Validity and reliability testing of the International Critical Thinking Essay Test form A (ICTET-A)

Helena Hollis
University College London
helena.hollis.14@ucl.ac.uk, orcid.org/0000-0003-1950-0657

Dr. Marina Rachitskiy
Regent's University London
marina.rachitskiy@regents.ac.uk, orcid.org/0000-0002-8736-029X?

Dr. Leslie van der Leer
Regent's University London
leslie.vanderleer@regents.ac.uk, orcid.org/0000-0001-7069-3760

Dr. Linda Elder
Foundation for Critical Thinking
elder@criticalthinking.org

Please cite this paper as:

Hollis, H., Rachitskiy, R., van der Leer, L., and Elder, L. (in press) Validity and reliability testing of the International Critical Thinking Essay Test form A (ICTET-A). *Psychological Reports*.

Abstract

This study aimed to assess the validity of form A of the International Critical Thinking Essay test as a tool for measuring critical thinking. To this end, we assessed the test for inter-rater reliability, internal reliability, and criterion validity. A self-selecting sample of participants (N = 100) completed the ICTET-A and a comparison test (the Ennis Weir Critical Thinking Essay Test) in an online, correlational, cross-sectional study. We found the ICTET-A items to have moderate to good levels of inter-rater reliability, and overall excellent inter-rater consistency for total test scores. The test had good internal reliability. There was a strong correlation between scores on the ICTET-A and the comparison test. Factor analysis showed that scores on ICTET-items were best explained with one factor, suggesting the test measures a single construct. The ICTET-A can therefore be considered a valid measure of critical thinking. Additionally, we propose a short form of the test for use in conditions where time constraints are a critical concern.

Introduction

The assessment of Critical Thinking (CT) is a concern in education (Bailin et al., 1999), business (Dwyer et al., 2015), and research (Meltzoff & Cooper, 2018). CT had historically been thought of as a set of logical reasoning abilities, concerned with bias detection and syllogistic reasoning (Ennis, 1964); however this conception of CT has been criticised for being too narrow, and stronger sense CT conceptions are more widely accepted (Paul, 1981). More modern conceptions of CT emphasise sensitivity to point of view, the ability to evaluate concepts and evidence, and other more sophisticated abilities (Ennis, 2018). With the recent focus on the issue of "fake news", the need to critically evaluate the contents of each piece of information we find has been brought to the fore (Batchelor, 2017). The ability to think critically has thus been emphasised as a wide-ranging societal need, and as being essential for democracy, freedom and autonomy (Holma, 2015). As Holma (2015) argues, this view of CT as necessary for democratic society depends upon a broad definition of CT that extends well beyond simple skills. Despite this, all currently available validated and up-to-date CT tests reflect a narrow conception of CT (Possin, 2008), and as such do not adequately measure the construct as it is more widely thought of. In this study we therefore seek to validate the International Critical Thinking Essay Test form A (ICTET-A), which we argue represents a means of measuring a broader and more impactful conception of CT.

A full review of CT tests falls outside the scope of this paper. However, all pre-existing and validated CT tests that we were able to identify at the time of writing fell into specific types with major limitations. Firstly, multiple-choice paradigms dominate in CT testing, and the current tests of this type are highly narrow to the logical and bias-spotting facets of CT (Ennis & Millman, 1985; Facione, 1990;

Watson & Glaser, 1980). Some tests seek to address this limitation by adding additional question types as well as multiple-choice, such as including some open text questions (Halpern, 2007), or including self-report disposition scales (Facione & Facione, 1992). Multiple choice formats and self-report scales have benefits in terms of simplicity of administration and avoiding assessment bias (Woodford & Bancroft, 2004), however they do not acknowledge the complexity of CT in a genuine information landscape (Thayer-Bacon, 2000). Therefore while multiple choice testing can be useful, it is too limited to fully assess CT in its entirety. Secondly, alternative means of assessing of CT are possible, such as using interviews that enable participants to fully explore and express their thinking (Kuhn, 1991). However, this is a highly time consuming method that would not be suitable for research with time and resource constraints. Finally, essay tests and rubrics can be used to assess CT by asking participants to read a piece of text and give their thoughts on it in an open form. Unlike interviewing, essay tests reduce the researcher workload required while still allowing for participants to express themselves more fully than in a multiple-choice paradigm. However, existing essay tests and rubrics are typically not validated (e.g. Gola, Ke, Creelman, & Vaillancourt, 2014), or are limited to a single specific text (Ennis & Weir, 1985).

In this study, we aim to validate an essay test we considered comprehensive in addressing a broad conception of CT: The International Critical Thinking Essay Test (ICTET) (The Foundation for Critical Thinking, 2019). The ICTET is a test of CT that requires participants to read a given text and answer questions based upon the Paul and Elder CT framework (Paul & Elder, 2005), demonstrating their ability to critically analyse it in a way that is reflective of wider, more robust CT conceptions. Answers are graded with a rubric, and this rubric can be used in conjunction with

various set texts on different topics. The test has two components: form A which asks specific questions about the text; form B which asks participants to write an open response to the text. This study is concerned with form A alone, as form A is more structured and offers the benefits of balancing free responses with efficient grading as previously discussed. Having been designed to match the Paul and Elder framework of CT, the ICTET has already met the requirements of face validity, and consequential validity (The Foundation for Critical Thinking, 2019). However, to date the ICTET has not been tested for reliability or criterion validity (i.e., to what extent it measures the same construct as other existing measures of CT) (Coolican, 2014). Therefore, it is the aim of this study to validate the ICTET form A (ICTET-A).

A test that assess the same construct (i.e. the same conception of CT) as the ICTET-A was needed for comparison so as to test the criterion validity of the ICTET-A. We opted for another essay test, as this would give the widest conception of CT. The Ennis-Weir Critical Thinking Essay Test (EWCTET) was selected as it has been previously validated (Ennis & Weir, 1985; Taube, 1997). At the time of writing, this was the only freely available and validated CT essay test we were able to locate. However, this test has limitations that are important to note, and these limitations further highlight the advantages of the ICTET-A. The EWCTET uses one specific text (a letter written by resident in a fictional town about parking regulations) which is somewhat outmoded in its language. Furthermore, the test may not be suitable for participants who are unfamiliar with driving and parking conventions in North American towns, as some of the arguments assume prior knowledge. Additionally, the text was written specifically for the test, and therefore lacks ecological validity. Other texts cannot be substituted in the EWCTET, as the grading rubric is specific to very concrete answers about features of the letter. By contrast, the ICTET-A can be

used with various texts and thus avoids these issues. These limitations lead us to conclude that the ICTET-A is a more suitable test than the EWCTET for most CT research needs. However, the EWCTET does provide a means of measuring a similar conception of CT as the ICTET-A, and therefore is taken to be a suitable comparison for assessing criterion validity.

Given the importance of CT, and the paucity of existing means of measuring it, the validation of the ICTET-A provides empirical evidence for its usefulness as a research and assessment tool for CT that is up-to-date and comprehensive. It fulfils a need for a validated CT assessment that matches the more modern conceptions of the construct of CT as a whole. In this study, we assess the validity of the ICTET-A in terms of inter-rater reliability, internal reliability, and criterion validity.

Method

An observational, cross-sectional study was conducted to test the validity of the ICTET-A as a measure of CT. An online survey was used to collect responses to both the ICTET-A and EWCTET. This study was pre-registered: https://osf.io/kdq4b

Participants

Participants were recruited using online calls for participation on social media (Twitter, reddit), study advertising platforms (callforparticipants, SurveyCircle), and via email newsletter distributed by the Foundation for Critical Thinking. Participants were self-selecting. A $10 amazon.com voucher was offered as a reward to each participant who completed the survey, with an additional $10 offered as a prize for the top 10% of scores across both tests.

A total of 128 participants responded to the online survey. Of these, 28 were removed due to incomplete answers; i.e. missing a scale, or submitting answers that

did not meet our minimum criteria of writing two complete and relevant sentences for each question. This yielded a total of 100 participants. Of these, 63 were female and 37 male. Education level was divided into three groups; no degree (n = 9), undergraduate degree (n = 53), postgraduate degree (n = 38). Age ranged from 18 to 67, with a mean age of 35 (SD=11.48).

Measures

International Critical Thinking Essay Test form A (ICTET-A).

In this test, participants were given a text to read, and then answered nine questions identifying the Purpose, Question at Issue, Information, Conclusions, Assumptions, Concepts, Implications, Point of View. Each question was graded on a scale of 1-10, resulting in a possible maximum score of 90. The sample text used in this study was an excerpt from Erich Fromm's *The Art of Loving,* as used in a previous test designed by The Foundation for Critical Thinking (Paul & Elder, 2012). This text was selected as the topic of "love" ought to be familiar and interesting to a majority of participants. Participants were instructed to write at least two complete and relevant sentences for each question. The test took approximately 30 minutes to complete. Table 1 shows the ICTET-A questions including explanatory notes, as presented to participants.

Table 1: ICTET-A questions

Number	Question
1	The main purpose of this text is... (Here you are trying to state as accurately as possible the author's purpose for writing the article. What was the author trying to accomplish?)
2	The key question(s) (whether stated or unstated) at issue is/are...

(Your goal is to figure out the key question that was in the mind of the author when s/he wrote the article. In other words, What was the key question which the article addressed?)

3 The most important information in this text is...

(You want to identify the key information the author used, or presupposed, in the article to support his/her main arguments. Here you are looking for facts, experiences, data the author is using to support her/his conclusions).

4 The main conclusion(s) in this text is/are...

(You want to identify the most important conclusions that the author comes to and presents in the article)

5 The main idea(s) we need to understand in order to understand this text is/are...

(To identify these concepts, ask yourself: What are the most important ideas that you would have to understand in order to understand the author's line of reasoning?)

6 Here is a short explanation of what the author means by this/these concept(s)...

(This refers to the concepts you answered the previous question with)

7 The main assumption(s) underlying the author's thinking is/are...

(Ask yourself: What is the author taking for granted (that might be questioned). The assumptions are generalizations that the author does not think s/he has to defend in the context of writing the article, and they are usually unstated. This is where the author's thinking logically begins).

8 The main implications of this line of reasoning is/are...

(What consequences are likely to follow if people take the author's line of reasoning seriously? Here you are to follow out the logical implications of the author's position. You should include implications that the author states, if you believe them to be logical, but you should do your best thinking to determine what you think the implications are.)

9 The main point(s) of view presented in this text is (are)…

(What is the author focused on and from what angle? The main question you are trying to answer here is: What is the author looking at, and how is s/he seeing it?)

Ennis Weir Critical Thinking Essay Test (EWCTET)

In this test, participants were asked to read a letter written by a fictional resident about parking regulations. They answered nine questions analysing the thinking in each paragraph of the letter, and gave a final analysis of the letter as whole. Each paragraph evaluation was graded with a maximum of 3 points, and the overall evaluation was graded with a maximum of 5 points. The maximum possible score was 29. The test took approximately 30 minutes to complete.

Procedure

The study was conducted online via Qualtrics. After obtaining informed consent, basic demographic information (age, gender, highest obtained educational level) was collected. Participants were then given overall instructions on how to answer the critical thinking tests. The tests were presented in a random order, and each came with specific instructions on how they would be graded. Both tests used open text entry spaces for each question. At the end of each test, participants were asked to confirm that they answered every question and were satisfied with their answers. Finally, participants were asked for their email address in order to receive their voucher reward for participation, and were given debriefing information.

Three independent graders assessed responses to the ICTET-A. One of the graders assessed the EWCTET responses, as inter-grader reliability of the EWCTET has already been assessed (Ennis & Weir, 1985), and was not a concern in this study.

Data was analysed using R 3.6.1. in R Studio version 1.2.1335; please see the reference list for packages used.

Results

Descriptive statistics

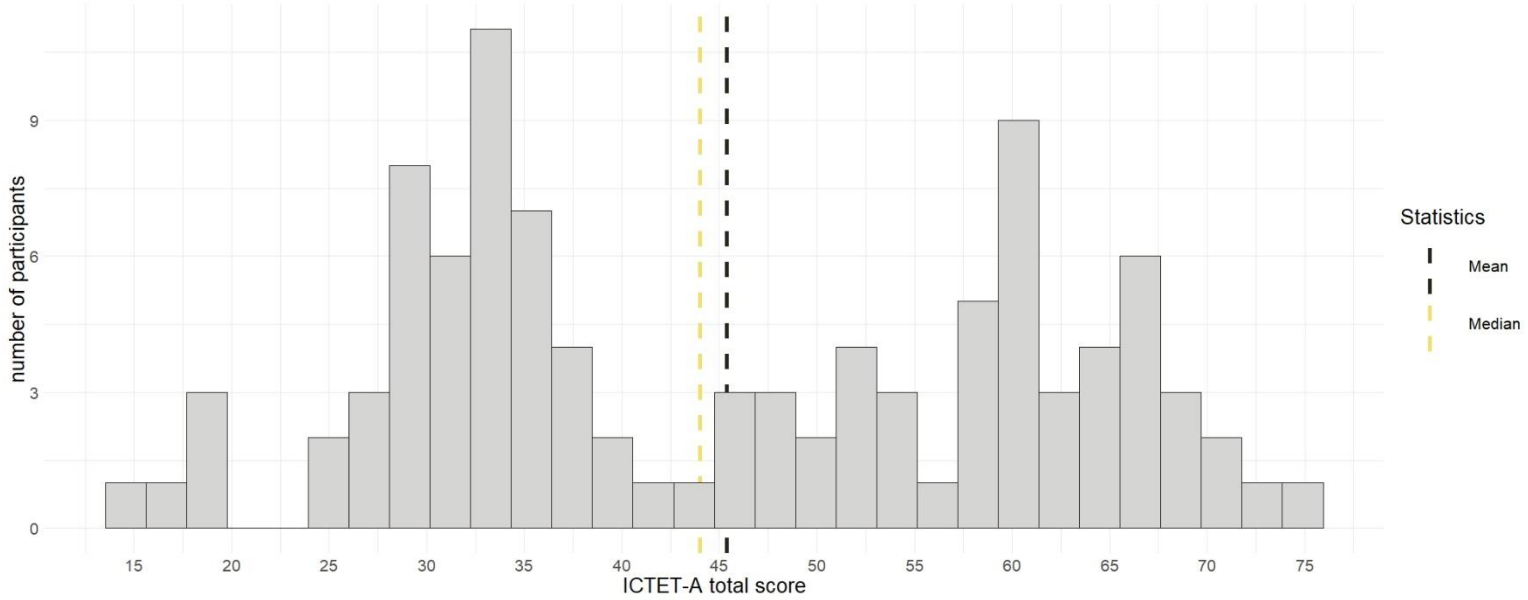
The mean of the three grader's total scores was taken as the overall ICTET-A total score for each participant. The mean score on the ICTET-A was 45.39 (*SD*=15.88), with a median of 44. The lowest score was 14.67 and the highest 75, giving a range of 60.33. As the maximum score available in the ICTET-A is 90, the sample mean represents 50% of the possible marks. See Figure 1 for a histogram of ICTET-A scores. Table 2 shows the mean scores on each ICTET-A question.

Table 2: ICTET-A question mean scores

Question number	Mean score
1	6.22
2	4.63
3	3.63
4	5.76
5	5.82
6	5.34
7	4.57
8	4.08
9	5.33

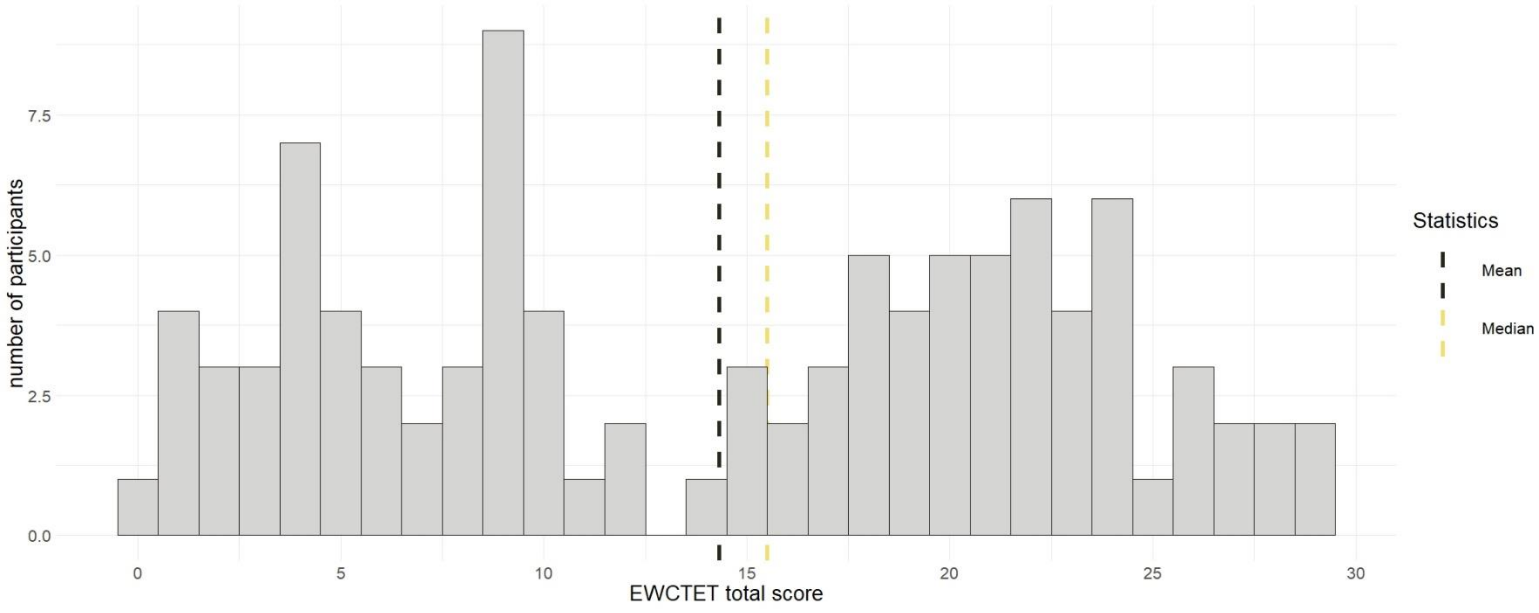
Note. Means are derived from all three graders' scores.

Figure 1: Histogram of ICTET-A total scores



The mean score on the EWCTET was 14.31 (*SD*=8.45), with a median of 15.5. The lowest score was 0, and highest 29. As the maximum score available in the EWCTET is 29, the sample mean represents 49% of the possible marks. See Figure 2 for a histogram of EWCTET scores.

Figure 2: Histogram of EWCTET total scores



Welch's unequal variance *t*-test found no significant difference in scores between men (M = 41.68, SD = 15.95) and women (M = 47.57, SD = 15.56) on the ICTET-A (p = .077). Men (M = 13.00, SD = 9.35) and women (M = 15.08, SD = 7.85) also did not significantly differ on the EWCTET (p = .260).

Levene's test for homogeneity of variance found no significant difference in the variance of ICTET-A scores and education (F(2,97) = 1.52, p = .223). Therefore, a one-way ANOVA was carried out, which showed a significant difference (F(2, 97) = 12.98, p <.001). Post-hoc Tukey's comparisons showed that the only significant difference (p<.001) was between UG (M = 39.26, SD = 12.95) and PG level (M = 54.59, SD = 14.68), with the ICTET score increasing on average from UG to PG by 15.33 [95% CI 22.54, 8.12]. The difference between ICET-A scores for participants in the no degree category (M = 42.67, SD = 19.4) and UG was not significant (p = .785), as was no degree to PG (p = .067).

Levene's test for homogeneity of variance found no significant difference in the variance of EWCTET scores and education (F(2,97) = 2.24, p = .112). Therefore a one-way ANOVA was carried out, which showed a significant difference (F(2, 97) = 8.506, p <.001)). Again, post-hoc Tukey's comparisons showed that the only significant difference (p<.001) was between UG (M = 11.51, SD = 8.23) and PG (M = 18.39, SD = 6.59) level with the EWCTET score increasing on average from UG to PG by 6.89 [95% CI 10.87, 2.90]. The difference between EWCTET scores for participants in the no degree category (M = 13.56, SD = 10.55) and UG was not significant (p = .752), as was no degree to PG (p = .227).

As the scores for the ICTET-A and EWCTET were not normally distributed (see Figures 1 and 2), Spearman's rank correlation coefficient was used to test for a correlation between test scores and age. ICTET-A scores and age had a rho of 0.021 (p = .838). EWCTET scores and age had a rho of 0.083 (p = .410). Therefore age was not significantly related with scores on either test.

Inferential statistics

ICTET-A reliability

Three graders marked responses on the ICTET-A in order to be able test inter-rater reliability, both in terms of consistency and agreement. Consistency is a measure of how similar in proportion the distances among graders are from each score a grader gives the same participant to that grader's mean score (i.e. each grader may give higher or lower scores, but with regularity in how much higher or lower), while agreement is a measure of the extent to which different graders give the same scores to the same participant (Koo & Li, 2016). Two-way intra-class correlation coefficient models were used. Table 3 shows the results for inter-rater agreement, and Table 4 shows inter-rater consistency.

Table 3: Inter-rater agreement

		Agreement		
Question	*F*	Lower 95% CI	ICC	Upper 95% CI
1	(99,184) = 9.18	0.641	0.725	0.796
2	(99,23.1) = 15.6	0.607	0.774	0.864
3	(99,6.04) = 19.7	0.288	0.713	0.867
4	(99,92.4) = 7.41	0.534	0.651	0.746
5	(99,200) = 8.66	0.634	0.718	0.79
6	(99,178) = 9.15	0.638	0.723	0.795
7	(99,108) = 13.2	0.704	0.785	0.847

8	(99,18.9) = 8.56	0.381	0.62	0.62
9	(99,12.7) = 5.5	0.153	0.441	0.643
Total score	(99,11) = 34.7	0.672	0.865	0.933

Note. All p values were < .001.

Deploying guidelines suggested by Koo and Li (2016), we take ICC values of less than 0.5 to indicate poor, 0.5 – 0.75 moderate, 0.75-0.9 good, and greater than 0.9 excellent reliability. In terms of inter-rater agreement, questions 1, 3, 4, 5 and 6 have moderate reliability; questions 2, 7, and the total scores have good reliability. Question 9 has poor reliability in terms of inter-rater agreement.

Table 4: Inter-rater consistency

		Consistency		
Question	*F*	Lower 95% CI	ICC	Upper 95% CI
1	(99,198) = 9.18	0.650	0.732	0.801
2	(99,198) = 15.6	0.773	0.83	0.876
3	(99,198) = 19.7	0.814	0.862	0.9
4	(99,198) = 7.41	0.590	0.681	0.761
5	(99,198) = 8.66	0.635	0.719	0.791
6	(99,198) = 9.15	0.650	0.731	0.8
7	(99,198) = 13.2	0.738	0.802	0.855
8	(99,198) = 8.56	0.631	0.716	0.789
9	(99,198) = 5.5	0.495	0.6	0.695
Total score	(99,198) = 34.7	0.888	0.918	0.942

Note. All p values were < .001.

In terms of inter-rater consistency, questions 1, 4, 5, 6, 8, and 9 have moderate reliability; questions 2, 3, and 7 have good reliability, and the total scores have excellent reliability.

Cronbach's alpha was used to test the internal reliability of items in the ICTET-A. Table 5 shows the results.

Table 5: Internal reliability

Grader	Lower 95% CI	Alpha	Upper 95% CI
1	0.942	.953	0.961
2	0.908	.925	0.938
3	0.774	.816	0.846
Mean	0.923	.936	0.946

Note. Bootstrap 95% CI based on 1000 samples.

Utilising Ponterotto and Ruckdeschel's (2007) reliability matrix, an alpha of .85 or above is deemed to be excellent, and an alpha between .80 and .85 is good. Therefore, graders 1 and 2 had excellent internal reliability in ICTET-A scores, and grader 3 had good reliability. The mean scores across the 3 graders had excellent reliability.

Additionally, the mean scores across the three graders were tested for internal reliability with the removal of ICET-A items, to determine if removing any items from the test would improve reliability. Table 6 shows Cronbach's alpha values with each item dropped from the test.

Table 6: Internal reliability with item removed

Item removed	Lower 95% CI	Alpha	Upper 95% CI
1	0.910	.928	0.941
2	0.921	.935	0.945
3	0.918	.932	0.942
4	0.916	.932	0.944
5	0.911	.927	0.939
6	0.907	.924	0.938

7	0.906	.924	0.937
8	0.906	.923	0.935
9	0.915	.931	0.943

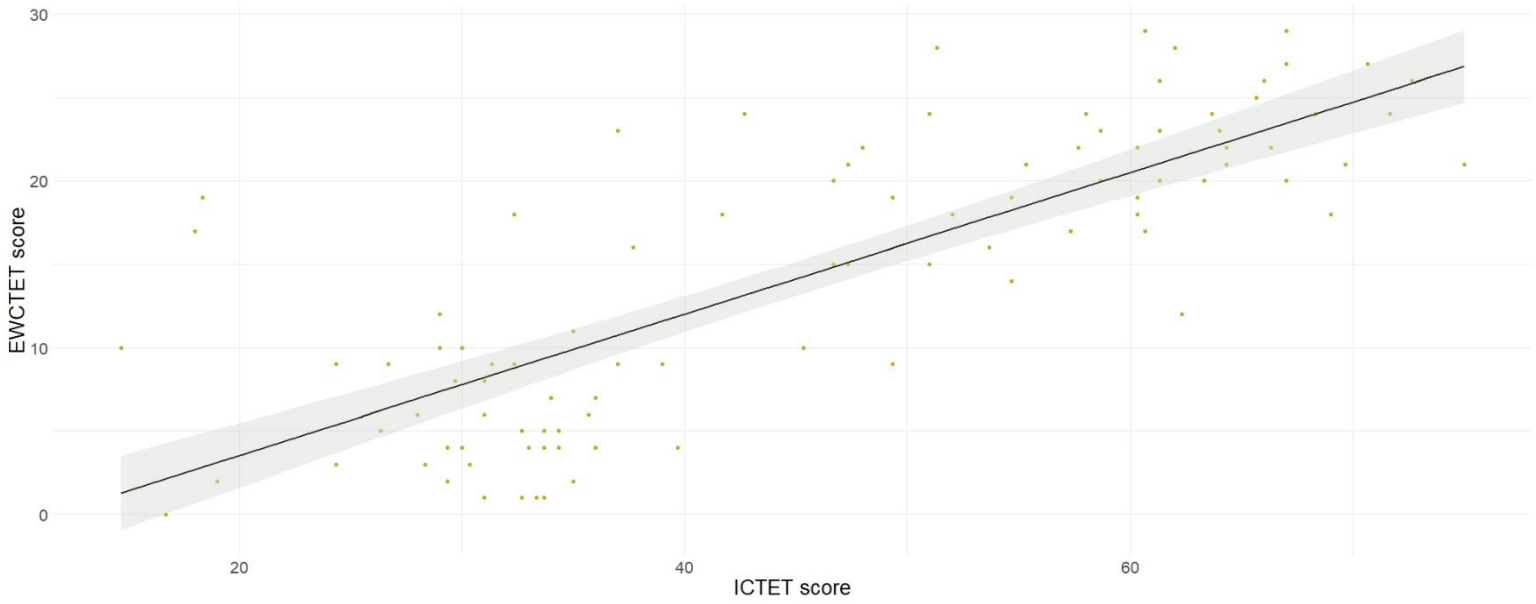
Note. Bootstrap 95% CI based on 1000 samples.

As the alpha value for all questions included was .936, the removal of any item does not improve reliability.

ICTET-A validity

In order to assess the criterion validity of the ICTET-A, the EWCTET was used as a comparison test. Scores on the ICTET-A and the EWCTET were compared using Spearman's rank correlation due to the data not being normally distributed. Results of the Spearman correlation indicated that there was a significant positive association between ICTET-A and EWCTET scores, ($r_s(98) = .78$, $p < .001$). Figure 3 shows a scatterplot of scores on the two tests.

Figure 3: Plot of EWCTET and ICTET-A scores



ICTET-A factor analysis

Exploratory factor analysis was conducted to assess the number of factors that make up the construct of critical thinking as measured by the ICTET-A. Parallel analysis was used to determine the optimal number of factors to retain, and this found that a single factor solution was best. Additionally, the optimal number of factors to retain according to optimal coordinates, the acceleration factor, and the Kaiser rule was also one. Table 7 shows the fit of a maximum likelihood common factor model for one factor.

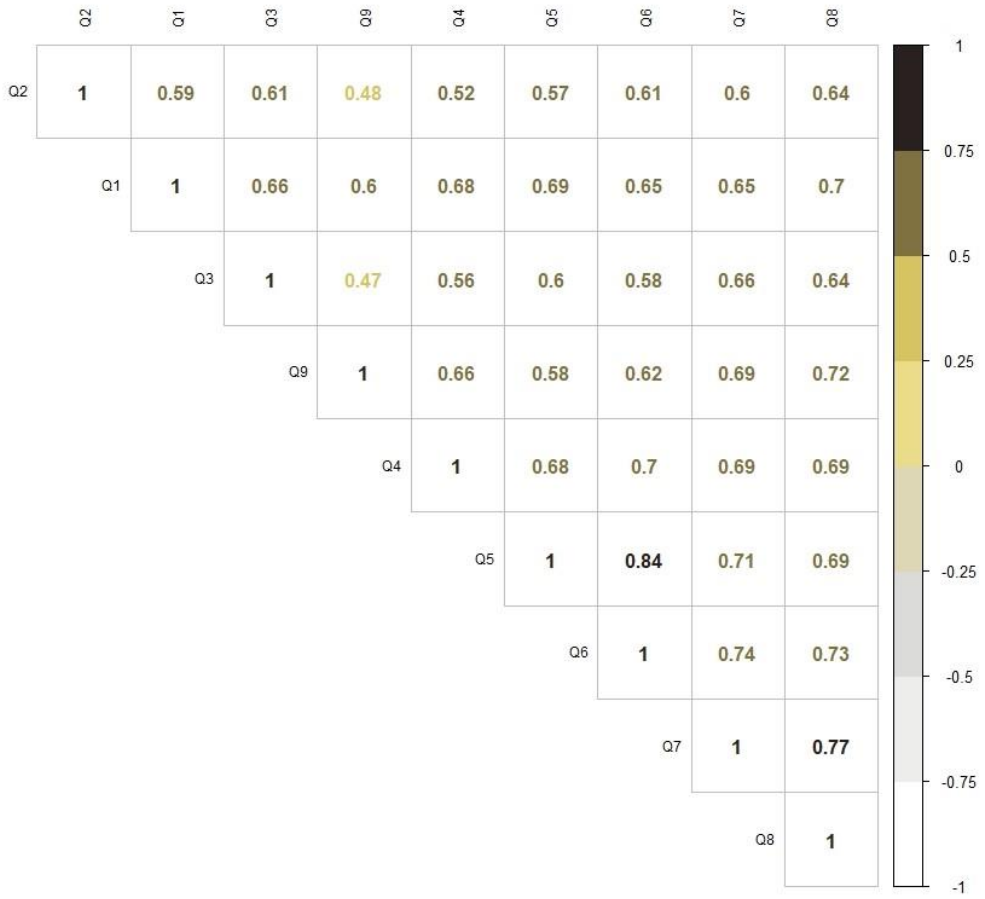
Table 7: One factor model fit

Question	Factor 1 loading	Uniqueness
1	.800	.360
2	.705	.502
3	.728	.470
4	.805	.352
5	.849	.279
6	.866	.250
7	.862	.257
8	.870	.242
9	.752	.434
SS loadings	5.853	
Proportion of variance	0.650	
$X^2 = 53.06$ [df 27], $p =$ .00198		

Questions 2, 3 and 9 have relatively lower loadings (below 0.8) for the single factor, and also show the highest uniqueness scores, suggesting these questions are least reflective of the one factor.

Figure 4 shows a correlation matrix, displaying correlation coefficients. The correlations between each question are displayed. Values are graded to indicate correlation strengths, as described in the key.

Figure 4: ICTET-A question correlation matrix



Question 8 had the strongest correlation across all questions with no values below 0.6, and question 2 had the weakest with four values below 0.6. The lowest correlations were between question 3 and 9 (0.47), and 2 and 9 (0.48). The highest correlation was between questions 5 and 6 (0.84).

Discussion

The ICTET-A has moderate to good inter-rater agreement, with the exception of question 9 for which agreement was poor. Inter-rater consistency was moderate to excellent. In terms of internal reliability as determined by Cronbach's alpha, two graders had excellent reliability and one grader had good reliably, and the average scores across the three graders had excellent reliability. A positive correlation was found between scores on the ICTET-A and EWCTET, demonstrating criterion validity for the ICTET-A. A factor analysis found that ICTET-A responses are best explained by a single factor.

As grading the ICTET-A is a subjective activity, with each grader interpreting the marking criteria and judging responses against them independently, we would expect differences in interpretation across graders. The overall range from moderate to excellent (with the exception of question 9) in inter-rater reliability is therefore better than we may expect from grading that is so subjective in nature. Inter-rater consistency was better than inter-rater agreement, which indicates that although different graders may be interpreting the rubric differently so as to give lower or higher marks in general, their interpretation as to which answers reflect higher or lower scores is stable. However, although consistency and agreement can diverge strongly (Hallgren, 2012), this was not the case in ICTET-A grading, where consistency had higher ICC values but the agreement values were similar; questions 1, 4, 5 and 6 had moderate reliability, while 2 and 7 had good reliability, for both agreement and consistency. For a test of this nature, we argue that inter-rater consistency is of more importance than agreement. The inter-rater consistency for total scores was excellent, suggesting excellent reliability in the overall scoring patterns of the different graders assessing the ICTET-A.

Scores on both the EWCTET and ICTET-A were not normally distributed, but rather had two peaks in their distribution. This suggests a distinction between low scoring and high scoring participants. This may be best explained due to the nature of the reward offered to incentive participation, however; there may be a distinction between participants who merely aimed to complete the tests and obtain the $10 voucher, and those who tried to complete the tests as well as possible with the aim of also obtaining the additional $10 reward for having a score in the top 10%. However, as test responses were anonymous we did not perform any follow-up questioning of participants, and therefore our existing data cannot explain this trend. It would be interesting to see if this distribution pattern is found in ICTET-A scores with different reward structures in future research.

Removing ICTET-A items

In grading responses, we found that answers to question 9 (asking for the author's point of view) were often repeating what participants had already written in their previous answers. One participant noted this themselves, stating:

> "I believe this question to be redundant with previous questions in this test, so that answering it here would not provide any new information. Please refer to my previous answers."

as their answer to question 9. Indeed, it seems inevitable that the author's point of view must already have been taken into account, at least to some extent, when answering questions about the purpose, conclusions etc. of the text. Alternatively, this may simply be due to this question falling last in the test, and thus being most subject to fatigue, rather than due to the content of the question. Nonetheless, this observation is reflected in the results, as question 9 had the lowest correlations with other questions, and had a relatively low loading to the single factor that best fit the

data, and a relatively high uniqueness value. Furthermore, question 9 had a poor ICC for inter-rater agreement, and also had the lowest score for inter-rater consistency. Removing question 9 yields a Cronbach's alpha of .931 for the mean ratings across all three graders. The alpha including question 9 is .936. This suggests that keeping question 9 does lead to slightly higher internal reliability, but as the difference is so small the question could be omitted. We suggest its omission in situations where participant fatigue is likely to be an issue.

Question 2 asks for the key questions addressed in the text. However, when grading answers to this question we noticed that participants often failed to state any questions, but rather gave answers that would be more suitable for question 1 that asks about the purpose of the text. This observation is supported by the results which showed that question 2 had the weakest correlation with the other questions, and the lowest loading and highest uniqueness value for the single factor that best fit data. Removing question 2 yields a Cronbach's alpha of .935 for the mean ratings across all three graders. The alpha including question 2 is .936. This difference is so small it suggests that questions 2 does not make a substantial difference to the internal reliability of the ICTET-A. Where a shorter form of the test is preferred, we suggest its omission.

Removing both questions 2 and 9 results in a Cronbach's alpha of .930. As the alpha including both questions is .936, this suggests that both can be omitted to form a briefer version of the ICTET-A without meaningfully impacting the internal reliability of the test. As both questions appeared to prompt repetition in answers, we believe the shorter version could maintain the essential components of the Paul and Elder CT framework (Paul & Elder, 2005), without explicitly asking for each of its elements in a distinct test question. As removing these questions does not improve

internal reliability, the short form of the test ought not be preferred to the full ICTET-A, however. The short version ought only to be deployed under circumstances where time and fatigue constraints are a likely concern. Furthermore, as the repetition in answers may be attributable to the presentation of the test in an online format (in which participants cannot easily see all the questions at once, and thus may be more likely to scroll and answer one by one without first reading all the questions), we recommend the use of the short form in these conditions, while the long form ought to be used in paper presentations of the test or in contexts where exam technique is expected of participants. Thus, the short form does not constitute a replacement for the full ICTET-A, but rather an option for specific circumstances.

Clarification of instructions

Question 3 asks for the key information in the text. However, the term "information" is ambiguous, and it can have different connotations across different contexts (Pawley, 2003). In grading we noted some participants interpreted "information" to mean different things; some participants identified the information the author was presenting, rather than the information he was utilising, which is what the further clarification notes of the question specify. Misunderstanding of the question may explain why question 3 had a relatively lower loading and higher uniqueness for the single factor that best fit the data. As shown in Table 2, question 3 also had the lowest mean score. Removing question 3 yields a Cronbach's alpha of .932 (as compared to .936 with it included), suggesting it could be omitted without meaningfully impacting the internal reliability of the test. However, rather than removing this question an amendment to its phrasing may be more appropriate, given that it appeared to be the way participants understood the word "information" that caused some to score poorly. The notes given specify that the question is

targeting “facts, experiences, data”; we suggest that this is better termed “supporting information or evidence” rather than just “information”. We therefore recommend rephrasing this question to use the words “supporting information or evidence”; however this will require further investigation as to whether this increases the question’s clarity and improves reliability.

Questions 5 and 6 are inter-related, with the former asking participants to identify the key concepts in the text and the latter to define them. We noticed that some participants did both in their answer to question 5, and then had little to add in their response to question 6. This may be due to the presentation of the test in an online form, which may encourage participants to scroll and answer questions one after another. As our sample were not recruited within an educational context (although the majority had degrees), we cannot expect exam technique knowledge to be common among them, and therefore cannot expect participants to read through all of the questions to begin with before they commence answering. We therefore suggest adding an explicit instruction to read all the questions prior to answering, and to add a note to question 5 alerting participants to the fact that question 6 will follow-on from it. However, it should be noted that questions 5 and 6 had moderate reliability, and were strongly correlated with each other as we would expect given their shared topic. Despite our observation that some participants struggled to divide their answers between the two questions, our data do not raise concerns over question 6.

Limitations

CT is considered to be an essential capacity across many contexts, including everyday life, the workplace and others beyond educational institutions (Siegel, 1997). Therefore, we aimed to recruit participants from the general population.

However, the majority of our participants had a university degree, and as such represent a highly educated sample. Thus the suitability of the ICTET-A for use with participants of different educational levels still requires further investigation.

One grader received training in the Paul and Elder (2005) framework upon which the ICTET-A is based, and then prepared the other two graders. The ICTET-A is not designed for use without such prior training. Furthermore, the three graders who assessed responses to the ICTET-A have all worked in the UK higher education context, and have experience of grading student work within its conventions. It may be the case that this has primed the graders to evaluate answers in similar ways, and that this contributes to the consistency in their grading beyond the instructions offered by the test's grading rubric. It would be worthwhile to test the ICTET-A for inter-rater reliability when graders come from different backgrounds.

As a result of our experience grading the responses to the ICTET-A, and from our findings, we have made suggestions for possible changes to the test (i.e. rephrasing question 3, and adding clarifying instructions). However, the efficacy of these changes remains to be tested. Furthermore, this study validated the ICTET-A using one text; additional validation studies with other texts would enable broader subject applications, as well as facilitating longitudinal use of the test.

Conclusion

The ICTET-A is a CT test that addresses the wider sense of what it means to think critically, not merely a test of logical or syllogistic reasoning. This study showed the test to be a valid instrument for testing CT. Although the test is graded subjectively, this study shows it to have a moderate to good level of inter-rater reliability, with excellent inter-rater consistency for total test scores. The test was

also shown to have good internal reliability, and to correlate with scores on a comparison test thus demonstrating it assesses the same CT construct. Scores on the nine test items load onto a single factor, indicating that the test is measuring a single construct. We suggest that a shorter form of the test for use in time constrained circumstances could also be deployed by omitting two questions. In either its full or shortened form, we find the ICTET-A to be a suitable measure for research into CT.

Notes

Ethical approval

The project was approved by the Psychology Research Ethics Committee at Regent's University London (reference: 19.36).

Informed consent

Informed consent was obtained from all participants included in the study.

References

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies*, *31*(3), 285–302. https://doi.org/10.1080/002202799183133

Batchelor, O. (2017). Getting out the truth: The role of libraries in the fight against fake news. *Reference Services Review*, *45*(2), 143–148. https://doi.org/10.1108/RSR-03-2017-0006

Dwyer, C. P., Boswell, A., & Elliott, M. A. (2015). An evaluation of critical thinking competencies in business settings. *Journal of Education for Business*, *90*(5), 260–269. https://doi.org/10.1080/08832323.2015.1038978

Ennis, R. H. (1964). A definition of critical thinking. *The Reading Teacher*, *17*(8), 599–612.

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi*, *37*(1), 165–184. https://doi.org/10.1007/s11245-016-9401-4

Ennis, R. H., & Millman, J. (1985). *The Cornell Critical Thinking Test Level Z*. Midwest Publications.

Ennis, R. H., & Weir, E. (1985). *The Ennis-Weir critical thinking essay test: Test manual, criteria, scoring sheet*. Midwest Publications.

Facione, P. A. (1990). *The California Critical Thinking Skills Test-College Level. Technical report #2. Factors predictiveof CT skills*. California Academic Press.

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

Gola, C. H., Ke, I., Creelman, K. M., & Vaillancourt, S. P. (2014). Developing an information literacy assessment rubric: A case study of collaboration, process,

and outcomes. *Communications in Information Literacy*, *8*(1), 131–144. https://doi.org/10.7548/cil.v8i1.238

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. PubMed. https://doi.org/10.20982/tqmp.08.1.p023

Halpern, D. F. (2007). *Halpern Critical Thinking Assessment using everyday situations: Background and scoring standards*. Claremont McKenna College.

Holma, K. (2015). The critical spirit: Emotional and moral dimensions of critical thinking. *Studier i Pædagogisk Filosofi*, *4*(1), 17–28. https://doi.org/10.7146/spf.v4i1.18280

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. PubMed. https://doi.org/10.1016/j.jcm.2016.02.012

Kuhn, D. (1991). *The skills of argument*. Cambridge University Press.

Meltzoff, J., & Cooper, H. (2018). *Critical thinking about research: Psychology and related fields* (2nd ed.). American Psychological Association. https://doi.org/10.1037/0000052-000

Paul, R. (1981). Teaching critical thinking in the 'Strong' sense: A focus on self-deception, world views, and a dialectical mode of analysis. *Informal Logic*, *4*(2). https://doi.org/10.22329/il.v4i2.2766

Paul, R., & Elder, L. (2005). *A guide for educators to critical thinking competency standards*. http://www.criticalthinking.org/files/SAM_Comp%20Stand_07opt.pdf

Paul, R., & Elder, L. (2012). *The international critical thinking reading and writing test*. The Foundation for Critical Thinking.

Pawley, C. (2003). Information literacy: A contradictory coupling. *Library Quarterly*, *73*(4), 422–452. lls.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills*, *105*(3), 997–1014. https://doi.org/10.2466/pms.105.3.997-1014

Possin, K. (2008). A field guide to critical-thinking assessment. *Teaching Philosophy*, *31*(3), 201–228. https://doi.org/10.5840/teachphil200831324

Siegel, H. (1997). *Rationality redeemed? Further dialogues on an educational ideal*. Routledge.

Taube, K. T. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. *Journal of General Education*, *46*(2), 129–164.

Thayer-Bacon, B. J. (2000). *Transforming critical thinking: Thinking constructively*. Teachers College Press.

The Foundation for Critical Thinking. (2019). *International Critical Thinking Test*. https://www.criticalthinking.org/pages/international-critical-thinking-test/619

Watson, G., & Glaser, E. M. (1980). *Watson–Glaser Critical Thinking Appraisal: Forms A and B*. PsychCorp.

Woodford, K., & Bancroft, P. (2004). Using multiple choice questions effectively in information technology education. In R. Atkinson, C. McBeath, D. Jonas-Dwyer, & R. Phillips (Eds.), *Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference* (pp. 948–955). http://www.ascilite.org.au/conferences/perth04/procs/woodford.html

References to R packages used

Bernaards, C. A. and Jennrich, R. I. (2005) Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis, *Educational and Psychological Measurement, 65*, 676-696. http://www.stat.ucla.edu/research/gpa

Fox, J. and Weisberg, S. (2019). *An {R} companion to applied regression* (3rd ed.), Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gamer, M., Lemon, J., Fellows, I., and Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Kassambara, A. (2019). ggpubr: 'ggplot2' based publication ready plots. R package version 0.2.1. https://CRAN.R-project.org/package=ggpubr

Peters, G. (2018). _userfriendlyscience: Quantitative analysis made accessible_. R package version 0.7.2. doi: 10.17605/osf.io/txequ

Raiche, G. (2010). nFactors: An R package for parallel analysis and non graphical solutions to the Cattell scree test. R package version 2.3.3.1.

Ram, K. and Wickham, H. (2018). wesanderson: A Wes Anderson palette generator. R package version 0.3.6. https://CRAN.R-project.org/package=wesanderson

Revelle, W. (2018) psych: Procedures for personality and psychological research, Version 1.8.12., Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses, *Journal of Statistical Software, 17*(5), 1-25. http://www.jstatsoft.org/v17/i05/

Schafer, J., Opgen-Rhein, R., Zuber, V., Ahdesmaki, M., Duarte Silva, A. P., and Strimmer, K. (2017). corpcor: Efficient estimation of covariance and (partial) correlation. R package version 1.6.9. https://CRAN.R-project.org/package=corpcor

Wei, T. and Simko, V. (2017). R package "corrplot": Visualization of a correlation matrix (Version 0.84). https://github.com/taiyun/corrplot

Wickham, H. (2011). The Split-Apply-Combine Strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1-29. http://www.jstatsoft.org/v40/i01/

Wickham, H. (2017). tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Yee, T. W. (2013). Two-parameter reduced-rank vector generalized linear models. *Computational Statistics and Data Analysis*. http://ees.elsevier.com/csda.