Assessing Critical Thinking in Higher Education: Validity Evidence for the Use of the HEIghten™ Critical Thinking Test in Ireland

Michael O'Leary
Dublin City University

Katherine Reynolds
Boston College

Guangming Ling
Educational Testing Service

Ou Lydia Liu
Educational Testing Service

Sarahjane Belton
Dublin City University

Naoimh O'Reilly
Dublin City University

John McKenna
Dublin City University

The HEIghten Critical Thinking Test (HCTT), developed by the Educational Testing Service (ETS), has gained traction both in the United States (US) and elsewhere. This study presents preliminary validity evidence for the use of the HCTT in Ireland. We provide evidence of the HCTT's overall structure, reliability, and relationships with student-level variables. While some item discrimination and reliability indices are not optimal, results of other analyses support the low stakes use of the HCTT in an Irish context. The procedures and outcomes of the study will be of interest to those planning to validate an existing high quality measure of critical thinking in local contexts.

Keywords: critical thinking, higher education, assessment, measurement, validity study

INTRODUCTION

Critical thinking (CT) is considered to be an essential competence for the 21st century workplace (McKinsey & Company, 2013). Because of this trend, many higher education institutions (HEIs) around the world explicitly name development of CT as a desired outcome for their students (see for example, Hill et al., 2016). However, details regarding how HEIs conceptualize CT or assess it are less prominent. There are several potential reasons for this, including lack of a widely accepted definition of CT in the literature (Nicholas & Labig, 2013) and a scarcity of assessment instruments/approaches designed specifically for use in HEIs.

Some instruments for assessing CT at HEIs do exist— this paper provides preliminary validity evidence for one such instrument in a context outside that for which the assessment was initially developed. The instrument in question is the HEIghten™ Critical Thinking Test (HCTT), developed by the Educational Testing Service (ETS) using a CT framework proposed by Liu, Frankel, and Roohr (2014). This test was initially developed for use in the United States, and the authors make it clear that explicit validation research is required to support its use in other contexts (ETS, 2017). Thus, the present study undertakes such validation research in the context of an Irish HEI.

Two issues frame this study. First, concern has been expressed that critical thinking is stifled by the educational experiences of Irish students at the end of post-primary education (see, for example, National Council for Curriculum and Assessment, 2019). The Leaving Certificate Examination (LCE) is taken by almost all secondary school students in Ireland and is a crucial element in gaining access to programs of study in most Irish HEIs. The two-year long Leaving Certificate program and terminal examination (known as *Senior Cycle*) are often criticized as encouraging rote learning and failing to adequately prepare students for further education (see, for example, Banks et al., 2018, O'Leary & Scully 2018). Second, while critical thinking is often included in lists of graduate skills that appear on the websites of Irish HEIs[1], direct measures of the skills that underlie critical thinking are not commonly used.

LITERATURE REVIEW

Critical Thinking and Assessment

While Facione (1990a) highlighted the potential benefits and uses of critical thinking assessment at varying educational levels in the early 1990s, assessment efforts have not been without their share of difficulties. Interest in such assessments at the postsecondary level has increased within the past several decades in response to greater emphasis on student outcomes in higher education (Klein et al., 2009). This push has led to the development of many different assessments purporting to assess critical thinking skills (for an overview of such assessments, see Liu et al., 2014). However, these efforts have not been without their challenges.

Ennis (2003) elucidates some of the many difficulties incumbent in assessing critical thinking. First and foremost is arriving at a commonly agreeable definition. Even if a definition is reached, identifying an appropriate tool with sufficient malleability and flexibility to assess critical thinking as it applies to a vast array of different content areas, professions, or domains of interest is a significant challenge. Some even argue that attempting large-scale assessment of critical thinking skills is of little value, citing difficulties in defining critical thinking, as well as arguments that critical thinking skills are domain-specific, rendering little value for the results of a general assessment (e.g., Rear, 2019). Tremblay et al. (2012) confront these difficulties, as well as issues of assessing critical thinking and other generic skills across multiple cultures, in their feasibility study for the Assessment of Higher Education Learning Outcomes (AHELO) sponsored by the Organisation for Economic Co-operation and Development (OECD). Although this project was eventually abandoned, initial efforts in its development reflect ongoing international interest in the postsecondary assessment of critical thinking and other so-called generic skills despite the difficulty in doing so.

Acknowledging the difficulty of developing critical thinking assessments in higher education, Liu et al. (2014) provide a comprehensive overview of critical thinking measures and their validity evidence.

Among the measures reviewed were, from oldest to most recent, the Watson–Glaser Critical Thinking Appraisal (Watson & Glaser, 1980), the Cornell Critical Thinking Test (Ennis et al., 1985), the Ennis–Weir Critical Thinking Essay Test (Ennis & Weir, 1985), the California Critical Thinking Skills Test (Facione, 1990b), the California Critical Thinking Disposition Inventory (Facione & Facione, 1992), the Halpern Critical Thinking Assessment (Halpern, 2010) and the Collegiate Assessment of Academic Proficiency Critical Thinking (CAAP Program Management, 2012). These assessments tend to cover areas related to reasoning, analysis, argumentation, or evaluation. They also highlight some of the more technical challenges arising in the assessment of critical thinking (in addition to the conceptual difficulties involved in construct definition). For example, should large-scale postsecondary critical thinking assessments prioritize instructional value (enabling the provision of actionable information for faculty) or standardization (enabling the comparison of students' critical thinking skills across institutions)? Similarly, should score utility be prioritized at the institutional or individual level?

The HEIghten Critical Thinking Test

The HEIghten Critical Thinking test (HCTT) is one of five tests within the HEIghten Suite, which also includes computer-based tests of written communication, quantitative literacy, intercultural competency and diversity, and civic competency and engagement (ETS, 2019a). HEIs may elect to purchase one or more of these tests; after the assessment, score reports are provided to both institutions and individual test-takers (ETS, 2019b). At the institutional level, ETS suggests that HEIghten scores may be used for several purposes, including documentation of general education skills, examining specific groups of interest, informing faculty development or training, and supporting accreditation or other accountability programs. Inappropriate uses of assessment scores are also highlighted, including using the tests as criteria for postsecondary graduation or other high-stakes decisions (ETS, 2017).

Liu et al. (2014) proposed an assessment framework for the HCTT. This framework suggests two overarching components of critical thinking: analytic and synthetic. The analytic component concerns students' ability to evaluate the quality of evidence and how it is used, as well as analyze and evaluate arguments. The complementary synthetic component concerns students' understanding of implications, consequences and ability to develop arguments. These skills within the analytic and synthetic components come together to represent students' overall understanding of causation and explanation.

The HCTT, administered via computer, features several types of items, including critical thinking sets, logical reasoning items, and analytical reasoning sets. The formats of these items include single- and multiple-selection multiple choice, drop-down menus, and select-in-passage (ETS, 2017).[2] Examinees are given 45 minutes to complete the assessment, which consists of 26 items (ETS, 2017). Students' overall scores on the HCTT (or other HEIghten exams) represent "an estimated statistical representation of a student's skill as represented by the HEIghten assessment module content" (ETS, 2017, p. 25). These scores range from 150 to 180. Subscores are also reported at the institutional level for the analytic and synthetic components of the assessment; these scores range from 1 to 10. Percentile ranks and performance level descriptors (i.e., Developing, Proficient, Advanced) are also provided (ETS, 2017). ETS (2019b) notes that the scaling process does allow for comparison of HCTT performance across institutions; however, such comparisons should be made cautiously.

Prior to institutional use of the HCTT, initial validity evidence was presented in Liu, et al. (2016). Five forms of the assessment (each containing 27-29 items) were evaluated using a sample of 3,036 postsecondary students in the United States. The authors present several pieces of promising validity evidence for the assessment, including unidimensional factor analysis solutions, Cronbach's alpha indices greater than .7 for three of the five forms, and positive correlations between HCTT scores and SAT/ACT scores. They also report that students in their fourth year of college outperformed those in their first year of college, aligning with the hypothesis that postsecondary experiences should bolster critical thinking skills. Finally, attending to the importance of motivation in low-stakes assessment (as demonstrated in Wise & DeMars, 2010, and others), Liu et al. (2016) found that students who reported giving their best effort on the HCTT score higher than students who indicated that they did not give their best effort.

Development and validation of the HCTT has allowed for large-scale examination of critical thinking skills in United States college students. Roohr et al. (2019) used HCTT scores to examine student and institutional characteristics that contribute to students' critical thinking skills. Using multi-level modelling, Roohr et al. (2019) found that there was generally more variation in HCTT scores within schools rather than between schools. They also found that the differences in HCTT scores between first and third/fourth year students were greater at institutions with lower retention rates and that student-faculty ratio and the percentage of African American or Black students enrolled at an institution had generally negative relationships with HCTT scores.

The present study seeks to gather preliminary validity evidence for the use of the HCTT with Irish postsecondary students. As discussed in the following section, best practice suggests that use of an assessment in a context other than that in which it was developed requires its own validation process. However, this does not always occur. Finnie et al. (2018) used the HCTT to measure critical thinking differences between entering and graduating students in a Canadian context. Finnie et al. (2018) cite Liu et al. (2016) to establish the validity of the assessment; however, they did not undertake any validation activities for their specific context. This is potentially problematic given the findings of Ku (2009) concerning reliability and validity of critical thinking tests. Ku (2009) found that for some commonly-used critical thinking assessments (including the aforementioned CCTST and CCTDI), validity and reliability evidence obtained by researchers other than exam developers tends to be much lower than what is reported in assessment documentation.

In contrast, Liu et al. (2016b, 2018) explicitly undertook validation activities for a Chinese version of the HCTT. Although validation efforts are considered essential whenever a test is used in a new context (such as with the present study), validation was particularly important in this case because the HCTT was translated into a new language and assessment format. Liu et al. (2018) piloted the Chinese version of the HCTT with 2,087 students (excluding students who did not complete at least 75% of the assessment) across 35 different Chinese universities. Both paper and computer versions of the HCTT were used in this pilot; however, most of the sample (93.6%) completed a paper version. (This is different from typical HCTT administration, which usually occurs via computer.) Similar to the initial United States validation study, the authors examined dimensionality, reliability, and relationships with other variables. They also conducted individual item analyses to examine difficulty and discrimination and looked for potential mode effects based on assessment administration format.

Beginning by examining the possibility of a mode effect, Liu et al. (2018) found statistically significant differences in performance between the paper- and computer-based versions of the assessment. Because of these, subsequent results were presented using data from the paper-based assessment only. The authors report evidence of assessment unidimensionality, as well as expected relationships with university elite status and students' self-ratings of critical thinking skills. While institution-level reliability for scores was quite high, individual-level reliabilities were lower than desired. With respect to item analysis, five items had discriminations below .3 and were flagged for further investigation.

Shaw et al. (2019) also carried out an international validation effort for a translated HCTT in Russia. The authors piloted a Russian version of the assessment with 1,060 university students enrolled in 34 universities. Similar to Liu et al. (2018) in the Chinese context, Shaw et al. (2019) found that most individual HCTT items fell within discrimination guidelines; however, two were flagged for further investigation. The Cronbach alpha reliability of the HCTT with the Russian sample (.67) was also deemed acceptable for a low-stakes assessment.

The selection of the HCTT as a measure of critical thinking to evaluate for Irish use was heavily influenced by the Liu et al (2014) review and its clear conceptual framework for critical thinking assessment and the validity evidence for the instrument presented by Liu et al. (2016a), in addition to the relatively successful adaptations of the assessment for other international contexts. The generic (rather than domain specific) definition of CT underlying the instrument was also appealing given the multidisciplinary nature of university programs at the sample institution. The HCTT also met practical requirements related to ease of administration and scoring, student testing time, and cost.

Validity Studies in International Contexts

Broadly speaking, test validity "(1) is *not* an inherent an inherent property of a test (2) refers to the interpretations or actions that are made on the basis of test scores, and (3) must be evaluated with respect to the *purpose* of the test and how the test is *used*" (Sireci, 2009, p. 20). This means that validity is not a property of a test itself; rather, it is a judgment made regarding the inferences drawn based upon test scores. Kane (2009) suggests that the validity of these inferences can be conceptualized as an argument, where different kinds of validity evidence are accumulated. Validity evidence collected in one context does not necessarily hold for different contexts; thus, it is necessary to go through the process of validation (i.e., the accumulation and reporting of validity evidence) whenever an assessment is to be used in a new context (Haladyna, 2006).

Given recent calls for quality assessment programs and "comprehensive validation" efforts to measure general skills (such as critical thinking) in higher education (e.g., Zlatkin-Troitschanskaia et al., 2015), there is great interest in gathering validity evidence for use of existing exams in this area. The need for localized validation efforts is also implied by the findings of Ku (2009) referenced above. Several studies concerning international use or validation of critical thinking assessments (among other generic skills) have been published within the past several years. Results and key takeaways from these studies are reviewed in the remainder of this section.

Al-Thani et al. (2016) sought to use the ACT's Collegiate Assessment of Academic Proficiency (CAAP) to evaluate how Qatari university students progressed in meeting general education goals set forth by a series of education reforms. One goal was the promotion of critical thinking, which is also featured as a component of the CAAP. The authors cited initial validity evidence for the CAAP, which was developed in the United States, as a justification for its use. However, they also gathered their own validity evidence by examining alignment among CAAP content and university course syllabi. Al-Thani et al. report a satisfactory degree of alignment, noting that "developing critical ability to analyze, evaluate, and extends arguments feature in the courses' description and pedagogical results" (p. 171). While this finding does constitute some promising validity evidence, examination of alignment only is likely not sufficient to support assessment inferences or score interpretations.

Verbaugh et al. (2013) added a more technical component to their study, which aimed to support the validity of two United States-based critical thinking tests for use with Flemish university students. Similar to Liu et al.'s (2018) work with the Chinese version of the HCTT, Verbaugh et al. (2013) note that "translating would not be sufficient to guarantee a valid instrument as cross-cultural assessment of generic skills as CT appears to be difficult" (p. 2). Specifically, the authors examined the Cornell Critical Thinking Test (CCTT) and the Halpern Critical Thinking Assessment (HCTA). A panel of higher education representatives was convened to evaluate correspondence between Flanders' accepted definition of critical thinking and content from the CCTT and HCTA; the authors report a degree of alignment. More technically, Verbaugh et al. also report unidimensionality for each assessment using principal components analysis and multidimensional item response theory. However, reliability for the assessments as captured by Cronbach's alpha was below acceptable levels (.53 and .64 for two HCTA administrations, and .52 and .52 for two CCTT administrations). This had implications for correlations between scores from the two assessments (both supposedly measuring the same critical thinking construct), which was weaker than expected.

Franco et al. (2018) also examined the validity of a translated critical thinking test-- specifically, a Portuguese adaptation of the HCTA. The authors used confirmatory factor analysis procedures to determine if a Portuguese administration would reveal the same five factors presented in the initial HCTA validity study: verbal reasoning, argument analysis, thinking as hypothesis testing, likelihood and uncertainty, and decision making and problem solving. This model also distinguished between constructed- and selected-response items within the assessment. Franco et al. report acceptable goodness-of-fit indices for the model with Portuguese data.

Although the studies presented above are somewhat positive in their findings regarding adaptation or use of United States-based critical thinking assessments in international contexts, this cannot be taken for granted. O'Hare and McGuinness (2015) examined the predictive validity of the multiple-choice California Critical Thinking Skills Tests (CCTST) and the Likert-type California Critical Thinking Disposition

Inventory (CCTDI) for psychology degree performance over and above what could be explained by A-level grades at a university in Northern Ireland. The authors found poor reliabilities for the three CCTST sub-domains of Analysis, Evaluation, and Inference. The Analysis subdomain score reliability was so poor (KR-20 = .02) that it was actually excluded from the study; however, the Evaluation (KR-20 = .52) and Inference (KR-20 = .40) sub-domains were retained because the study was intended to be exploratory. O'Hare and McGuinness did not set out to do a validation study for the CCTST; in fact, they note that the content of the test has been critiqued elsewhere (e.g., Fawkes et al., 2005). Their sample size was also relatively small (n = 109), which may contribute to the low reliabilities. However, these results highlight the importance of undertaking validation activities for assessments when they are to be used in new (particularly international contexts). Lack of adequate validation compromises not only one's confidence in inferences obtained from test scores, but also confidence in the relationships between those test scores and other variables of interest.

RESEARCH QUESTIONS

Whenever use of a test is proposed in a new context, new validity evidence must be gathered in order to establish that the test functions as intended (Kane, 2006). The overarching purpose of this study is to provide validity evidence for use of the HCTT within an Irish postsecondary institution. To that end, four primary research questions were addressed:

RQ1: *What is the internal structure of the HCTT when administered to an Irish sample?*

RQ2: *Is there evidence for meaningful relationships between HCTT scores and other Irish measures of academic achievement, most notably performance on the Leaving Certificate Examination (LCE) taken by all Irish students at the end of their secondary education?*

RQ3: *Do first and fourth-year university students score differently on the HCTT?*

RQ4: *Are Irish students motivated to do well on HCTT? Is there evidence for a relationship between levels of motivation and HCTT performance in the sample?*

A secondary research question concerns evidence of measurement invariance for the HCTT across United States and Irish samples. Because the sample used for this study was not designed to be nationally representative of Irish university students, strong claims of invariance cannot be made. However, exploratory analyses were conducted related to this issue.

STUDY DESIGN

Test and Sample Information

Two forms of the HCTT were administered to Irish students using ETS' secure on-line platform. The forms contain 26 items each and are designed to be psychometrically equivalent, meaning that the results on one form are comparable to the results of the other (ETS, 2017). Test items are constructed based on the critical thinking framework proposed by Liu et al. (2014), covering two dimensions of critical thinking: analytical and synthetic. The HCTT was piloted with a United States sample and yielded promising validity evidence (Liu et al., 2016a). The test is now fully operational within the United States.

Data for this study come from a large publically funded Irish university with over 17,000 students. Students come from every part of Ireland, with approximately 8% coming from socio-economically disadvantaged backgrounds. International students make up approximately 15% of the student body. Data were collected from volunteer student samples at three different time points: from incoming first-year students in September 2017 (n = 213), graduating fourth-year students in May 2018 (n = 63), and a mixture of first- (n = 231) and fourth-year (n = 54) students in October 2018. Students represented three different areas of study within the university: computer science, business, and health/human performance.[3] Details

regarding the samples at each administration are presented in Table 1. Test results had no direct consequences for students, although they were told that they could report critical thinking scores on their resumes or to potential employers if desired. Relevant data for the examination of critical thinking test *performance's relationships with other variables were obtained directly from the university. Ethical* approval was sought and obtained from the university's research ethics committee (equivalent to an Institutional Review Board in the United States).

TABLE 1
SAMPLE CHARACTERISTICS (COUNT DATA)

	Fall 2017 Administration		Spring 2018 Administration		Fall 2018 Administration	
	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2
Student level						
• *First year*	106	107	-	-	116	115
• *Fourth year*	-	-	32	31	27	27
Degree Program						
• *Computer science*	94		-		105	
• *Business*	49		-		109	
• *Science/health*	70		63		71	
Total	213		63		285	

Methods

At its core, the current study was designed replicate Liu et al.'s (2016a) analyses from the HCTT pilot in an Irish context. Specifically, this includes the examination of unidimensionality, reliability, measurement invariance, relationships with other variables, and analysis of student effort. However, unlike Liu et al. (2016a), who tested five potential HCTT forms, the present study uses only the two operational HCTT forms.

The internal structure of the critical thinking assessment (RQ1) was evaluated using three methods: individual item analysis, exploratory factor analysis (EFA), and a Cronbach reliability analysis. The item analysis involved calculating the item difficulties and the point-biserial correlation of each item with total HCTT score. EFA using oblique rotation was carried out using robust weighted least squares estimation (WLSMV) in Mplus (Muthén & Muthén, 2017) with a complementary parallel analysis performed in SAS. Critical thinking test reliability was assessed through the use of the Kuder-Richardson 20 (KR-20) statistic, which was calculated using Microsoft Excel. The KR-20 is comparable to the Cronbach's alpha reliability coefficient for binary data. Reliability analyses were performed separately for each form of the test to account for the fact that different forms contain different items and, while yielding psychometrically equivalent scale scores, may not function in the exact same way.

The secondary research question of measurement invariance is also addressed here through PISA linking procedures (OECD, 2017 using multidimensional latent trait modelling (MDLTM) software (von Davier, 2006). All items parameters and item-data fit indices were evaluated between the Irish sample and a comparison sample of 1,000 US students obtained through collaboration with ETS.[4] Those items with fit indices below the preset criterion value of 0.15 were accepted as the common items (OECD, 2014). Those items with a fit index value higher than the 0.15 value were released from the constraint of having the same item parameters between the United States and Irish samples in the next iterative step. This process ended when all item-data fit indices were below the .15 criterion.

Correlational analyses were used to investigate the relationship between the HCTT and Irish-specific measures of academic achievement (RQ2). The relationship between critical thinking test score and overall Leaving Certificate Examination (LCE) performance (represented by a points total based on grades from a

student's six best subjects – referred to in Ireland as Central Application Office [CAO] points) was examined visually using a scatterplot. Analysis involving CAO points derived from grades on the English LCE was also of interest as the HCTT is text-based and heavily language dependent. English is also one of the few compulsory subjects in the LCE, meaning all study participants had an English LCE score. The English analyses were performed using a one-way ANOVA procedure in SPSS. Although Leaving Certificate points are numeric, they represent performance categories and not a continuous variable; this premise guided our analytical choices for this research question.

Comparison of critical thinking test performance between first and fourth-year students (RQ3) was addressed using a t-test to check for differences in mean performance between the two groups, while the final piece of analysis conducted was focused on student effort on the HCTT (RQ4). Performance on the critical thinking test did not have direct consequences for students; research has shown that student motivation can be an issue in low-stakes exams and that motivation is associated with performance (Wise & DeMars, 2010; Wise & Kong, 2005). As such, it is important to evaluate the relationship between motivation and critical thinking test performance. Students provided a self-report of critical thinking test effort by indicating whether or not they tried their best during the assessment. It was expected that students who responded "yes" to this question would have higher score than students who responded "no". This was tested through the use of a one-way ANOVA performed in SPSS.

RESULTS

Descriptive Statistics

Table 2 presents the descriptive statistics for each form of the HCTT. 268 cases were used for Form 1 and 253 cases were used for Form 2. These numbers are slightly lower than the total number of students who completed the test; students who were flagged for exclusion from score reports (i.e., students who did not complete at least 75% of test items) were not included in any of the analyses presented below.

TABLE 2
MEAN DIFFERENCES IN SCALE SCORE BETWEEN TEST FORMS

Test Form	Mean HCTT Score (SE)	SD	t-value	p-value	d
Form 1 (n = 268)	165.50 (.38)	6.39	1.659	.98	.14
Form 2 (n = 253)	164.66 (.33)	5.57			

The item analysis revealed some differences between the two test forms. The range of item difficulties for Form 1 was .00 to .96, with an average difficulty of .56. For Form 2, difficulties ranted from .00 to .87, with an average difficulty of .49. The average item discrimination (as measured through point-biserial correlation with total score) was .25 for Form 1 and .11 for Form 2. Both test forms had one item where no students selected the correct answer, meaning that both forms had one item with a discrimination of 0. For Form 1, only one additional item had a discrimination less than .10; however, for Form 2, ten items had a discrimination below .10, and one item actually had a negative discrimination. Due to restrictions imposed by ETS in order to keep test content secure at this time, it was not possible to review the items themselves for why these differences occurred. It should be noted that items with low and/or negative discrimination values were also found in other validity studies involving the HCTT (e.g., Shaw et al., 2019) and further research on the issue will need to be undertaken by ETS.

Exploratory Factor Analysis

Figures 1 and 2 present scree plots of eigenvalues from the exploratory factor analysis, along with corresponding parallel analysis results. It can be seen that each form does yield a relatively strong first factor compared to the parallel analysis. These results should be interpreted cautiously due to the relatively

small sample size relative to the number of items used to conduct these analyses. It can also be seen that a negative eigenvalue was produced for Form 2; this is an artifact of the estimation procedure.

FIGURE 1
EIGENVALUES AND PARALLEL ANALYSIS FOR FORM 1

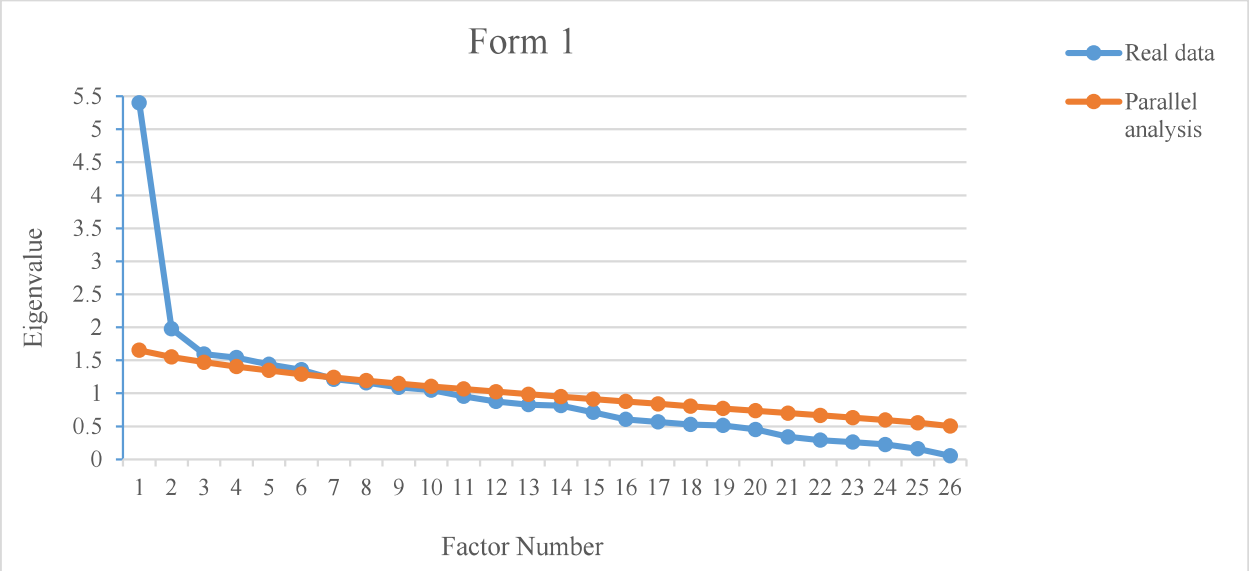
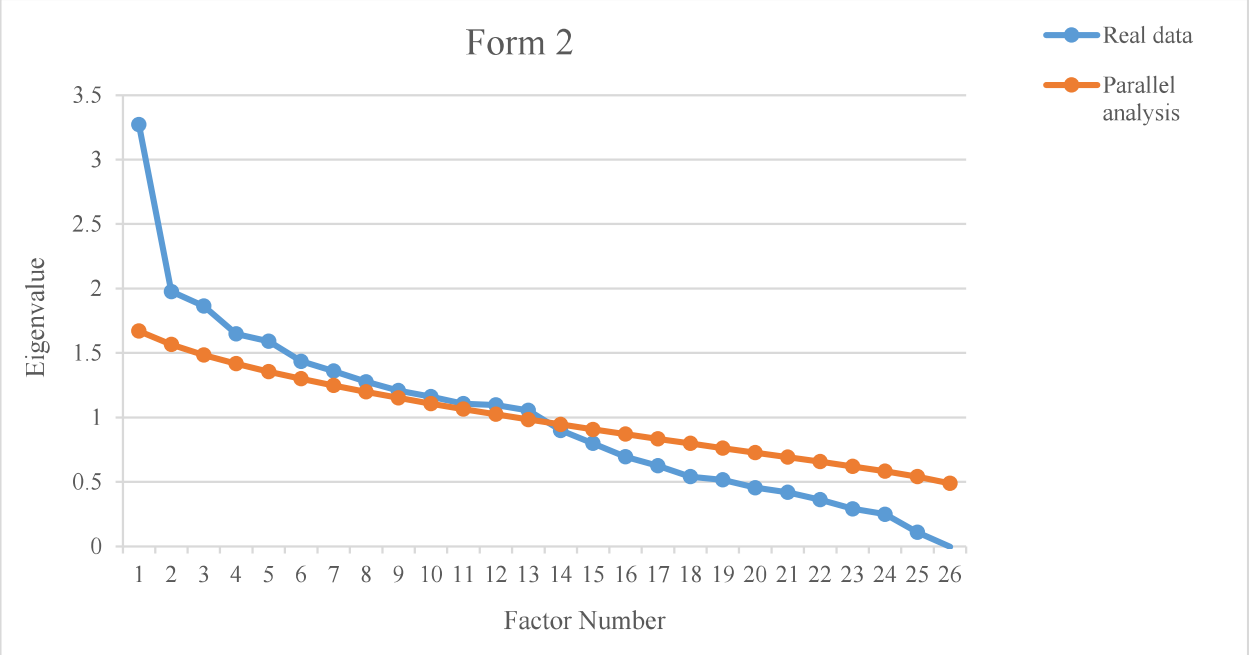


FIGURE 2
EIGENVALUES AND PARALLEL ANALYSIS FOR FORM 2



Both one- and two-factor models were fit to the data. While the test is hypothesized to represent the unidimensional construct of critical thinking ability, based on the results of Liu et al. (2016a), it was suspected that two factors (one representing the analytical dimension and one representing the synthetic

dimension) might emerge. Fit indices were obtained for both of these models. These indices are presented in Table 3.

TABLE 3
FIT INDICES FOR ONE- AND TWO-FACTOR MODELS (WSLMV ESTIMATION)

	Form 1 (n = 268)	Form 2 (n = 253)
One-factor model		
• CFI	.964	.807
• TLI	.961	.791
• RMSEA	.016	.020
Two-factor model		
• CFI	1.000	.895
• TLI	1.000	.876
• RMSEA	<.001	.015
Factor correlation	.326*	.299

* Significant at the .05 level.

The results for Form 1 suggest a unidimensional solution is appropriate. The CFI and TLI fit statistics for a two-factor model in Form 1 are suggestive of possible overfit. Other criteria support a single-factor solution for Form 1. A dominant single factor is supported by the scree plot. Additionally, there is a statistically significant correlation between the factors extracted in the two-factor model. Based on the rule of parsimony, a one-factor solution seems appropriate.

Form 2 presents a slightly different story. The scree plot does suggest the existence of a primary factor, although not as strongly as within Form 1. Additionally, the fit statistics for Form 2 suggest room for improvement; even the two-factor model does not meet the acceptable criterion for the CFI and TLI. The small sample size may also be an explanatory factor for these findings.

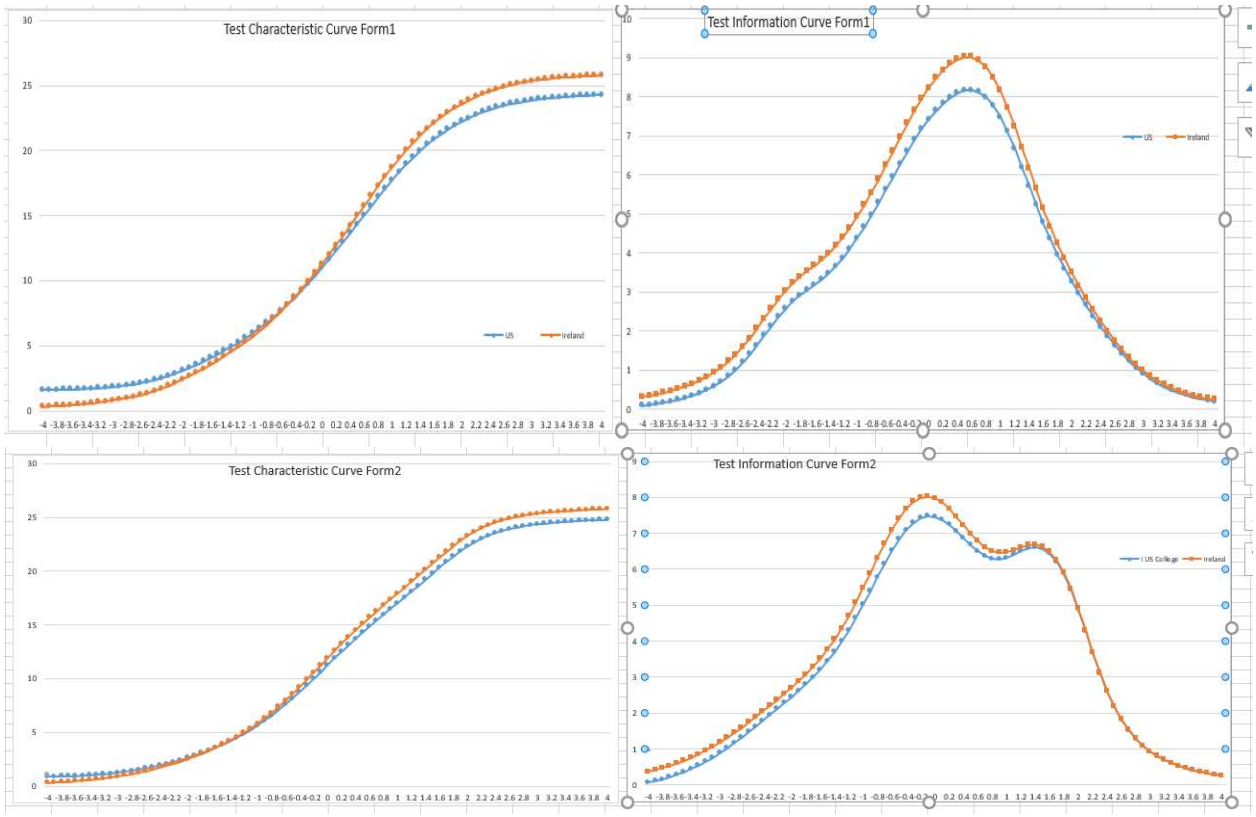
Reliability

KR-20 reliabilities were calculated for each form of the critical thinking test. The reliability for Form 1 was .725 (n = 268) and the reliability for Form 2 was .566 (n = 253). Reliabilities greater than .7 are generally considered acceptable (Cortina, 1993). This criterion was achieved for Form 1, but not for Form 2. This is not unexpected, given the results of the item analysis for Form 2.

Our secondary research question concerned the equivalence of item parameters across United States and Irish samples. Equivalent parameters across all items would provide evidence of measurement invariance, meaning that the items function comparably across the United States and Ireland. Results indicated that 24 of 26 items in Form 1 and 25 of 26 items in Form 2 can be considered the same (common) between the two samples. While complete measurement invariance was not established, the results do support a condition of partial invariance and are encouraging. This conclusion is also supported by an inspection of the test characteristics curves and test information curves for both forms of the HCTT in each sample as presented in Figure 3.

The test characteristic curves show the relationship between ability estimates and test scores, while the test information curves show the precision of measurement for individuals with different ability estimates. Although the curves for the United States and Irish samples are not exactly overlaid, they are not far apart and show similar patterns. This indicates that the relationship between test score and ability and the precision of measurement at different ability levels is comparable across the United States and Irish samples, providing further exploratory evidence of partial invariance.

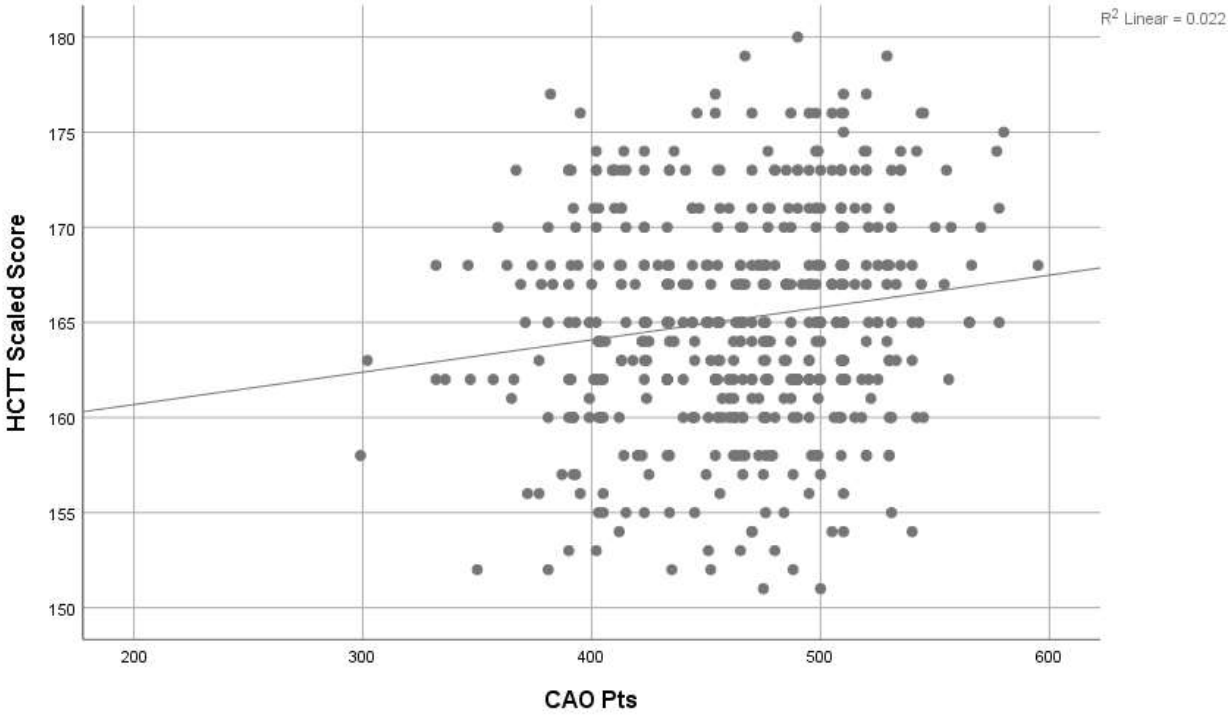
FIGURE 3
IRELAND AND US TEST CHARACTERISTIC AND INFORMATION CURVES FOR FORM 1 AND FORM 2 OF THE HCTT



Relationships With performance on the Irish Leaving Certificate Examination

The total number of LCE points (CAO points) information was available for 510 students in this study, 474 of whom answered at least 75% of assessment items. Given the criticisms of the Senior Cycle program in Ireland, it was hypothesized that the relationship between these points and HCTT performance would be positive, but not strong. The correlation between HCTT performance and CAO points was .149 ($\alpha < .05$) and is shown visually through the scatterplot presented in Figure 4.

FIGURE 4
SCATTERPLOT OF CAO TOTAL POINTS AND HCTT SCALED SCORES



In addition to overall CAO points, the relationship between HCTT scores and points awarded based on grades for the LCE in English were examined. The twelve grade/points categories possible were collapsed into low, medium, and high levels due to the small numbers of students in some categories. While presented numerically, the CAO points represent performance categories, justifying their collapse. The relevant data are presented in Table 6.

TABLE 6
DESCRIPTIVE STATISTICS FOR ENGLISH CAO POINTS AND HCTT SCALE SCORES

English CAO Points	Number of Students	Mean HCTT Score (SD)
0, 12, 20, or 28	6	159.67 (4.93)
33, 37, 46, or 56	84	163.65 (5.78)
66, 77, 88, or 100	410	165.61 (5.69)
Total	500	165.21 (5.76)

The ANOVA that was run was statistically significant ($F = 6.999$, $p < .001$), indicating that at least one statistically significant difference occurs among English points levels on HCTT scores. Post hoc tests indicated that these statistically significant differences occurred between the mean HCTT scores from the highest- and lowest-scoring English test groups. The partial eta-squared effect size was .027.

Relationships With University Year Status

The t-test comparing mean critical thinking test performance between first- and fourth- year students was not statistically significant. The results of this test are presented in Table 7.

TABLE 7
T-TEST RESULTS

Group	Mean HCTT Score (SD)	*n*	t	*p*	*d*
First-years	164.84 (5.58)	416	-1.977	.080	.204
Fourth-years	166.10 6.75)	105			

Effort

Only 321 students provided responses to the effort question presented at the end of the critical thinking test. The relevant descriptive statistics are presented in Table 8.

TABLE 8
DESCRIPTIVE STATISTICS FOR STUDENT EFFORT

Group	*n*	%	Mean HCTT Score	SE	Eta-squared
Tried their best	297	92.5	165.95	.323	.023
Did not try their best	24	7.5	162.75	1.020	

Most students indicated that they tried their best while completing the test. Although results should again be interpreted cautiously due to low response rates (within an already-small sample size), a one-way ANOVA yielded a statistically significant difference in mean critical thinking test score between students who reported trying their best and those who did not ($F = 7.417$, $p = .007$).

CONCLUSION

This study sought to gather initial validity evidence for the use of the Heighten Critical Thinking Test in an Irish context and the results will be of interest to a wider audience given the international push to assess critical thinking skills within higher education and the growing use of this particular test beyond the confines of the Unites States (e.g., China and Russia). Overall, the findings are generally in congruence with what was reported by Liu et al. (2016a) in their initial US validity study although it must be acknowledged that factor and reliability analyses for Form 2 are somewhat problematic and require further investigation. However, on balance, assuming low stakes use to guide teaching and learning around critical thinking, the findings suggest that the HCTT is suitable can be employed for assessment purposes in HEI contexts similar to the study site. This conclusion is also supported by the invariance results obtained from the multidimensional latent trait modelling of the data. The fact that the HCTT seems to measure something different to the LCE is a positive, given criticisms that performance on the LCE may not require high levels of critical thinking. The difference in performance between first- and fourth-year students was not statistically significant, although fourth year students did have a higher mean score. Additionally, the vast majority of the students who responded to the question about effort indicated that they tried their best and this effort was associated with higher HCTT scores.

A major limitation of study is the relatively small sample size for each form of the HCTT. This limitation is particularly salient for the factor analysis results. As noted, it was not possible in this study to review content in the search for an explanation for why the psychometric properties of certain items was less than ideal and this work remains an important undertaking for future studies involving larger samples and perhaps linked to similar investigations in other countries where the HCTT is being validated. However, the other validity evidence presented here provides some confidence for the Irish use of the HCTT to measure critical thinking among post-secondary students. In addition to replicating the analyses presented here with larger samples, future research will need to examine the utility of the HCTT for specific programmatic purposes, including informing the development of curricular interventions to bolster

students' critical thinking skills. Long term, as Shaw et al. (2019) note, there will be value in planning for studies that investigate how well the HCTT predicts outcomes beyond academic performance in college such as on-the-job decision making or life and career success.

ENDNOTES

1. See, for example, https://www.maynoothuniversity.ie/study-maynooth/maynooth-education/graduate-attributes; https://www.tcd.ie/TEP/graduateattributes.php; https://ceim.su.nuigalway.ie/about/graduate-attributes/
2. Sample items can be viewed at https://www.ets.org/s/heighten/pdf/critical-thinking-sample-questions.pdf
3. During one of their regular class periods, a faculty member from each of the three schools provided their student groups with a detailed description of the research and what it hoped to achieve, as well as information about what participation entailed in terms of time, commitment, confidentiality, risk/benefits and right to withdraw at any time. Students over the age of 18 were then invited to participate, with no inducements being offered to them other than drawing their attention to the fact that they could use their test data to provide evidence of their critical thinking skills to others e.g. employers. Information about the testing date, time and location was provided at a later stage to all students utilizing student academic group email lists.
4. The US sample used for the invariance study was based on a selected national pool of institutional users who voluntarily use HCT as an outcome assessment between 2016 and 2018. These institutions varied by control type, size, selectivity, geographic locations, as well as Carnegie type (e.g., doctoral/research universities, master level institutions, baccalaureate/liberal arts colleges, associate colleges, and specialized schools. The total sample size was 5329, with a stratified random sample of 1000 students (500 for each form) drawn for the study (G. Ling, email communication, February 17, 2020).

REFERENCES

Al-Thani, S.B.J., Abdelmoneim, A., Cherif, A., Moukarzel, D., & Daoud, K. (2016). Assessing general education learning outcomes at Qatar University. *Journal of Applied Research in Higher Education*, *8*(2), 159-176.

Banks, J., McCoy, S., & Smyth, E. (2018). *The Leaving Certificate Program as Preparation for Higher Education: The Views of Undergraduates at the End of their First Year in University*. Working Paper No. 607. Dublin: Economic and Social Research Institute (ESRI).

CAAP Program Management. (2012). *ACT CAAP technical handbook 2011–2012*. Iowa City, IA: Author. Retrieved from http://www.act.org/caap/pdf/CAAP-TechnicalHandbook.pdf

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

Educational Testing Service. (2017). *HEIghten™ Outcomes Assessment Suite Guide to Score Interpretation*.

Educational Testing Service. (2019a). *The HEIghten outcomes assessment suite*. Retrieved from https://www.ets.org/heighten

Educational Testing Service. (2019b). *HEIghten outcomes assessment suit scores*. Retrieved from https://www.ets.org/heighten/scores/

Ennis, R.H. (2003). Critical thinking assessment. In D. Fasko (Ed.), *Critical thinking and reasoning* (pp. 293–310). Cresskill, NJ: Hampton Press.

Ennis, R.H., &Weir, E. (1985). *The Ennis–Weir Critical Thinking Essay Test*. Pacific Grove, CA: Midwest Publications.

Ennis, R.H., Millman, J., & Tomko, T.N. (1985). *Cornell Critical Thinking Tests*. Pacific Grove, CA: Midwest Publications.

Facione, P.A., & Facione, N.C. (1992). *The California Critical Thinking Dispositions Inventory.* Millbrae, CA: California Academic Press.

Facione, P.A. (1990b). *The California Critical Thinking Skills Test-college level*. Technical report #2. Factors predictive of CT skills. Millbrae, CA: California Academic Press.
Facione, P.A. (1990a). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction executive summary.* Fullerton, CA: California State University.
Faulkes, D., O'Meara, B., Weber, D., & Flage, D. (2005). Examining the exam: A critical look at the California Critical Thinking Skills Test. *Science & Education*, *14*, 117-135.
Finnie, R., Dubois, M., Pavlic, D., & Suleymanoglu, E. (2018). *Measuring critical thinking Skills of postsecondary students.* Toronto: Higher Education Quality Council of Ontario.
Franco, A.R., Costa, P.S., & da Silva Almeida, L. (2018). Translation, adaptation, and validation of the Halpern critical thinking assessment to Portugal: Effect of disciplinary area and academic level on critical thinking. *Anales de Psicología*, *34*(2), 292-298.
Haladyna, T. (2006). Roles and the importance of validity studies in test development. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 739-755). Mawah, NJ: Lawrence Erlbaum Associates.
Halpern, D.F. (2010). *Halpern Critical Thinking Assessment manual*. Vienna, Austria: Schuhfried GmbH.
Hill, J., Walkington, W., & France, D. (2016) Graduate attributes: implications for higher education practice and policy. *Journal of Geography in Higher Education*, *40*(2), 155-163.
Kane, M. (2009). Validating the interpretations and uses of test scores. In R. Lissitz (Ed.), *The concept of validity* (pp. 39-64). Charlotte, NC: Information Age Publishing.
Klein, S., Liu, O.L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., . . . Steedle, J. (2009). *Test Validity Study (TVS) Report*. Supported by the Fund for Improvement of Postsecondary Education (FIPSE). Retrieved from http://www.voluntarysystem.org/index.cfm?page=research
Ku, K.Y.L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, *4*, 70-76.
Liu, O.L., Frankel, L., & Roohr, K.C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, *i-23*.
Liu, O.L., Mao, L., Frankel, L., & Xu, J. (2016a). Assessing critical thinking in higher education: the HEIghten™ approach and preliminary validity evidence. *Assessment & Evaluation in Higher Education*, *41*(5), 677-694.
Liu, O.L., Mao, L., Zhao, T., Yang, Y., Xu, J., & Wang, Z. (2016b). Pilot testing the Chinese version of the ETS® proficiency profile critical thinking test. *ETS Research Report Series*. ISSN 2330-8516
Liu, O.L., Shaw, A., Gu, L., Li, G., Hu, S., Yu, N., . . . Loyalka, P. (2018). Assessing college critical thinking: preliminary results from the Chinese HEIghten® Critical Thinking assessment. *Higher Education Research & Development*, *37*(5), 999-1014.
McKinsey & Company. (2013). *Voice of the graduate*. Philadelphia, PA: Author. Retrieved from http://mckinseyonsociety.com/downloads/reports/Education/UXC001%20Voice%20of%20the%20Graduate%20v7.pdf
Muthén, L.K., & Muthén, B.O. (2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.
National Council for Curriculum and Assessment. (2019). *Senior Cycle Review*. Retrieved from https://www.ncca.ie/en/senior-cycle/senior-cycle-review
Nicholas, M.C., & Labig, C.E. (2013). Faculty approaches to assessing critical thinking in the humanities and the natural and social sciences: Implications for general education. *The Journal of General Education*, *62*(4), 297–319.
O'Leary, M., & Scully, D. (2018). *The Leaving Certificate Program as preparation for higher education: The views of undergraduates at the end of their first year in university*. Dublin: CARPE/DCU. Retrieved from http://transition.ie/files/The_Leaving_Certificate_Programme_as_Preparation_for_Higher_Education.pdf
OECD. (2017*). PISA 2015 technical report*. Retrieved from https://www.oecd.org/pisa/data/2015-technical-report/

O'Hare, L., & McGuinness, C. (2015). The validity of critical thinking tests for predicting degree performance: A longitudinal study. *International Journal of Educational Research*, *72*, 162-172.
Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking. *Assessment & Evaluation in Higher Education*, *44*(5), 664-675.
Roohr, K., Olivera-Aguilar, M., Ling, G., & Rikoon, S. (2019). A multi-level modeling approach to investigating students' critical thinking at higher education institutions. *Assessment & Evaluation in Higher Education*, *44*(6), 946-960.
Shaw, A., Liu, L.O., Gu, L., Kardonova, E., Chirikov, I., Li, G., . . . Loyalka, P. (2019). Thinking critically about critical thinking: validating the Russian HEIghten® critical thinking assessment. *Studies in Higher Education*. https://doi.org/10.1080/03075079.2019.1672640
Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing.
Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education leaning outcomes feasibility study report volume 1: Design and implementation.* Organisation for Economic Cooperation and Development.
Verburgh, A., François, J., Elen, J., & Janssen, R. (2013). The assessment of critical thinking critically assessed in higher education: A validation study of the CCTT and the HCTA. *Education Research International*. https://doi.org/10.1155/2013/198920
von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models*. Princeton, NJ: Educational Testing Service.
Watson, G., & Glaser, E.M. (1980). W*atson–Glaser Critical Thinking Appraisal, forms A and B manual*. San Antonio, TX: The Psychological Corporation.
Wise, S.L., & DeMars, C.E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment*, *15*, 27-41.
Wise, S.L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163-183.
Zlatkin-Troitschanskaia, O., Shavelson, R.J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, *40*(3), 393-411.