the analysis compared students who indicated that, at their previous institution, their primary language of instruction was English (n = 808), who we term Experienced with the English Language, or just 'Experienced' with those who indicated that English was *not* their primary language of instruction, who we term English Language Learning, or 'Learning', in accordance with LaCelle-Peterson and Rivera [65] (n = 250), with non-respondents removed (n = 8).

The regression line of item difficulties, shown in Fig. 5, has a slope of 0.62 instead of the targeted 1.00, which suggests that English Experienced students perform better than English Learning by a large margin.

In addition, we found the difficulty of item 16 to be 1.15 logits higher for English Learning compared to English Experienced students. This item is well outside of the control limits used to identify DIF items [57]. The item (''Which of the following statements made by the authors is least reliable?'') requires students to compare between different citation sources. The correct response, which was the most selected by both subpopulations, is a statement that contains no citation at all. Of the Experienced students, 76% correctly answered the item. However, only 41% of the Learning correctly

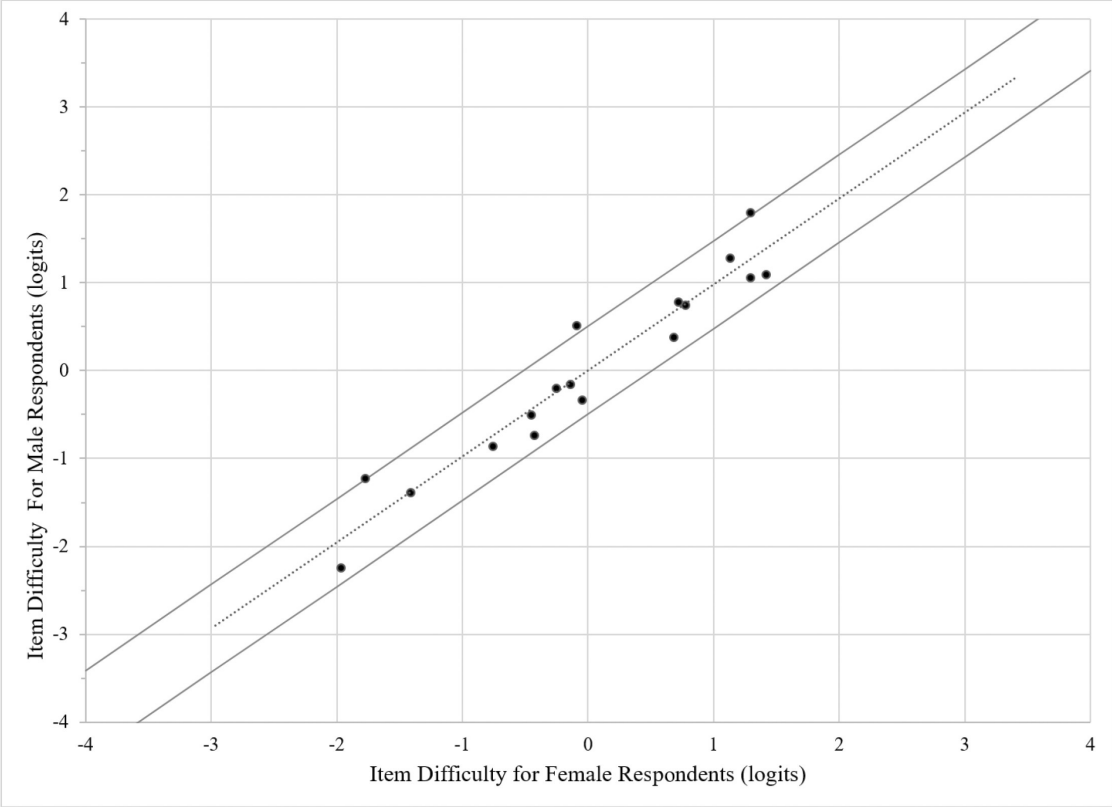

Fig. 4. Differential item function regression for Male (y-axis) and Female (x-axis).

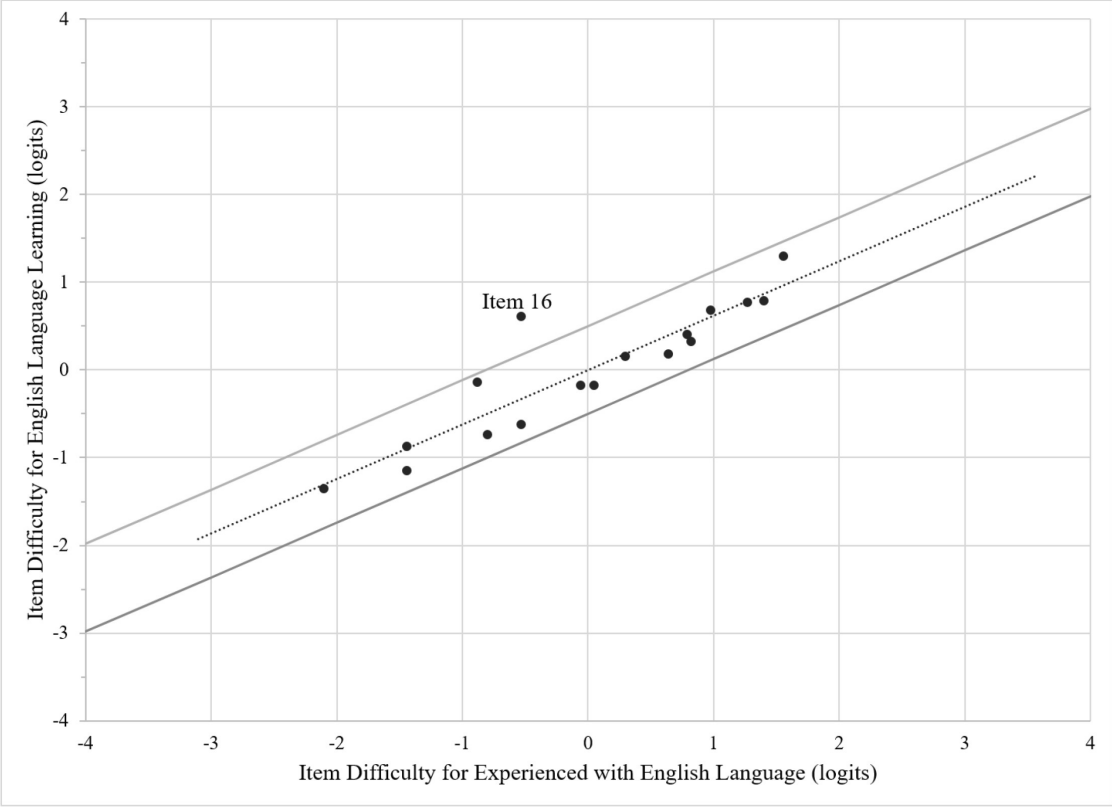

Fig. 5. Differential item function regression for Experienced with the English Language (x-axis) and English Language Learning (y-axis).

answered the item. As a further point of comparison between the subgroups: of the 82 overall participants impacted by the ceiling effect identified in the difficulty results, only 8 (9%) belonged to the Learning group even though this group makes up 24% of the overall respondent pool. As with the gender DIF, we checked the DIF result against the Experienced (M = 10.78, SD = 2.26) and Learning (M = 9.26, SD = 2.63) raw scores. The raw score difference was significant with a large effect size ($t(1031)$ = 8.86, p = <0.001, d = 0.62).

5. Discussion

In this study, we used a Rasch model to examine validity evidence for the use of CELT as an assessment instrument of first-year engineering students' evaluation and application of information skills. Overall, the results indicate CELT performs within acceptable fit values and is appropriate for use to assess first-year engineering students' evaluation and application of information skills. However, within this appropriate use, instructors and researchers must interpret students' scores as potentially biased by their experience with the English language.

The infit statistics of all items and the outfit statistics for 14 of 17 items in the test fall within the 0.8–1.2 mean square values that are suggested for high stakes testing using a Rasch analysis [58], a standard that is more rigorous than the intended applications of CELT. The three outfit statistics that fall outside this range still fall within the lower bound of the 0.7–1.3 range that is suggested as acceptable for lower stakes testing. The lower outfit statistics indicate less than expected variance in responses, meaning less test information rather than poor fit. These results demonstrate that students' performance on CELT was based on their ability and item difficulty, rather than outside sources of variance [58, 68, 69]. In addition, we found that for approximately 9% of students, CELT may be a relatively easy assessment of evaluation and application of information.

While the Standards for Educational and Psychological Testing [17] raise evidence of fairness to be on par with evidence of validity and reliability, few studies have examined assessment instrument use on minority populations. Through examination of differential item functioning based on gender, we found no meaningful differences in item or overall test behavior between males and females. This provides evidence that although CELT was administered to a largely male sampling group, as is common in engineering classes, CELT can be used fairly for both male and female students. However, we did find substantial differences in how CELT performed with English Experienced and English Learning student groups.

CELT is written in English and as a whole is much more difficult for English Learning than English Experienced students. Further, there is a large and significant difference in how difficult one item is for members of the two groups. This is not an unexpected result given that reading and understanding information is a core part of information literacy. Multiple choice tests in information literacy inherently rely on the gathering and comprehension of information, most often presented through text, which creates a potential to under report ability based on language issues [70]. In CELT, students draw answers from two passages, which are each approximately 800-words long. This test design inherently requires students to have English reading comprehension skills; indeed, comprehending information is prerequisite to evaluating and applying it. Therefore, it is important that users of CELT interpret the results of international students accordingly, with the role between comprehension and information literacy in mind. Yet, this finding does not suggest CELT is inappropriate for diverse classroom settings. Performing critical reading comprehension and using information literacy skills *in English* is necessary for success in the classrooms where we conducted this study and for U.S. engineering programs, which typically function primarily in English. In that way, the DIF presents a limitation in use but not an underlying problem in the test and may actually be useful to some instructors.

Given the DIF results, we suggest the use of CELT as a formative tool rather than a more summative assessment. Instructors could use CELT to give feedback to students and help identify techniques and instructional resources to support evaluating and applying information. In regard to English ability, a formative use of CELT would allow them opportunity for feedback and discussion around both their English comprehension and information literacy. While CELT is text-heavy, its use reinforces that engineers read and comprehend text and written documents. It is important, particularly for those new to English language, to receive feedback on their information literacy skills when the information is in the English language.

5.1 Fairness in Assessment

The significantly different performance of international students on CELT highlights the need for an ongoing and increased appreciation of fairness in engineering education assessment. We cannot simply assume that assessment instruments work equally and fairly with diverse student groups, and we cannot assume that the core constructs we seek

to measure are independent, at a fundamental level, from the models we select to measure them or from other constructs that are intertwined with the measurement model.

Information literacy is built upon the foundation of literacy and goes beyond the skill of just finding the information, to encompass the use or application of the information [71, 72]. One study has shown that differing information literacy skills frequently have roots in differing levels of ability to read and write [73]. Given this close relationship between general literacy and information literacy, any assessment targeted at information literacy must consider test takers' level of language proficiency, especially in reading. Hence, the evaluation of CELT included an examination of test scores with regards to students' experience with classroom instruction in the English language.

At a practical level, our results reinforce the need to ensure that we, as engineering educators, fairly assess all students and seek to identify bias, whether the bias is language or other characteristics not relevant to the purpose of the assessment. One of the authors of CELT is not a native English speaker and many people of diverse backgrounds read CELT during its development process. Yet still, our results indicate that CELT is significantly more difficult for non-native English speakers at both a test and item level. While most engineering programs require students to 'pass' an English language literacy test for admission (e.g., receive a TOEFL score above a certain cutoff), more work must be done to understand whether this level of proficiency is sufficient for students to be successful in the local language of instruction. As the engineering education community continues to emphasize equal opportunities for all students, it is imperative that our assessment instruments fairly assess students of diverse backgrounds as well.

5.2 Limitations and future research

Instrument validation is an ongoing process [38]. The version of CELT presented in this paper is the culmination of the results of several previous versions administered and tested at a variety of universities. This study reports analysis of data from the first semester of a first-year engineering program; hence, it is plausible that many of the non-native English speakers are relatively new to primarily English instruction. It is unknown how differently students would perform if CELT were offered in their native language or later in students' academic career with deeper exposure to English language.

Another limitation is that because of the one-dimensional model, evaluation and application items cannot be scored separately. This differentiation of scores would provide instructors more information at the sub-construct level and provide potential utility for classroom intervention. However, CELT is designed as a relatively short instrument with very few items, limiting the reliability of a purely evaluation or application factor. Scoring student performance on CELT overall may not fully represent the differences between how students evaluate and apply information. However, the CFA results demonstrate the sub-constructs are highly correlated. Briggs and Wilson [74] note that the risk of misrepresenting student ability in a one-dimensional model is decreased when constructs are moderately correlated and used for low stakes testing. Before CELT is recommended for high-stakes use of individual student consequence, an additional scenario should be added with additional evaluate and apply items and then again tested for whether evaluate and apply function as multidimensional.

While the fit statistics provide generally positive indicators of scale quality, one limitation is that 9% of the participants displayed a ceiling effect (i.e., person ability above the maximum item difficulty). Using the criteria from Fisher [59], this would rank as 'poor' as the criteria for 'good' ceiling effect is suggested to be below 2%. This ceiling effect indicates that the current version of CELT may have difficulty measuring longitudinal or post-intervention improvement in the highest performing students. Future research should consider the addition of more difficult items to reduce potential ceiling effect [75]. The additional items could also serve to support a two-factor model of CELT, where evaluation and application might be able to be scored separately.

A final limitation is that the majority of studies of CELT have been conducted at one institution. Future studies should consider generalizability of our findings with first-year engineering students attending other universities. In the ongoing process of validation, future work must also focus on extending the appropriate uses of CELT. Because CELT was normed on first year engineering students, advanced students may have skill levels that surpass the current version of CELT, expanding the ceiling effect. To improve CELT's efficiency and efficacy, future research should consider; (1) an expansion of the conceptual framework and blueprint to target higher level information literacy skills, (2) adding additional, more difficult, test items capable of discriminating top ability students, and (3) creation of alternate versions to allow the tracking of person growth and the effect of specific interventions without potential instrumentation effects [49].

6. Conclusions

We developed the CELT instrument to measure engineering students' ability to evaluate and apply information in a technical context. Based on the development of CELT and the results of our validation work, we conclude that in its current form, CELT is a useful tool for formative assessment in the classroom and as a research instrument to evaluate the information literacy ability of students.

The formative uses include identifying and giving feedback to students who do not demonstrate necessary information literacy skills, or providing targeted instruction based on individual skills. The results of CELT can provide evidence to students of their tested skill level and in which areas they are particularly strong or weak. In addition, at an aggregate level, instructors can use CELT to identify aspects of evaluating and applying information in which students may require additional intervention.

CELT shares a similar use case at a research level. Researchers may appropriately use CELT as assessment instrument in studies examining students' information literacy knowledge. We recommend that researchers examine whether there are differences in scores based on native language in their sample group. When interpreting CELT scores, researchers must understand that the test inevitably measures English language comprehension while measuring information literacy ability. Additionally, while CELT does test students' implementation of literacy skills, it does so in an environment abstracted from their actual work.

Our work also highlights the need to increase fairness of the test for non-native English speaking students. This may take the form of both item level exploration of wording and test level exploration of the reading-heavy assessment model and making analyses such as DIF as a critical process in instrument development.

Finally, at a time when programs may struggle with how to document life-long learning outcomes, CELT has potential as one indicator of how well students evaluate and apply information in an engineering-related context. As ABET accreditors seek to streamline the ability to measure student outcomes, we argue that when life-long learning is conceptualized as consisting of self-directed learning, reflective judgment, critical thinking, information literacy, domain knowledge, and continuous active learning, then assessment measures can be developed and used for evaluative purposes.

Acknowledgments—This work was made possible by a grant from the National Science Foundation (DUE-1245998) Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would also like to acknowledge Prof. Ruth Wertz's contributions to the design of CELT. CELT can be found at: https://purr.purdue.edu/publications/1118/2.

References

1. ABET, Criteria for accrediting engineering programs 2016-2017, 2014. [Online]. Available: http://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2016-2017/.
2. L. R. Lattuca, P. T. Terenzini and J. F. Volkwein, *A Study of the Impact of EC2000*, Baltimore, MD, 2006.
3. ABET, *Proposed changes to the criteria*, Baltimore, MD, 2016.
4. Association of College and Research Librarians, Information Literacy Competency Standards for Higher Education, *Association of College and Research Libraries*, 2000. [Online]. Available: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/standards/standards.pdf.
5. S. Allard, K. J. Levine, and C. Tenopir, Design engineers and technical professionals at work: Observing information usage in the workplace, *J. Am. Soc. Inf. Sci. Technol.*, **60**(3), 2009, pp. 443–454.
6. E. Collins and G. Stone, Understanding patterns of library use among undergraduate students from different disciplines, *Evid. Based Libr. Inf. Pract.*, **9**(3), 2014, pp. 51–67.
7. R. E. H. Wertz, Ş. Purzer, M. J. Fosmire and M. E. Cardella, Assessing information literacy skills demonstrated in an engineering design task, *J. Eng. Educ.*, **102**(4), 2013, pp. 577–602.
8. K. A. Douglas, A. S. Van Epps, B. Mihalec-Adkins, M. Fosmire and Ş. Purzer, A Comparison of Beginning and Advanced Engineering Students' Description of Information Skills, *Evid. Based Libr. Inf. Pract.*, **10**(2), 2015, pp. 127–143.
9. J. C. Blixrud, Project SAILS: Standardized assessment of information literacy skills, *ARL A Bimon. Rep. Res. Libr. Issues Actions*, **230**, 2003, pp. 18–19.
10. I. R. Katz, Testing information literacy in digital environments: ETS's iSkills assessment, *Inf. Technol. Libr.*, **26**(3), 2007, pp. 3–12.
11. L. Cameron, S. L. Wise and S. M. Lottridge, The development and validation of the information literacy test, *Coll. Res. Libr.*, **68**(3), 2007, pp. 229–237.
12. M. Mathison, S. Mitchell and R. Andrews, 'I don't have to argue my design—The visual speaks for itself': A case study of mediated activity in an introductory mechanical engineering course, in *Learning to argue in higher education*, S. Mitchell and R. Andrews, Eds. Portsmouth, NH: Boynton: Cook Publishers, 2000, pp. 74–84.
13. R. E. H. Wertz, M. C. Ross, S. Purzer, M. J. Fosmire and M. E. Cardella, Assessing Engineering Students' Information Literacy Skills: An Alpha Version of a Multiple-Choice Instrument, in *ASEE Annual Conference and Exposition*, 2011, p. AC2011-1273.
14. M. T. Kane, An argument-based approach to validity, *Psychol. Bull.*, **112**(3), 1992, p. 527.
15. M. T. Kane, Explicating validity, *Assess. Educ. Princ. Policy Pract.*, **23**, 2015, pp. 1–14.
16. S. Messick, Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *Am. Psychol.*, **50**(9), 1995, pp. 741–749.
17. J. W. Young, Y. So and G. J. Ockey, *Guidelines for best test development practices to ensure validity and fairness for internatinal English language proficiency assessments*, Educational Testing Service, 2013.
18. M. Kane, Validity and fairness, *Lang. Test.*, **27**(2), 2010, p. 177.
19. N. Jorion, B. D. Gane, K. James, L. Schroeder, L. V. Dibello and J. W. Pellegrino, An analytic framework for evaluating the validity of concept inventory claims, *J. Eng. Educ.*, **104**(4), 2015, pp. 454–496.
20. A. S. Van Epps, R. E. H. Wertz, M. J. Fosmire and S. Purzer, Measuring student's ability to find and use high quality

information: Developing standardized assessments, in *Professional Communication Conference (IPCC), 2013 IEEE International*, 2013, pp. 1–5.
21. R. E. H. Wertz, M. C. Ross, M. Fosmire, M. E. Cardella and S. Purzer, Do students gather information to inform design decisions? Assessment with an authentic design task in first-year engineering, in *Annual Conference and Exposition of the American Society for Engineering Education*, 2011, p. AC 2011–2776.
22. L. J. Shuman, M. Besterfield-Sacre and J. McGourty, The ABET 'professional skills'—Can they be taught? Can they be assessed?, *J. Eng. Educ.*, **94**(1), 2005, pp. 41–55.
23. K. M. Bursic and C. J. Atman, Information gathering: A critical step for quality in the design process, *Qual. Manag. J.*, **4**(4), 1997, pp. 60–75.
24. E. M. Glaser, *An Experiment in the Development of Critical Thinking*, New York, 1941.
25. R. Paul and L. Elder, The Miniature Guide to Critical Thinking—Concepts and Tools, Dillon Beach, CA, 2004.
26. M. D. Burghardt and M. Hacker, Informed design: A contemporary approach to design pedagogy, *J. Technol. Educ.*, **64**(1), 2004, pp. 6–8.
27. C. J. Atman, R. S. Adams, M. E. Cardella, J. Turns, S. Mosborg and J. Saleem, Engineering design processes: A comparison of students and expert practitioners, *J. Eng. Educ.*, **96**(4), 2007, pp. 359–379.
28. M. Fosmire and D. F. Radcliffe, *Integrating information into the engineering design process*, West Lafayette, IN: Purdue University Press, 2014.
29. B. Williams, P. Blowers and J. Goldberg, Integrating information literacy skills into engineering courses to produce lifelong learners, in *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*, 2004.
30. B. MacAlpine and M. Uddin, Integrating information literacy across the engineering design curriculum, in *American Society for Engineering Education*, 2009.
31. H. Nerz and L. Bullard, The literate engineer: Infusing information literarcy skills throughout an engineering curriculum, in *Proceedings of the American Soceity for Engineering Education Annual Conference & Exposition*, 2006.
32. D. Denick, J. Bhatt and B. Layton, Citation analysis of Engineering Design reports for information literacy assessment, in *ASEE Annual Conference and Exposition, Conference Proceedings*, 2010.
33. J. Younker and A. Mckenna, Ac 2009-680: Examining Student Use of Evidence To Support Design Decisions, in *Proceedings of the American Society for Engineering Education*, 2009.
34. Ş. Purzer and R. Wertz, Scaffold and assess: Preparing students to be informed designers, in *Integrating Information Into the Engineering Design Process*, M. Fosmire and D. Radcliffe, Eds. West Lafayette, IN: Purdue University Press, 2013, pp. 185–193.
35. S. N. Williamson, Development of a self-rating scale of self-directed learning, *Nurse Res.*, **14**(2), 2007, pp. 66–83.
36. M. Fisher, J. King and G. Tague, Development of a self-directed learning readiness scale for nursing education, *Nurse Educ. Today*, **21**(7), 2001, pp. 516–525.
37. L. M. Gugliemino, Development of the self-directed learning readiness scale, University of Georgia, 1977.
38. K. A. Douglas and Ş. Purzer, Validity: Meaning and Relevancy in Assessment for Engineering Education Research, *J. Eng. Educ.*, **104**(2), 2015, pp. 108–118.
39. R. F. DeVellis, *Scale development: Theory and applications*, Newbury Park, CA: SAGE Publications, Inc., 2003.
40. T. M. Haladyna and M. C. Rodriguez, *Developing and validating test items*, New York: Routledge, 2013.
41. W. J. Boone and K. Scantlebury, The role of Rasch analysis when conducting science education research utilizing multiple-choice tests, *Sci. Educ.*, **90**(2), 2006, pp. 253–269.
42. R. G. Netemeyer, W. O. Bearden and S. Sharma, Scaling procedures: Issues and applications, *Stat. Med.*, **23**(15), 2003, pp. 2480–2481.
43. R. E. H. Wertz, M. J. Fosmire, S. Purzer, A. I. Saragih, A. S. Van Epps, M. R. Sapp Nelson and B. G. Dillman, Work in Progress: Critical Thinking and Information Literacy: Assessing Student Performance, in *Proceedings of the Annual ASEE Conference*, 2013.
44. Center for Assessment and Improvement of Learning at Tennessee Technological University, *CAT Technical Information*, 2010.
45. M. J. Fosmire, R. E. H. Wertz and S. Purzer, Critical Engineering Literacy Test (CELT), 2013. [Online]. Available: https://purr.purdue.edu/publications/1118.
46. A. W. Meade and S. B. Craig, Identifying careless responses in survey data, *Psychol. Methods*, **17**(3), 2012, pp. 437–455.
47. S. L. Wise and X. Kong, Response time effort: A new measure of examinee motivation inf computer-based tests, *Appl. Meas. Educ.*, **18**(2), 2005, pp. 163–183.
48. T. M. Haladyna, A. State and S. M. Downing, Construct-irrelevant variance in high-stakes testing, *Educ. Meas. Issues Pract.*, **23**(1), 2004, pp. 17–27.
49. T. G. Bond and C. M. Fox, *Applying the Rasch Model*, 2nd ed. New York: Rutledge, 2007.
50. M. Swain, D. L. Sundre, and K. Clarke, *The Information Literacy Test (ILT) Test Manual*, Harrisonburg, VA, 2014.
51. C. Bruce, Information literacy as a a catalyst for educational change, in *Paper commissioned for UNESCO Information Literacy Leadership Conference*, 2003.
52. R. Catts, *Information Skills Survey Technical Manual*, 2005.
53. Y. Rosseel, Lavaan: An R Package for Structural Equation Modeling, *J. Stat. Softw.*, **48**(2), 2012.
54. H. Jiao and S. Wang, Modeling local item dependence with the hierarchical generalized linear model, *J. Appl. Meas.*, **6**(3), 2005.
55. P. Baghaei, The effects of the rhetorical organization of texts on the C-TEst construct: A Rasch modelling study, *Melb. Pap. Lang. Test.*, **13**(2), 2008, pp. 32–51.
56. P. Mair, R. Hatzinger, M. Maier and T. Rusch, eRm: Extended Rasch Modeling, 0.15-5, [Online]. Available: http://erm.r-forge.r-project.org/.
57. B. D. Wright and G. N. Masters, *Rating Scale Analysis*, Chicago, IL: MESA Press, 1982.
58. B. D. Wright, J. M. Linacre, J. E. Gustafson and P. Martin-Lof, Reasonable mean-square fit values, *Rasch Meas. Trans.*, **8**(3), 1994, p. 370.
59. W. P. Fisher, Rating scale instrument quality criteria, *Rasch Meas. Trans.*, **21**(1), 2007, p. 1095.
60. J. M. Linacre and B. D. Wright, Chi-square fit statistics, *Rasch Meas. Trans.*, **8**(2), 1994, p. 350.
61. B. J. Fraser, C. Johnson and R. A. Templeton, *Applications of Rasch measurement in learning environments research*, Rotterdamn, Netherlands: Sense Publishers, 2011.
62. D. W. Forbes, The use of Rasch logistic scaling procedures in the development of short multi-level arithmetic achievement tests for public school measurement, in *Proceedings of the American Educational Research Association*, 1976.
63. K. Sijtsma, On the use, the misuse, and the very limited usefullness of Cronbach's alpha, *Psychometrika*, **74**(1), 2009, pp. 107–120.
64. G. J. Clauser and J. M. Linacre, Relating Cronbach and Rasch reliabilities, *Rasch Meas. Trans.*, **13**(2), 1999, p. 696.
65. J. M. Linacre, KR/Cronbach alpha or Rasch person reliability: Which tells the 'truth'?, *Rasch Meas. Trans.*, **11**(3), 1997, pp. 580–581.
66. J. D. Scheuneman and R. G. Subhiyah, Evidence for the validity of a Rasch model technique for identifying differential item functioning, *J. Outcome Meas.*, **2**(1), 1997, pp. 33–42.
67. I. Paek and M. Wilson, Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel-Haenszel Procedure in Short Test and Small Sample Conditions, *Educ. Psychol. Meas.*, **71**(6), 2011, pp. 1023–1046.
68. R. M. Smith, The distributional properties of Rasch Item fit statistics, *Educ. Psychol. Meas.*, **51**, 1991, pp. 541–565.
69. R. M. Smith, Person fit in the Rasch model, *Educ. Psychol. Meas.*, **46**(2), 1986, pp. 359–372.
70. M. Oakleaf, Dangers and opportunities: A conceptual map of information literacy assessment approaches, *Libr. Acad.*, **8**(3), 2008, pp. 233–253.

71. R. Audunson and R. Nordlie, Information Literacy: The case or non-case of Norway, *Libr. Rev.*, **52**(7), 2003, pp. 319–325.
72. C. C. Kuhlthau, *Information skills for an information society: A review of research*, 1987.
73. M. Warschauer, The paradoxical future of digital learning, *Learn. Inq.*, **1**(1), 2007, pp. 41–19.
74. D. C. Briggs and M. Wilson, An introduction to multidimensional measurement using Rasch models, *J. Appl. Meas.*, **4**(1), 2003, pp. 87–100.
75. J. M. Linacre, Sample size and item calibration stability, *Rasch Measurement Transactions*, 1994. [Online]. Available: http://www.rasch.org/rmt/rmt74m.htm.

Kerrie A. Douglas is an Assistant Professor of Engineering Education at Purdue University. Her research is focused on supporting high-quality assessment practice in engineering education. This focus includes what evidence and rationale are used to justify educational data use and the consequences of that intended use. She earned her PhD in Educational Psychology, with a concentration on evaluation and assessment, from Purdue University in 2012.

Todd Fernandez is a PhD Candidate in Engineering Education at Purdue University. His research is focused on entrepreneurship and design in modern engineering education efforts. He has authored and co-authored multiple publications focused on understanding and assessing student cognition. Todd is active in the Entrepreneurship and Engineering Innovation division within the American Society for Engineering Education. He holds BS and MS degrees in Mechanical Engineering from the Rochester Institute of Technology. Before returning to graduate school he worked in the semiconductor industry and founded several companies.

Şenay Purzer is an Associate Professor in the School of Engineering Education at Purdue University and the Director of Assessment Research at the INSPIRE Institute for Pre-college Engineering Research. Senay is a NAE/CASEE New Faculty Fellow and the recipient of a 2012 NSF CAREER award. Her research focuses on assessment of design learning with a specific focus on information literacy, innovation, and decision-making processes. She received a B.S.E with distinction in Engineering at Arizona State University in 2009 as well as a B.S. degree in Physics Education in 1999. Her Ph.D. degree is in Science Education with a focus on engineering education from Arizona State University.

Michael Fosmire is a Professor of Library Science and Head, Physical Sciences, Engineering, and Technology Division of the Purdue University Libraries. He has written over 40 articles and chapters on the role of information in active-learning pedagogies and the integration of information literacy in STEM curricula, including co-editing *Integrating Information into the Engineering Design Process*, authoring the *Sudden Selector's Guide to Physic*s, and writing chapters on "Research in the Sciences" and "Engineering Research" in *Research in the Disciplines: Foundations for Reference and Library Instruction*.

Amy Van Epps is Director of Sciences and Engineering Services at Cabot Science Library, Harvard University. With over 20 years of experience as a subject librarian, Amy has extensive experience teaching engineering and technology students, and in 2017 won the Purdue Libraries' Excellence in Teaching award. Her research has focused in part on developing effective methods for integrating information literacy into the undergraduate engineering curriculum, particularly on the use of information in design settings. Amy is a long-time active member of the Engineering Libraries Division (ELD) of the American Society for Engineering Education (ASEE), and in 2014 won the Homer I. Bernhardt Distinguished Service award from that organization. Amy has an BA from Lafayette College in Engineering Science, with a focus on Mechanical engineering, a MSLS from Catholic University of America, and a Master's in Industrial Engineering from Rensselaer Polytechnic Institute. She is currently a doctoral candidate in Engineering Education at Purdue.