Validation of the CEU-Lopez Critical Thinking Test Using Multidimensional IRT Model

Johnny T. Amora
De La Salle-College of Saint Benilde, Philippines

Marcos Y. Lopez
Centro Escolar University-Malolos, Philippines

Abstract

The CEU-Lopez Critical Thinking Test was developed as a measure of the critical thinking dimensions (Lopez, 2012b). The present study adds to the emerging literature on critical thinking by assessing the psychometric properties of The CEU-Lopez Critical Thinking Test using the three IRT models, namely: unidimensional, consecutive, and multidimensional models. Emerged in the present study is a shorter version of The CEU-Lopez Critical Thinking test, which is fitted statistically to the multidimensional IRT model. The results from the analyses of the five dimensions of the short version of The CEU-Lopez Critical Thinking Test provided evidence on the reliability, validity, and measurement precision of the Critical Thinking Test.

Keywords: Critical thinking, Unidimensional IRT model, Consecutive IRT model, Multidimensional IRT model.

Introduction

One among the twenty first century skills that are crucially essential to be cultivated by our learners is critical thinking which is not just necessary for the mastery of academic contents but more so for the execution of daily life-changing activities and work-related matters (Fisher, 2011; Halpern, 2014; Huber & Kuncel, 2015; Shaheen, 2016; Wagner, 2014). If critical thinking is a necessity in every aspect of human existence then educators need to do something on how this could be taught and tested effectively which supports what Ennis (2003, 2011) stated that much work needs to be done not only on how critical thinking can be taught but also in the field of assessment, an area that has not been given due attention in critical thinking research. This purports that there is a dearth of existing critical thinking tests and this is true in Asian context which can be borne out by the list of critical thinking tests prepared by Ennis and Chattin (2015) which shows that majority of the available tests specifically written in English language are developed by Western scholars from North America, United Kingdom, Canada and only one from Asia specifically designed for tertiary learners in the Philippines. Hence, the need to develop more sophisticated critical thinking tests designed for Filipinos and other Asian nationals has to be addressed.

In response to the call for the need to develop a critical thinking test, Lopez (2012a, 2012b) developed and validated a critical thinking test designed for Filipino learners in tertiary level. It is called The CEU-Lopez Critical Thinking Test (2012a) which consists of 87 items and is considered a multi-aspect general knowledge critical thinking test (Lopez & Asilo, 2014). Ennis, Millman, and Tomko (2005) explained that the said type of test uses content that is based on daily life experiences which are assumed familiar to the target examinees. The said test developed by Lopez adopted the conception of Ennis (1996) as regards aspects of critical thinking for purposes of item construction. Considering the relevance of critical thinking for individuals to become more functional citizens of the world, the need to assess it periodically is undeniable. If there is no valid and reliable test of critical thinking available, there will be problems on how the schools and other learning organizations can monitor the progress of their efforts on enhancing critical thinking of learners.

The CEU-Lopez Critical Thinking Test, which is a multiple-choice type of test that consists of 87 items, deals with five aspects of critical thinking, such as: deduction, credibility judgment, assumption identification, induction, and meaning and fallacies with 19, 17, 16, 16, and 19 items, respectively. Though each aspect was tested separately, in practice, they are interdependent (Ennis, 1996, 2011; Facione & Gittens, 2013; Fisher, 2011; Norris & Ennis, 1989). The test focuses only on aspects of critical thinking that can be tested objectively. The disposition aspects of critical thinking, which are attitudinal in nature, cannot be tested objectively (Ennis, 2003; Facione & Gittens, 2013). Hence, they were not directly covered as part of test items.

In addition, the test has been normed locally using the students in the three campuses of Centro Escolar University (CEU), namely: Manila, Makati, and Malolos as samples. The results of such study yielded score interpretations that are categorized into six, namely, unreflective thinker, challenged thinker, beginning thinker, practicing thinker, advanced thinker, and master thinker which are based on the concept of Paul and Elder (2001) as regards stages of critical thinking development of individuals. Three types of norms were generated such as university norms, curricular year level norms, and program type norms (science vs. non science). Though locally done, the said norms for score interpretation purposes can be adopted by other academic institutions that have similar academic setup with that of CEU. Other institutions that have entirely different academic structure with that of CEU can come up with their own norms for research and other purposes (Lopez, Mendoza, Lucero, Opina, 2014).

Concerning the internal consistency measure of reliability of the said test, KR 20 was used. The obtained reliability is .68 which is within the acceptable range for a critical thinking test considering that a critical thinking is a multidimensional construct. Reliability estimates that can be considered adequate tend to range from about .65 to .75 and may increase with the examinees' level of sophistication and that very high reliability on tests that deal with critical thinking should not be automatically regarded as better than more moderate ones (Norris & Ennis, 1989). Since The CEU-Lopez Critical Thinking Test is a mulitidimensional type of test, it can be expected that its internal-consistency index is lower than that of other unidimensional standardized tests (Lopez, 2012b).

In reference to its uses, The CEU-Lopez Critical Thinking Test can be utilized to determine the strengths and weaknesses of the students in certain aspects of critical thinking. The results of diagnosis can be used for the development of interventions or remedial measures regarding the enhancement of learners' critical thinking. Since the said test has been normed locally, it can be used to determine and categorize the level of critical thinking of the learners and can come up with a comparative study regarding the level of critical thinking of the learners who belong to different curricular year levels and different curricular areas like science or non-science-related courses. It can also be used as pretest and posttest in certain experimental study specifically in determining the effectiveness of infusing critical thinking into curriculum or to look into the comparison of effectiveness of the four approaches in the teaching of critical thinking, such as, general, infusion, immersion, and mixed. In addition, academic institutions may utilize the test as part of selective and retention process for courses in which board examinations are required as well as industries that value critical thinking as one of the most necessary and desirable attributes employees must possess in certain companies or some other organizations.

Furthermore, critical thinking experts from Association of Informal Logic and Critical Thinking (AILACT) were consulted regarding the content of the test and accuracy and clarity of items. Out of twenty-four recognized critical thinking consultants listed in AILACT website, three experts expressed willingness to review each item of the test. They were David Hitchcock (McMaster University, Hamilton, Canada), Dona Warren (University of Wisconsin-Stevens Point, Wisconsin, USA), and Susana Nuccetelli (St. Cloud University, Minnesota). Their insightful comments were looked into and considered thoughtfully as part of a number of series of revisions of test (Lopez & Asilo, 2014).

Regarding its construct validity, Norris' (1992) procedure in the use of verbal reports of thinking to determine the construct validity of each item was adapted. There were 13 verbal reports of thinking gathered for each test item. A total of 1,131 verbal reports of thinking were gathered with the additional 260 verbal reports of thinking for some items that called for further validation. The said verbal reports were tape-recorded and transcribed verbatim and carefully analyzed. Two types of scores were generated from this activity, such as, performance and thinking scores. The former is based on the examinees' selected answer from the three choices provided in each test item. The latter is based on the justification of the examinees as to why a certain option was their answer. The two scores were correlated to determine whether certain item needs to be retained, revised, or replaced in addition to the insights given by examinees based on their verbalized thoughts regarding test item (Lopez, 2013).

It is interesting to note that The CEU-Lopez Critical Thinking Test has been used by other Filipino researchers. Some of these studies were conducted by Agraan, Amado, Lumunsad, Manalo, and Vidad (2016), Galvez (2015), Gracia, Jose, Espiritu, Geronimo, Estrella, and Gulinao, (2013), Grafilo (2013), Moreno, Braza, De Villa, and Refugido (2016), Parico (2015), Tayao (2014), Reyes (2017), Tayao, Daez, Inigo, Cruz, Nievera, Francisco, and Avinante (2016), and Viray (2014) in which the said test was utilized as their measure in determining the critical thinking level of their students. Unfortunately, all of these researches did not report the reliability of the test for their gathered data. Though the reliability of the original 87-item of The CEU-Lopez Critical Thinking Test was analyzed (which yielded an overall KR-20 coefficient of .68), the reliability coefficient for each dimension was not reported. The present paper assesses the reliability of The CEU-Lopez Critical Thinking Testat the dimension level based on the IRT context.

The aforementioned researchers (i. e., those who utilized The CEU-Lopez Critical Thinking Test) were interviewed as regards their feedback in administering the test. They were one in saying that their examinees found the test too long and tiresome and might have the tendency not to take seriously the latter part of the test due to fatigue. Hence, it was deemed necessary to reduce the number of test items to make the test more manageable on the part of the examinees to finish answering the test without experiencing so much mental or physical exhaustion. The present paper attempts to develop a valid and reliable shorter version of The CEU-Lopez Critical Thinking Test based on the criteria of IRT.

It can be noted that the measurement properties of The CEU-Lopez Critical Thinking Test was established using the Classical Test Theory (CTT) approach. The use of CTT in test development has some limitations, which are mainly related to the validity and reliability of the results (Embretson & Reise, 2000; Singh, 2004). Some major limitations of CTT, which are described as circular dependency, are as follows: (1) Person's ability is test-dependent (i. e., for a fixed test, person's ability is high if the test is easy; and person's ability is low if the test is difficult); (2) item/test difficulty is group-dependent (i. e., item/test is easy if test takers have higher ability; and item/test is difficult if test takers have lower ability); and (3) Test-oriented (i. e., score is given at the test level, but there is no basis in determining how well a person perform on a particular item). According to Fan (1998), the circular dependency poses some theoretical difficulties in CTT's application in some measurement situations such as test equating and computerized adaptive testing. On the other hand, many researchers (e. g., Amora & Bernardo, 1999) stressed that it is difficult to judge whether a person with certain ability level will have a problem on a particular item because, in CTT, the items and persons are calibrated in different scales. All the weaknesses of The CEU-Lopez Critical Thinking Test that arise due to the limitations of the CTT can be fixed using the IRT and/or Rasch modeling approaches.

The present paper aims to analyze The CEU-Lopez Critical Thinking Test using the three IRT models, namely: multidimensional, unidimensional, and consecutive IRT models. The main goal is to find the best IRT model that fits The CEU-Lopez Critical Thinking Test. Then, based on the chosen best IRT model, a valid and reliable shorter version of The CEU-Lopez Critical Thinking Test is produced. The three IRT models are discussed in the next section.

Method

Participants

Served as respondents of the study were 1400 college students from the three campuses (CEU-Malolos, CEU-Makati, and CEU-Manila) of Centro Escolar University in the Philippines who enrolled in courses such as Medical and Health-related courses (e. g., Nursing, Medical Technology, Dentistry), Natural Sciences (e. g., Biology), Social Sciences (e. g., Psychology, Social Work), Education, Business-related courses (e. g., Accountancy, Business Administration), and Computer-related courses (e. g., Information Technology/Computer Science). Distribution of the samples taken was based on degree program, year level, gender, age, and SES (social economic status).

The Instrument

The instrument under study is the 87-item multiple-choice type Critical Thinking Test (Lopez, 2012a and 2012b) which consists of five interdependent dimensions of critical thinking such as deduction, credibility judgment, assumption identification, induction, and meaning and fallacies. Deduction dimension, which consists of 19 items, refers to several principles of critical thinking such as fallacy of affirming the consequent, fallacy of division, fallacy of bandwagon, modus ponens, contraposition, and post hoc fallacy. In the test, the examinees are asked to decide on the given argument or item based on the following options: If the underlined statement follows necessarily from the other statements given, mark letter A. If the underlined statement contradicts the other statements given, mark letter B. If the underlined statement neither follows necessarily nor contradicts the other statements given, mark letter C. A sample item is "People say that Sex Education is not effective in teaching elementary pupils to be reflective and reasonable individuals, so, it should be eliminated from the list of subjects being studied by this young people. Although they may be right in saying it, Sex Education needs to be taught to elementary pupils."

The credibility judgment dimension consists of 17 items. The criteria used in the item construction for judging credibility of sources and observation statements are expertise, lack of conflict of interest, agreement with other sources, reputation, careful habits, use of established procedures, ability to give reasons, minimal inferring involved, provision of records, and corroboration. In the test, each item has two characters who present two conflicting observation statements and examinees are requested to judge which of the two contradicting statements is more credible by choosing the following options: If you think the first statement is more believable, mark letter A. If you think the second statement is more believable, mark letter B. If neither statement is more believable than the other, mark letter C. A sample item is “Ace said, “Oliver and I visited the Population Census Office in municipal hall two weeks ago and the chairman of the Population Census informed us that there is a total population of 2,558 living in the barrio Maasin. Searching for a sheet of paper, from his pocket, Oliver said, “After our conversation with the chairman of the Population Census, I immediately jotted down the shared information that there is a total population of 2,988 living in the barrio of Maasin.”

The assumption dimension consists of 16 items. The types of assumption tested in this dimension are presupposition, needed assumption, and used assumption. Each item has one character who makes a proposition that is taken for granted in a situation and that supports a conclusion. Sample item: “Farmers are hardworking individuals, so people from this place must be hardworking. Which is most probably taken for granted?” The response options for such item are as follows: A. Farmers in this place really work hard. B. Hardworking people are usually farmers. C. People from this place are farmers.

The fifth dimension, which is called meaning and fallacies, consists of 19 items. In this dimension, the items deal with difference on the use of necessary and sufficient condition language, judging provided definitions, negation and double negation, such logical words as only, if and only if, some, all, and some fallacies such as non-sequitur, post hoc, straw person, and circularity. There are three options to choose from after every stem of the item. Sample item: “To say that a person is healthy, is to say that a person can move gracefully, work tirelessly, and think rationally. These are all characteristics of a healthy individual. Of the following, which is the best statement that comes closest to the definition of a healthy person? A. To move gracefully, work tirelessly, and think rationally are necessary but not sufficient characteristics of a healthy person. B. To move gracefully, work tirelessly, and think rationally are not necessary but sufficient characteristics of a healthy person. C. To move gracefully, work tirelessly, and think rationally are each a necessary component, with all three being jointly sufficient characteristics of a healthy person.”

The CEU-Lopez Critical Thinking Test was developed and validated by adapting the eight-phase test development model designed by Norris (1992). Included in the eight phases are: (1) Test conceptualization, (2) Development of a test plan, (3) Development of the test items, (4) Face and content validity of the test, (5) Revision of the test items, (6) Pre- try-out of the test, (7) Actual try-out of the test, and (8) construct validation of the test using verbal reports of thinking. The 87-item of The CEU-Lopez Critical Thinking Test yielded an overall KR-20 coefficient of .68. No reliability coefficients were reported for each dimension.

As discussed in the previous section, there are a number of studies that utilized The CEU-Lopez Critical Thinking Test. Unfortunately, all of these researches did not report the reliability of the test for their gathered data.

Data Analysis

The gathered data were fitted to the three different IRT models, namely: unidimensional, consecutive unidimensional, and multidimensional models. The goal is to determine the model that fits best The CEU-Lopez Critical Thinking Test. The unidimensional model is the standard Rasch/IRT model where unidimensionality is the fundamental assumption; that is, all the items of the instrument should measure one single trait (Lord, 1980). In this paper, the unidimensionality model is tested because one goal of The CEU-Lopez Critical Thinking Test is to measure the overall critical thinking of the students.

The consecutive approach (Davey and Hirsch, 1991), on the other hand, is simply a unidimensional model repeated a number of times using subsets of the full range of items on a given instrument (Briggs and Wilson, 2003). The consecutive unidimensional can be an appropriate modeling approach to analyze The CEU-Lopez Critical Thinking Test because the goal also of the test is to measure each of the five dimensions of critical thinking.

Multidimensional IRT model, or simply multidimensional model, simultaneously calibrates several dimensions and capture the complexity of tests that measure several traits (Adams et al., 1997; Kelderman, 1996; Rost & Carstensen, 2002; Yao & Schwarz, 2006; Baghaei, 2012). Multidimensional IRT model, also known as multidimensional random coefficients multinomial logit (MRCML) model, is an extension of the Rasch family of item response models (Briggs & Wilson, 2003). There are two types of multidimensional model, namely: between-item multidimensional and within-item multidimensional models (Adams et al., 1997). In the between-item multidimensional model, each item in the instrument belongs to only one dimension, while in the within-item multidimensional model, each item is designed to measure simultaneously more than one dimension. The present paper utilized the between-item multidimensional model. For simplicity, multidimensional model is used throughout this paper, instead of between-item multidimensional model. Multidimensional model is considered as another option to analyze The CEU-Lopez Critical Thinking Test because the test consists of five interrelated dimensions and the purpose of the test is to measure the five dimensions and then report them as separate scores of critical thinking performance and/or as a single critical thinking performance.

The parameters of the multidimensional model were estimated through the use of the ConQuest version 4.5.2 using the Monte Carlo method with 1000 nodes and .005 convergence. Monte Carlo method was used because The CEU-Lopez Critical Thinking Test consists of five interrelated dimensions. According to Adams and Wu (2010), the Monte Carlo method is generally the preferred approach for problems of more than three dimensions, so that the goodness-of-fit statistics are comparable across the three IRT models. The same estimation method (method=monte carlo, nodes=1000, convergence=.005) was utilized in the estimation of the parameters of the unidimensional and consecutive unidimensional models.

Results and Discussion

The gathered data of the 87-item of The CEU Lopez Critical Thinking Test were subjected to analysis using the multidimensional model and subsequently using both the unidimensional and consecutive dimensional models. The multidimensional model was performed in two steps using the Conquest software: (1) All the 87 items of The CEU Lopez Critical Thinking Test were subjected to analysis and then the statistically fitted items were identified from the results. (2) All the statistically fitted items identified in step 1 were subjected again to analysis. Step was repeated by including only the fitted items in the analysis. Stop the analysis if all items in step 2 are already fitted statistically. Statistically fitted items are those with mean square (MNSQ) unweighted or weighted fit statistics that are inside the 95% confidence interval or items with MNSQ unweighted or weighted fit statistics with absolute T-values of less than 2.0. Results of the analysis revealed that 56 of the 87 items of The CEU-Lopez Critical Thinking Test statistically fitted the multidimensional model. In this study, the resulting 56-item instrument is called the short version of The CEU-Lopez Critical Thinking Test. Subsequently, the short version was subjected to further analysis using both the unidimensional and consecutive unidimensional models. It is worth noting that all the 56 items of the short version statistically fitted with both models. The distribution of the 56 items of the short version across the five dimensions is shown in Table 2. Proceed to Table 5 for the item measures (or difficulty statistics) and item fit statistics of the short version.

Table 2
Distribution of the fit and misfit items of The CEU-Lopez Critical Thinking Test across dimensions

Dimension	Original version (# of items)	Short version (# of fitted items)	# of Misfitted items
1. Deduction	19	15	4
2. Credibility	17	8	9
3. Assumption	16	5	11
4. Induction	16	13	3
5. Meaning	19	15	4
Total	**87**	**56**	**31**

Model fit Statistics

Model fit statistics such as Akaike Information Criterion (AIC) and Deviance (G^2) of the multidimensional, unindimensional, and consecutive unidimensional models were computed. Generally, the smaller the AIC and G^2, the better the model fit. As shown in Table 3, the multidimensional model (AIC=87,633.99) has the smallest AIC among the three models, indicating that the multidimensional model fits better than the consecutive dimensional and unidimensional models. Because multidimensional model is hierarchically related to the unidimensional model, the model fit can be compared relative to the change in the deviance (G^2) value, where the difference in deviance between the two models is approximately distributed as a chi-square (Briggs & Wilson, 2003) and the degrees of freedom is the difference in the number of parameters of the models. As indicated in Table 3, the difference in deviance is statistically significant (χ^2=348.50, df = 14, p<0.00), indicating that the multidimensional IRT model significantly fits the data better as compared to both unidimensional and consecutive unidimensional models. The multidimensional model is expected to be a better model since it accounts for the interrelated dimensions of The CEU-Lopez Critical Thinking Test. As shown in Table 4, the five dimensions under multidimensional model are significantly and positively correlated (ranging between .06 and .45), except for deduction and credibility (r =.01, p>.05). Having a positive correlation coefficient indicates that higher scores on one dimension tend to be paired with higher scores on the other dimensions. Conversely, the lower the scores on one dimension so does the other dimensions. It is worth noting that of the five dimensions critical thinking, deduction and credibility emerged without significant correlation. This finding contradicts with the results of the previous studies (e. g., Ennis, Millman, & Tomko, 2005; Ennis, 1987, 1996) which connote that deduction and credibility judgments are correlated. The aforementioned authors argued that observation and credibility judgments call for the application of principles of deductive process. This suggests that when an individual is asked to give his judgment on the credibility of a person, deductive thinking may be applied. This may explain the correlation between deduction and credibility judgments. Moreover, there were no studies in the literature reporting that deduction and credibility are not correlated.

Table 3
Fit statistics of the three models

Models	**AIC**	**Deviance (G^2)**	**No. of Parameters**
Unidimensional	87,954.41	87,840.49	57
Consecutive Unidimensional	87,754.96	n/a	61
Multidimensional	87,633.99	87,491.99	71

Notes. The deviance difference between multidimensional and unidimensional models is chi-square distributed with 71-57 = 14 degrees of freedom: χ^2=348.50, df = 14, p<.000. Data are deviance and AIC of the short version of The CEU-Lopez Critical Thinking Test.

Table 4
Correlations between the dimensions of The CEU-Lopez Critical Thinking Test short version

	Deduction	Credibility	Assumption	Induction	Meaning
1. Deduction	1.00				
2. Credibility	.01	1.00			
3. Assumption	.06*	.35**	1.00		
4. Induction	.31**	.14*	.29**	1.00	
5. Meaning	.06*	.33**	.35**	.45**	1.00

Note. ** significant at .01 level; * significant at .05 level.

One of the advantages of the present study in comparison with the early development of the scale by Lopez (2012a), which utilized the classical test theory, is the model fit statistics. These model fit statistics play a very important role in the present study because IRT modelis a probabilistic model. Having a better model fit statistics implies that the model estimates (i.e., the results generated by the model, e.g., items measures) can be used in making inferences about the items of the scale and the scale as a whole with high degree of confidence, an important property of IRT model that cannot be provided by the classical test theory.

Item Fit Statistics

Table 5 presents the item difficulty for each of the test items of the short version along with the corresponding fit statistics such as MNSQ fit statistics, the 95% confidence interval for the expected value of the MNSQ, and the T-statistics. According to Wu, Adams, Wilson, and Haldane (2007), MNSQ fit statistics are residual-based indices that are similar in conception and purpose to the weighted and unweighted fit statistics that were developed by Wright and Stone (1979) for Rasch's simple logistic model and Wright and Masters (1982) for partial credit model. An MNSQ fit statistic of 1.0 indicates that the item conforms perfectly with the multidimensional model. If a test item has a MNSQ statistic that lies outside the corresponding confidence interval, then the test item does not conform with the multidimensional model. Moreover, if the MNSQ statistic lies outside the confidence interval then the absolute value of the corresponding T-statistic exceeds 2.0 (Wu, Adams, Wilson, & Haldane, 2007). As shown in Table 5, the MNSQ fit statistics of the 56 items lie within the 95 confidence interval and the absolute values of the T-statistics are less than 2.0, indicating that all the 56 items of the short version of The CEU-Lopez Critical Thinking Test fit statistically with the multidimensional model.

The item measures for the test items are presented in the second column of Table 5. An asterisk next to a parameter estimate indicates that it was constrained. One test item for each dimension needs to be constrained so that the mean of the item measures on each dimension is zero (Adams and Wu, 2010). Theoretically, the values of the item measures range between negative infinity and positive infinity. Practically, the values of the item measure range between -3.0 and +3.0.

The item-person map (Figure 1.0) depicts the plot of the persons (represented by Xs) and test items (represented by item numbers) with the five dimensions of the shorter version of The CEU-Lopez Critical Thinking Test. In IRT, both items and persons are calibrated on a common continuum based on the amount of traits possessed by each other (Bond & Fox, 2007). In the map of the present study, the values in the continuum approximately range from -2.0 to +2.0. The items are hierarchically arranged in terms of items measures from very easy (bottom) to very difficult (top), while the least able respondents are placed at the bottom and the most able at the top. It can be noted that the bunch of test items (around more than 57%) have difficulty measures that fall within the moderate levels. Only few items are very easy (5%), easy (10%), difficult (23%), and very difficult (4%). The ability levels of the majority of the respondents match with the below average test items.

The quality of the items of The CEU-Lopez Critical Thinking Scale was assessed using the item fit statistics. Items that did not qualify for the item fit statistics criteria were removed from the IRT model. Such quality criteria in selecting the good items for the shorter version of the scale is one of the advantages of the IRT modeling in comparison with the classical test theory utilized in the early development of the scale. The 56-item shorter version of The CEU-Lopez Critical Thinking Scale is an improvement of the 87-item original version which was developed based on the classical test theory. Unlike the longer version, the measures in the shorter version are item-free and person-free.

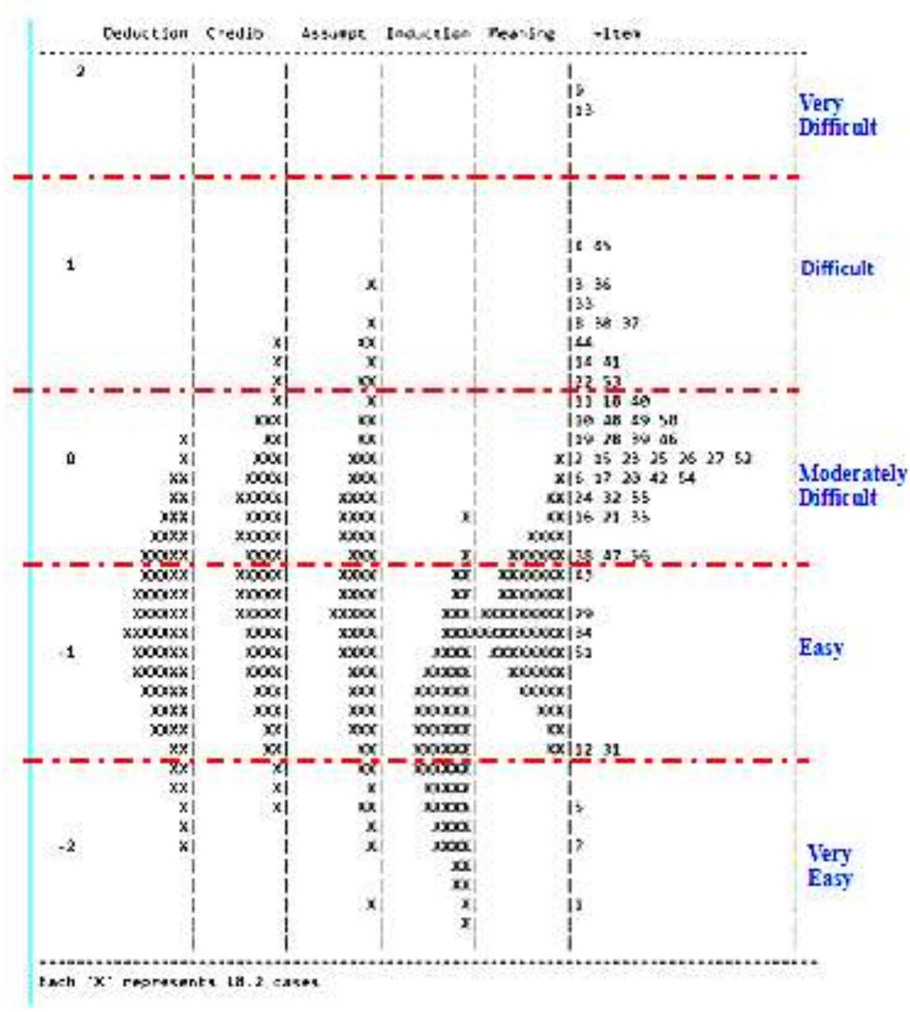
In terms of item difficulty, the 87-item original version have difficulty levels that range from easy to very difficult with a bunch of items (about 77%) falling within difficulty to very difficult levels, the 56-item short version have more dispersed difficulty levels ranging between very easy to very difficult with only few (27%) difficult and very difficult items. This means that a lot of difficult and very difficult items in the long version were pulled down to the moderately difficult items in the short version. This finding is not surprising since the IRT and CTT are two different methodologies.

Table 5

Item measure and item fit statistics of The CEU-Lopez Critical Thinking Test short version

Item#	Item Measure	MNSQ	95%CI	T-Value	**Item#**	Item Measure	MNSQ	95% CI	T-Value
Q01	-2.29	1.00	(.93,1.07)	0.10	**Q53**	-0.81	1.04	(.96,1.04)	1.80
Q03	0.05	1.00	(.95,1.05)	-0.10	**Q54**	0.74	1.00	(.87,1.13)	0.00
Q04	0.97	1.01	(.90,1.10)	0.20	**Q56**	-1.54	1.02	(.98,1.02)	1.80
Q05	1.11	1.00	(.89,1.11)	0.00	**Q57**	-0.16	0.99	(.93,1.07)	-0.20
Q06	-1.78	1.03	(.95,1.05)	1.20	**Q58**	0.83	1.00	(.87,1.13)	0.00
Q08	-0.11	1.00	(.96,1.04)	-0.20	**Q59**	-0.89	1.00	(.97,1.03)	-0.20
Q09	-1.96	1.00	(.94,1.06)	0.10	**Q60**	-0.30	1.03	(.94,1.06)	0.90
Q10	0.73	1.00	(.91,1.09)	0.10	**Q61**	0.97	1.00	(.85,1.15)	0.00
Q11	1.95	1.02	(.81,1.19)	0.20	**Q62**	0.72	1.00	(.87,1.13)	0.00
Q12	0.21	0.99	(.94,1.06)	-0.40	**Q65**	-0.53	0.98	(.95,1.05)	-1.00
Q13	0.32	1.02	(.94,1.06)	0.70	**Q66**	0.12	1.00	(.92,1.08)	0.00
Q16	-1.51	1.04	(.96,1.04)	2.30	**Q67**	0.36	1.00	(.90,1.10)	-0.10
Q17	1.83	1.00	(.82,1.18)	0.00	**Q68**	0.483*	1.01	(.89,1.11)	0.10
Q18	0.51	0.97	(.93,1.07)	-0.80	**Q70**	-0.05	0.98	(.96,1.04)	-0.90
Q19	-0.009*	0.98	(.95,1.05)	-0.70	**Q71**	-0.56	0.99	(.98,1.02)	-1.10
Q20	-0.26	1.00	(.97,1.03)	0.00	**Q73**	0.61	1.01	(.92,1.08)	0.40
Q21	-0.07	1.01	(.96,1.04)	0.40	**Q75**	1.10	0.99	(.89,1.11)	-0.10
Q22	0.27	1.02	(.95,1.05)	0.80	**Q77**	0.13	0.99	(.95,1.05)	-0.30
Q23	0.10	0.97	(.96,1.04)	-1.40	**Q78**	-0.50	0.99	(.98,1.02)	-0.50
Q24	-0.11	0.98	(.96,1.04)	-1.30	**Q79**	0.17	1.03	(.95,1.05)	1.00
Q29	-0.32	1.02	(.97,1.03)	1.10	**Q80**	0.21	1.01	(.95,1.05)	0.20
Q30	0.38	1.00	(.95,1.05)	0.10	**Q81**	0.18	1.02	(.95,1.05)	0.80
Q35	0.006*	1.00	(.96,1.04)	0.20	**Q82**	-0.97	1.01	(.98,1.02)	0.80
Q37	-0.15	1.01	(.96,1.04)	0.50	**Q83**	0.02	1.01	(.95,1.05)	0.20
Q39	0.00	1.00	(.96,1.04)	0.10	**Q84**	0.42	1.01	(.93,1.07)	0.20
Q40	0.02	1.02	(.96,1.04)	0.90	**Q85**	-0.11	1.00	(.96,1.04)	-0.10
Q44	0.02	0.99	(.96,1.04)	-0.50	**Q86**	-0.19	1.00	(.96,1.04)	-0.20
Q46	0.113*	0.99	(.95,1.05)	-0.40	**Q87**	-0.467*	1.01	(.98,1.02)	0.80

Note: 95%CI = 95% Confidence Interval; MNSQ = Mean square fit statistic; the 95%CI, MNSQ, and T-value are based on the weighted fit statistics.



```
        Deduction  Credib    Assumpt  Induction  Meaning    +Item
--------------------------------------------------------------------------
   2              |          |          |          |            |
                  |          |          |          |            |9
                  |          |          |          |            |13
                  |          |          |          |            |
                  |          |          |          |            |
                  |          |          |          |            |
                  |          |          |          |            |
                  |          |          |          |            |
                  |          |          |          |            |6 45
   1              |          |          |          |            |
                  |          |         X|          |            |3 36
                  |          |          |          |            |33
                  |          |         X|          |            |8 38 37
                  |         X|        XX|          |            |44
                  |         X|         X|          |            |14 41
                  |         X|        XX|          |            |22 53
                  |         X|         X|          |            |11 18 40
                  |       XXX|        XX|          |            |10 48 49 50
                 X|        XX|        XX|          |            |19 28 39 46
   0             X|       XXX|       XXX|          |           X|2 15 23 25 26 27 52
                XX|      XXXX|       XXX|          |           X|6 17 20 42 54
                XX|     XXXXX|      XXXX|          |          XX|24 32 55
               XXX|      XXXX|      XXXX|         X|          XX|16 21 35
              XXXX|     XXXXX|      XXXX|          |        XXXX|
             XXXXX|      XXXX|       XXX|         X|      XXXXXX|38 47 56
             XXXXX|     XXXXX|      XXXX|        XX|     XXXXXXX|43
            XXXXXX|     XXXXX|      XXXX|        XX|     XXXXXXX|
            XXXXXX|     XXXXX|     XXXXX|       XXX|XXXXXXXXX   |29
           XXXXXXX|      XXXX|      XXXX|       XXXXXXXXXXXXXXXX|34
  -1        XXXXXX|      XXXX|      XXXX|      XXXX|    XXXXXXXX|51
            XXXXXX|      XXXX|       XXX|     XXXXX|      XXXXXX|
             XXXXX|       XXX|       XXX|    XXXXXX|       XXXXX|
              XXXX|       XXX|       XXX|    XXXXXX|         XXX|
              XXXX|        XX|       XXX|    XXXXXX|          XX|
                XX|        XX|        XX|    XXXXXX|          XX|12 31
                XX|         X|        XX|    XXXXXX|            |
                XX|         X|         X|     XXXXX|            |
                 X|         X|        XX|     XXXXX|            |5
                 X|          |         X|      XXXX|            |
  -2             X|          |         X|      XXXX|            |7
                  |          |          |        XX|            |
                  |          |          |        XX|            |
                  |          |         X|         X|            |1
                  |          |          |         X|            |
                  |          |          |          |            |
==========================================================================
Each 'X' represents 18.2 cases
```

Figure 10. Person Item Map

Reliability of the The CEU-Lopez Critical Thinking Test

As discussed in the previous section, the model that fits best the CEU-Lopez Critical Thinking Test is the multidimensional IRT model. As shown in Table 6, the model that yields the large reliability coefficients for each of the five dimensions is the multidimensional IRT model. The reliability coefficients vary from the lowest of .60 for credibility dimension to the highest of .69 for the induction dimension. Such reliability coefficients are not far from the .68

overall reliability coefficient of the original version of The CEU-Lopez Critical Thinking Test.

The findings of the present study indicate that each of the five dimensions of The CEU-Lopez Critical Thinking Test is reliable. The scores yielded by The CEU-Lopez Critical Thinking Test are definitely not reliable if both conservative and unidimesional IRT models are utilized. The multidimensional IRT model is expected to yield the more reliable coefficients as compared to the conservative and unidimensional models because the former is the appropriate model for the five interdependent dimensions of the scale. The appropriateness of the multidimensional model for the scale under investigation is supported by the model fit statistics as discussed in the previous section. On the other hand, result of the present study in the aspect of reliability coefficient per dimension is an additional information about the psychometric properties of the scale since its early development reported only the reliability coefficient of the overall scale.

Table 6

Reliability of The CEU-Lopez Critical Thinking Test

Dimension	#of Items	Multidimensional IRT	Consecutive IRT
Deduction	15	.66	.34
Credibility	8	.60	.38
Assumption	5	.61	.37
Induction	13	.69	.32
Meaning	15	.64	.25
Total	56		

Unidimensional = .46; KR-20 (original version) = .68

Conclusion

The present study provides additional evidence on the reliability and validity of The CEU-Lopez Critical Thinking Test as a measure of the five critical thinking dimensions within the context of item response theory. The CEU-Lopez critical thinking test was successfully reduced from its original version with 87 items to the shorter version with 56 items. The five-interrelated dimensions of the shorter version of The CEU-Lopez critical thinking test can be best modelled using the IRT multidimensional model rather than the IRT unidimensional and consecutive models. Results of the analysis using the multidimensional IRT model show that the shorter version of the test has sound psychometric properties with item measures ranging between easy and difficult, where more than 55% of the items have average difficulty level. Further study about the shorter version may be conducted at the item level (e. g., item bias). On the basis of the findings of the present study, researchers are encouraged to use the shorter version of The CEU-Lopez Critical Thinking Test as a measure of the critical thinking dimensions.

References

Adams, R. and Wu, M. (2010). *Multidimensional models. Chapter 10 of the notes and tutorial.* ACER ConQuest Version 4 [computer software]. Camberwell: Australian Council for Educational Research.

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Agraan, S. M., Amado, A. K., Lumunsad, K. R., Manalo, G. A., & Vidad, M. L. (2016). *An assessment of critical thinking skills among radiologic technology students of De La Salle Health Sciences Institute* (Unpublished undergraduate thesis). De La Salle University-Dasmarinas, Cavite, Philippines.

Amora, J. T., & Bernardo, A. S. (2009). Testing and reducing L2 vocabulary learning strategies inventory using Rasch model. *Philippine ESL Journal, 3,* 38-73

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology*, *10*, 233-252.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch Model: Fundamental measurement in the human sciences (2nd ed.). NJ: Lawrence Erlbaum.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Davey, T., & Hirsch, T. M. (1991). Concurrent and consecutive estimates of examinee ability profiles. Paper presented at the annual meeting of the Psychometric Society, New Brunswick, NJ

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for psychologists. Mahwah, NJ: Erlbaum.

Ennis, R. H. (2011). Critical thinking. Reflection and perspective Part II. *INQUIRY: Critical Thinking Across the Disciplines, 26*(2), 5-19. doi:10.5840/inquiryctnews201126214

Ennis, R. H. (2003). Critical thinking assessment. In D. Fasko Jr (Ed.), *Critical thinking and reasoning* (pp. 293-313). NJ: Hampton Press, Inc.

Ennis, R. H. (1996). *Critical thinking.* Upper Saddle River: NJ: Prentice Hall.

Ennis, R. H., & Chattin, G. S. (2015, May 3). An annotated list of English-language critical thinking tests [Updated list of available critical thinking tests written in English]. Retrieved from www.criticalthinking.net/TestListDraft050315.docx

Facione, P., & Gittens, A. G. (2013). *Think Critically.* USA: Pearson Education, Inc.

Fisher, A. (2011). *Critical thinking: An introduction.* UK: Cambridge University Press.

Galvez, V. B. (2015, April). *The relationship between the level of critical thinking skills of B.S. Psychology students of Centro Escolar University-Malolos and their academic performance in professional Psychology subjects.* Paper presented at the annual research presentation of faculty of Centro Escolar University, Malolos, Philippines.

Garcia, M. A., Jose, S. J., Espiritu, K., Geronimo, A., Estrella, C., & Gulinao, A. (2013). *The relationship of critical thinking skills and Mathematics performance of freshmen and sophomore Bachelor in Secondary Education major in Mathematics students of polytechnic University of the Philippines-Sta. Maria, Bulacan campus.* (Unpublished undergraduate thesis). Polytechnic University of the Philippines-Sta. Maria, Bulacan, Philippines.

Grafilo, E.B. (2013). *Critical thinking and reading competence of senior high school students.* (Unpublished master's thesis). Bulacan State University, Malolos, Philippines.

Halpern, D. F. (2014). *Thought & knowledge: An introduction to critical thinking.* NY: Psychology Press.

Huber, C. R., & Kuncel, N. R. (2015). Does college teach critical thinking? A meta-analysis. R*eview of Educational Research, 20*(10), 1-38. doi: 10.3102/0034654315605917

Kelderman, H. (1996). Loglinear multidimensional item response theory models for polytomously scored items. In R. K. Hambleton, & W. J. van der Linden (Eds.), *Handbook of Modern Item Response Theory* (pp. 287-304). New York: Springer Verlag.

Lopez, M. Y. (2013). Determining the construct validity of a critical thinking test. *Educational Measurement and Evaluation Review, 4,* 87-99.

Lopez, M. Y. (2012a). *The CEU-Lopez Critical Thinking Test.* Research and Evaluation Office, Centro Escolar University, Manila.

Lopez, M. Y. (2012b). *The CEU-Lopez Critical Thinking Test Manual.* Research and Evaluation Office, Centro Escolar University, Manila.

Lopez, M. Y., & Asilo, M. V. (2014). Development and validation of The CEU-Lopez Critical Thinking Test. *Inquiry: Critical Thinking Across the Disciplines, 29(1),* 32-55.doi: 10.5840/inquiryct20142914

Lopez, M. Y., Mendoza, E. R., Lucero, R. D., & Opina, A. S. (2014). Establishing the local norms of The CEU-Lopez critical thinking test. *E-Journal of European Academic Research, 2(3),* 3895-3928.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Moreno, R., Braza, M., De Villa, R. D., & Refugido, K. M. (2016). *Intervening factors among grade 10 students' level of critical thinking skills in Mathematics and Science at Manuel S. Enverga University Foundation-Candelaria, Inc.* (Unpublished undergraduate thesis). Manuel S. Enverga University Foundation-Candelaria, Inc., Quezon Province, Philippines.

Norris, S. P. (1992). A demonstration of the use of verbal reports of thinking in multiple-choice critical thinking test design. *Alberta Journal of Educational Research, 38,* 155-176.

Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking.* Pacific Grove, CA: Midwest Publications.

Paul, R., & Elder, L. (2001). *Critical thinking: Tools for taking charge of your learning and your life.* Upper Saddle River, NJ: Prentice Hall.

Reyes, E. C. (2017). *Critical thinking and academic performance of junior high school students: Basis for the proposed instructional model in Chemistry* (Unpublished doctoral dissertation). Centro Escolar University, Mendiola, Manila.

Rost, J., & Carstensen, C.H. (2002). Multidimensional Rasch Measurenment via Item Component Models and Faceted Designs. *Applied Psychological Measurement, 26*(1), 42-56

Shaheen, N. (2016). International students' critical thinking-related problem areas: UK university teachers' perspectives. *Journal of Research in International Education, 15*(1), 18-31. doi: 10.1177/1475240916635895

Singh, J., 2004. Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research57* (2), 184–208.

Tayao, J. C. (2014). An assessment on the level of critical thinking skills of BSIT students at Centro Escolar University-Malolos in relation to their Computer Programming and Mathematics academic performance. *E-Journal of European Academic Research, 2*(3), 4394-4411.

Tayao, J. C., Daez, F. D. P., Inigo, A. N., Cruz, F. D., Nievera, M. L. T., Francisco, N. V., & Avinante, M. D. P. (2016, March). *The relationship of critical thinking skills and decision-making skills of selected Higher Education Institutions' senior Management Accounting, Finance, and Accounting Technology students in the province of Bulacan, Philippines.* Paper presented at International Conference on Business and Social Sciences (ICBASS2016) at Kyoto Research Park, Kyoto, Japan.

Wagner, T. (2014). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need-and what we can do about it.* New York: Basic Books, A Member of Perseus Books Group.

Wright, B. D., & Masters, G. N. 1982. *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press

Wright, B. D., & Stone, M. 1979. *Best Test Design: Rasch Measurement.* Chicago: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S.A. (2007). ACER ConQuest Version 2: Generalised item response modelling software [computer program]. Camberwell: Australian Council for Educational Research.

Yao, L., & Schwarz, R.D. (2006). A Multidimensional Partial Credit Model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30,* 469-492.