Educational Measurement and Evaluation Review (2013), Vol. 4, 87-99


# Determining the Construct Validity of a Critical Thinking Test

**Marcos Y. Lopez**
*Centro Escolar University-Malolos, Philippines*

**Abstract** This study deals with verbal reports of thinking in relation to determining the construct validity of each item of *The CEU-Lopez Critical Thinking Test* (Lopez, 2012) which is multi-aspect general knowledge critical thinking test designed for Filipino tertiary students. The procedure is based on the adapted methodology of Norris (1992) in validating his co-authored*Test on Appraising Observations* (Norris & King, 1983). Seven methodological phases were utilized in this study: determining the participants who would be part of sets of interviewees, adapting an interviewing methodology for eliciting trustworthy reports of thinking, collecting verbal reports of thinking, scoring verbal reports of thinking and choice of answer, comparing performance and thinking scores as basis for judging items, modifying suspected items, and retrying the revised or the replaced items in accord with steps 3 to 6. Items were revised, retained, and discarded based on the correlation of thinking and performance scores together with looking into the relevant insights of verbal reports of thinking given by examinees to establish construct validity of the test. It is concluded that verbal reports of thinking are useful in establishing the construct validity of a multi-aspect general knowledge critical thinking test.

**Keywords:** validity, test, construct, critical thinking, verbal reports

Norris (1992) said that evidence on the thinking processes of the examinees used to answer critical thinking test using multiple-choice format has a direct bearing on construct validity. Ebel and Frisbie (1991) defined construct validation as the process of gathering evidence to support the notion that a given test item really measures the psychological construct the test developers intend to measure. Alderson, Clapham, & Wall (1995) stated that construct validity is a form of test validation which fundamentally involves assessing to what extent each item in the test is a successful operationalization of the theory. Verbal reports of thinking can be

considered a relevant process to gather direct evidence that a test item is a successful operationalization of critical thinking construct.

The verbal reports of thinking collected for this study helped determine if a trial item measures what it is supposed to measure. If choosing the unkeyed option is associated with thinking well and choosing the keyed option is associated with thinking poorly, then the item is problematic. Conversely, if choosing the keyed option is associated with thinking well and choosing the unkeyed option is associated with poor thinking then the item is considered valid(Norris, 1992). However, these two premises are subject to empirical investigation by collecting a pool of verbal reports from a number of examinees for every test item of *The CEU-Lopez Critical Thinking Test*. Based on previous researches, verbal reports of thinking explicitly show how the test item works as intended by test developer.

This study examined the construct validity of the 87 trial items of *The CEU-Lopez Critical Thinking Test* (2012), a multi-aspect general knowledge critical thinking test measuring five aspects of critical thinking: deduction, credibility judgment, assumption identification, induction, and meaning. Specifically, verbal reports of thinking of the students were looked into to investigate the processes of thinking of examinees that led them to their chosen answers using Norris' (1988; 1989; 1990) argument that verbal reporting does not alter the course of thought and chosen answer of examinees while answering a multiple-choice critical thinking test.

Its main objective is to determine which item needs to be retained, revised, and discarded.

## Method

### Participants

A total of 2,412 students from all branches of Centro Escolar University (Malolos, Mendiola, Makati) took the experimental version of *The CEU-Lopez Critical Thinking Test* using paper-and-pencil format. Only students enrolled in courses with board examinations in all curricular year levels were considered part of research participants. A stratified random sampling was employed in which one intact class for every curricular year level of every course was randomly selected as samples.

### Procedures and Data Analysis

This presents the methodology on how the construct validity of the 87-item test was established using verbal reports of thinking. The seven phases of research methodology were based on adapted procedure of Norris (1992) in determining the construct validity of his co-authored test entitled *Test on Appraising Observations* (Norris & King, 1983). These methodological phases were: determining the participants who would be part of sets of interviewees, adapting an interviewing methodology for eliciting trustworthy reports of thinking, collecting verbal reports of thinking, scoring verbal reports of thinking and choice of answer, comparing performance and thinking scores as basis for judging items, modifying suspected answers, retrying the revised or the replaced items in accord with steps 3-6.

**Phase 1: Determining the participants who would be part of sets of interviewees**. These students were informed two days before test administration in order to pre-condition them mentally and psychologically that the test is a preparation for their future board examinations. It was done as a motivation for them to take the test seriously so that the validity of test results would not be adversely affected

The entire test can be thematically subdivided into eight parts, wherein, one student in a group which composed of eight students wasasked to verbalize his/her thoughts on the assigned number of items allocated to him/her. There are 13 groups included in this study. Hence, a total of 104 students who were chosen randomly from all CEU campuses participated in giving their verbal reports of thinking. Since these students came from different parts of the Philippines with different mental abilities, levels of maturity, interest, and culture, it is believed that they represented individuals who have different background beliefs, critical thinking sophistication, and intellectual ability that might have relevance to determine whether an item be retained or needs revision, modification, or even deletion based on their verbal reports of thinking.

**Phase 2: Adapting an interviewing methodology for eliciting trustworthy reports of thinking.** For interviewing methodology, the think aloud procedure in eliciting verbal reports of thinking was adopted. Norris (1990) described think aloud as a method for the elicitation of verbalized thoughts in which the subjects were instructed to verbalize all their thoughts as they answer the items assigned to them and to mark their answers on the provided scantron sheet.

Two stages of interview were adapted using standardized procedure espoused by Norris (1992). In the first stage, the interviewer informed the examinees of the main purpose of interviews and the procedure on how the interview would be conducted. The examinees were informed that their verbalized thoughts would be tape-recorded. Each examinee answered all items in the test but there were selected items that need to be answered through verbal reports of thinking.

In second stage, the interviewer asked examinees to say everything that goes on in their mind as they answer the item. The interviewer can only interrupt if there was ambiguous reference of demonstratives or third person pronouns as examinees explain their chosen answer. Interruptions can also be done to probe for obvious reading mistakes of the questions. No additional information provided to examinees if they asked for reasons or facts provided in the question. This was done to determine the clarity of the question and the possibility that the question is problematic.

All research interviewees were aware that their verbalized thoughts were tape-recorded.

**Phase 3: Collecting verbal reports of thinking.** Thirteen verbal reports of thinking were collected for each of the 87 items of the given test. Hence, a total of 1,131 verbal reports of thinking were collected excluding the other 260 verbal reports of thinking for the items that were revised for further validation.

Each examinee was assigned a certain number of items for him/her to verbalize his/her thoughts as the examinee worked through the items. The number of items

assigned to be answered verbally by each examinee was based on the logical break of the story line. An interview and paper-and-pencil formats were interlaced as a procedure to collect verbal reports of thinking.

Once the first student in a group is done with answering the first part of the test through thinking aloud the student would continue answering the rest of the items through paper-and-pencil format. The next student assigned to answer the second logical break of the story line of the test should finish answering the first part of the test through paper-and-pencil format and continue answering the next set of items through thinking aloud and would continue answering the rest of the parts through paper-and-pencil format. This procedure was done up to the last logical break of the set of items.

The interlacing of thinking aloud and paper-and-pencil formats was done to make it not too tiresome for every randomly selected student to answer the 87-item test through thinking aloud.

**Phase 4: Scoring verbal reports of thinking and choice of answer.** All tape-recorded verbal reports were transcribed verbatim and analyzed carefully. Two types of scores were assigned for each item: performance and thinking scores. These two types of scores were considered empirically and conceptually independent. Hence, it is possible that the student got the correct answer but gave an unjustifiable reason for the chosen answer. Conversely, student may arrive at the unkeyed option by giving justifiable reason.

The scoring allotted for performance score is 0 and 1. Zero for choosing the unkeyed option and a score of 1 for choosing the keyed option. The same scoring was allotted for thinking score: 0 and 1. Zero for a reason considered unjustifiable and a score of 1 for a justifiable reason. The justifiability of reasons was based on the approximation of the critical thinking principles and criteria from the critical thinking taxonomy of Ennis (1987,1996, 2011a) as used by the students when giving their verbal explanation for their chosen answer.

Performance score is based on the answer chosen from the three options provided in every item of the testwhereas the thinking score is based on the justification of the student for his chosen answer.

**Phase 5: Comparing performance and thinking scores as basis for judging items.** The relationship between performance and thinking scores for each item across subjects sampled was determined as basis for retention, revision or replacement of item. The two scores were correlated through point biserial correlation. The relationship between two scores should be strong for each other as basis to consider that choosing the keyed answer is associated with thinking critically and choosing the unkeyed answer is associated with thinking uncritically. If found that there is a low or no correlation between performance and thinking scores, then it serves as basis for item revision or item replacement. This purports that the student chose the unkeyed answer by thinking well or student thinks uncritically yet arriving at the keyed answer. Conversely, if there is a correlation between performance and thinking scores, then it can serve as a basis that the item be retained. This implies

that the student chose the keyed answer by thinking critically by invoking critical thinking criteria or principles.

**Phase 6: Modifying suspected items.** Items with low and no correlation between performance and thinking scores were looked into as basis for item revision or item replacement. The contents of transcribed verbal reports served as a guide on how revisions should be done or whether an item is problematic and needs replacement. Revision of items usually includes changes in words used and addition, deletion, and changes of information both in item stem and given options of certain questions.Hence, thinking and performance correlations along with transcribed verbal reports of thinking were used as basis for retention, modification, or replacement of problematic items.

**Phase 7: Retrying the revised or the replaced items in accord with steps 3-6.** Verbal reports of thinking were collected, transcribed, and analyzed to examine the revised and replaced items. The same scoring procedure was done for performance and verbal reports of thinking for each modified and replaced items. Items were considered valid when there is high correlation between performance and thinking scores. Revised items that did not have correlation between performance and thinking scores were repeatedly modified and underwent comparison between performance and thinking scores till all revised and replaced items have high correlation regarding scores on thinking and performance.

## Results

### Proportion of Accepted Items

Out of 87 trial items, 59 items were accepted without revision and twenty eightitems were accepted but with revision. Noitem was replaced with new one. The use of verbal reports of thinking in a multiple-choice type of critical thinking test determines whether items be retained, modified, and discarded. Hence, construct validity of the test items was established.

### Performance and Thinking Scores of Selected Items

Results of the verbal reports of thinking from selected four items from *The CEU-Lopez Critical Thinking Test* are presented below:

Item 17shown below is categorized as deduction in which post hoc fallacy isinstantiated. The said item is in a context of debate in which the speaker is presenting his argument whether fraternity be banned or allowed in universities and colleges. The examinee decides whether the underlined statement follows necessarily from or contradicts the other statements given or neither.

**Item 17:** The speaker is Congressman Salisi who argues that: *"A month after the imposition of dealth penalty, a group of researchers from one neighboring Asian country conducted a study regarding its effectiveness. They found that the crime rate in their country drops by about two percent in a matter of 30 days immediately after its imposition. This only shows that death penalty is effective.*

The following is a set of pairs of scores on performance and thinking of all students who verbalized their thoughts in item 17 : [ (0,0) (1,1) (0,0) (0,0) (0,0) (0,0) (0,0) (0,0) (0,0) (0,0) (0,0) (0,0) (1,1) ]. Noticeably, the thirteen pairs of performance and thinking scores of the said item coincided.The computed point biserial correlation coefficient between performance and thinking scores for item 17 is r= 1.0 (p <.01). Hence, the item was retained as is. This purports that an item works as desired implying high construct validity.

Item 23 which is given below deals with credibility judgment. Two contradictory statements are presented and examinees should decide which statement is more credible.

**Item 23:**
*A. Ace who was curiously looking at the unexplained movements in the nearby rice-field, said, "There is a pair of rats that hastily entered a small hole near that paddy."*
*B. Tying his shoelace, Oliver who was interminttenly looking at the same thing, which Ace has seen also, said " That is just one rat and not a pair of rats that immmediately entered a small hole near that paddy".*
*C. Neither statement is more believable*

The performance and thinking scores of students on this item is given below with seven mismatches between performance and thinking scores either combinations (0,1) or (1,0).

[ (0,0) (0,1) (0,0) (0,1) (0,0) (0,1) (0,1) (0,1) (0,0) (1,0) (0,0) (1,1) (0,1) ]

The computed point biserial correlationbetween performance and thinking scores for item 23 is r=.-.032 (n.s.) which suggests that there is a negative and low correlation between performance and thinking scores. This negative and low correlation implies that the item is problematic, thus, needs revision and lacking of construct validity.

Item 67 is categorized as induction item in which each item focuses on explanatory conclusion, specifically causal claims. For every conclusion, examinees decide whether the given information supports the conclusion, goes against the conclusion, and neither.

*67. A group of people from a competing company that also manufactured the same formulation claimed that the ABC pellets dried up rice plants in one of the farms somewhere in the province of Batanes. They concluded that ABC pellets are not safe to rice plants.*

A set of accumulated pairs of performance and thinking scores is given below for item 67.

[ (1,1) (0,0) (0,0) (0,0), (0,0) (1,1) (1,1) (1,1) (0,0) (1,1) (0,0) (1,1) (1,1) ]

The computed point biserial correlation between performance and thinking scores for item 67 is r= 1.0 ($p < .01$) which implies that the item has construct validity. However, it does not mean that the item is free from any problems and does not need any modification. Item 67 exemplifies that even though a correlation is high, it possible that an item may need revision due to the use of some words in test item that lead examinees to easily guess the correct answer.

Item 87 is part of the last aspect of the test that deals with meaning and fallacies. There are three options to choose from after every stem of the item.

**87. MARY: If it is not true that Filipinos do not work hard enough then it is false that our country will not become more progressive one. It is not the case that Filipinos do not work hard.**

**OLIVER: You keep on saying that our country will not become progressive one. Are you not having faith in Filipinos' determination and hard work to make this country a better one? Could you explain why you said so?**

**Pick the one best reason why some of this thinking is faulty.**

**A. Mary is offering a proposition she is arguing for as a reason for itself.**
**B. Oliver misdescribed and challenged the position of Mary.**
**C. Mary's and Oliver's conclusions do not necessarily follow from their respective reasons.**

A set of accumulated pairs of scores between performance and thinking for this item is shown below.

[ (1,0) (1,1) (1,1) (0,0) (0,0) (0,0) (1,0) (0,0) (1,1) (1,0) (1,1) (1,1) (1,0) ]

The computed point biserial correlation between performance and thinking scores for this item is r=.53 ($p <.01$) which shows moderate correlation. As shown above, there are four pairs of scores that are mismatched.

## Discussion

The construct validity of an item was established using verbal reports of thinking as espoused by Norris (1988,1990,1992). An item that was found to have construct validity through verbal reports of thinking is given below.

The following is the verbatim transcription of verbal reports of thinking of Student A with its English translation in parenthesesoriginally given in Filipino or Taglish (combination of Tagalog and English) on item 17.

Student A: *Letter A kasi based on what the speaker said ahhh… yun talaga yung justification kung bakit yung ano po death penalty is effective. Kasi after nung inimposeyung death penalty in a matter of thirty days daw bumaba yung crime rate ng bansa. So clearly parang yun din yung, death penalty nga ang dahilan kung bakit ahhh… bumababa yung crime rate. I mean effective siya. Kasi nga after nung inimpose 'yon dun sa bansa, bumaba yung percentage ng crime rate nila. So letter A.*

(*Letter A because based on what the speaker said that is the real justification why death penalty is effective. Because after the imposition of death penalty in a matter of thirty days crime rate of the country goes down. So clearly, death penalty is the reason of the decrease of crime rate, I mean that is effective. Because after its imposition the percentage of crime rate goes down. So letter A.*)

Student A chose the unkeyed response by thinking poorly. He received zero score for both thinking and performance scores for he attributed the two percent decrease of crime rate to the imposition of death penalty. He did not think that there may be other plausible explanations that can be justifiably attributed to the decrease of crime rate.

Another student verbalized his thoughts in answering the said item. His answer is given below:

Student B: *Letter C kasi sa second statement ano parang inassume na nilayung two %na nag drop yung crime rate ay dahil yon sa imposition of death penalty. Walang sinabing statement dito para sa first statement na pag bumaba yung crime rate, ibig sabihin death penalty is effective. Saka pwede din na may ibang dahilan kung bakit bumaba yung death penalty in that month.*

( *Letter C because in second statement they somewhat assumed that the 2 % decrease in crime rate is due to the imposition of death penalty. There is nothing given in the first statement that when crime rate goes down that means death penalty is effective. Besides, there may be other explanations why death penalty goes down in that month.*)

Student B received a score for both thinking and performance. He arrived at the keyed response by thinking critically. His justification by saying that there may be other explanations as to the decrease of 2 % to the crime rate of death penalty shows that he could recognize post hoc fallacy in an argument though he did not use that Latin phrase during his verbalization of thoughts but his explanations captured the essence of the said fallacy. Hence, its construct validity was established.

Concerning item that was accepted but with revision is exemplified by item 23.

Student C verbalized his thoughts as he worked on item 23. His chosen answer and its explanations are given below:

*Letter B. Parang ito ang mas believable. Kasi mas mababa yung pwesto ni Oliver so mas believable siya kesa kay Ace. Mas naniniwala ako sa kanya kasi nung nagta tie siya ng shoelace syempre parang nakayuko siya so parang mas kita niya yung rat na pumasok don sa small hole na sinasabi kesa kay Ace na nakatayo na di masyadong malinaw ang pagkakakita.*

(*Letter B. This one seems more believable. Because the body position of Oliver is lower than that of Ace. Oliver is more believable because he is tying his shoelace and his body position is stooped that makes him clearly see the rat that entered the small hole. Ace is less believable for he is in standing position that makes it difficult for him to see clearly the rats that entered the small hole.*)

Based on the test manual, the keyed answer is A because the attention of Oliver is distracted due to tying his shoelace. Hence, student C received a zero score for performance but one score for thinking. His given reason is justified because he is implicitly invoking other critical thinking criterion on judging observation statement which is the condition of the observer must be conducive to observation. He is assuming that the one who is in a sitting position can better observe the rat in a paddy than a person who is in standing position. These verbalized thoughts call for revision of the item.

Furthermore, student D has the same chosen answer and very much the same explanation as regards item 23. His verbal report of thinking is given below.

*Mas pinaniniwalaan ko yung letter B. Sabi niya isa lang daw yung rat na nakita niya. Mas naniniwala ako sa kanya kasi nung nagta-tie siya ng shoelace syempre parang nakayuko siya so parang mas kita niya yung rat na pumasok doon sa small hole na sinasabi kesa kay Ace na nakatayo siya. Though, nagta tie ng shoes si Oliver at divided ang attention niya, para sa akin mas na view niya ng maayos ang rat dahil halos ka level ng eyes niya ang paddy.*

*(I tend to consider letter B as more believable. He said that he just spotted one rat. He is more believable because when he is tying his shoelace his body is somewhat bent downward and this makes him view more clearlythat there is just one rat that entered the small hole than Ace who is in standing position. Though, Oliver is tying his shoelace and has a divided attention, for me he really views the rat more clearly because his eyesight level is nearer the paddy than that of Ace.)*

Similarly, student D received a zero score for performance and one score for thinking. His justification for his chosen answer is the same with that of student C. Though their chosen option does not agree with the keyed response, but their reason justifies their selected answer.

Hence, item 23 was revised by making Ace in sitting position which makes his body position similar to that of Oliver who is tying his shoelace. This makes Ace's and Oliver's body position quite similar and somewhat differ only in which one of them has a divided attention which makes Oliver statement less credible.

The revised version of this item makes option A the conclusive answer with justification that the attention of Oliver is distracted due to tying his shoelace whereas Ace attention is more focused than that of Oliver. Thus, Ace is considered more believable than Oliver.

High correlation between performance and thinking scores through verbal reports of thinking is not without a problem. It is possible to have a high correlation between performance and thinking scores in an item but a revision or modification of item is needed. An illustrative example is item 67 which is part of induction item of the said test. Each item in this section of the test focuses on explanatory conclusion, specifically, causal claims.

Student G chose letter C as her answer in item 67 which means that the information does not help us decide that ABC pellets lethally poison the rats. His detailed answer is given below.

*Take note sabi dito "competing." Sa pagkabasa ko pa lang sa word na competing hint nasa akin yon para letter C sagot ko. Kasi opinion yun ng competing company na manufacturer din ng same product. Syempre may bias don kasi may conflict of interest.*

*(Take note it says here "competing." Just merely reading the word "competing" is enough hint for me to choose letter C as my answer. Because that is the opinion of the competing company that is a manufacturer also of the same product. For sure there is bias there because of conflict of interest.)*

Student G chose the keyed answer by thinking critically because she invoked one critical thinking criterion in judging the credibility of information which is *conflict of interest*. His justification is tantamount to saying that suspension of judgment is needed for there is a presence of conflict of interest. However, he also implied that the item can be answered easily without reading the entire question because of the hint of the word "competing" as part of the question. With his reasonable justification, he received a score for both performance and thinking.

In the same way, student H has similar answer in the said item. His brief explanation and chosen answer is given below.

*Obviously, the answer is letter C because of the word "competing"….A group of people from a competing company, from there you could say automatically that there is bias in it. So my answer is letter C which is neither.*

In a like manner, student H chose the keyed answer by thinking well. He was able to justify his answer by stating the word bias in which the competing company tends to be biased because it also manufactures similar product. Like student G, this student received a score for both performance and thinking.

Despite the high correlation between performance and thinking scores of item 67, it was revised by changing the word *competing* into other term for the said word led examinees easily to determine the keyed correct answer.

Furthermore, it is possible that despite the moderate correlation between performance and thinking scores the item was retained as is because there was no clear indication from verbal reports of thinking of examinees that the item needs revision or replacement. This is instantiated by item 87 in which verbal reports of thinking are shown below.

Inthis item, student I chose letter B as his answer. He explained his justification with the following statements: *Letter B sagot ko. Kasi parang si Oliver di naniniwala sa sinabi ni Mary na hindi hardworking ang mga Filipino. Yun ang naisip ko.)*

*(Letter B is my answer. It seems that Oliver does not believe in what Mary said that Filipinos are not hardworking. That is what I thought about.)*

Giving a vague explanation to his chosen answer, student I was asked by the interviewer to take note and analyze the negative words in the propositions of Mary. After which, a follow-up question was asked to student I: *Why did you choose letter B which states that Oliver misdescribed and challenged the position of Mary?* Student I

simply replied, *Basta letter B sagot ko. Mahirap ang tanong di ko din alam explain. (I just simply answered letter B. The question is difficult. I don't know how to explain.)*

Student I was able to get the keyed correct answer but was not able to justify his chosen answer. Thus, he received a score of 1 in the performance but no score on thinking. His answerjust shows that it is highly probable for an examinee to choose the keyed correct answer without any valid reason at all or by mere sheer guessing in a multiple-choice type of test which is one of the big drawbacks of this type of examination.

Student J answered the same number of test item. He said that: *Ang hirap nito. Ang sagot ko is letter A. Kaya hindi B ang sagot ko kasi si Oliver ang sinasabi niya hindi naman na misdescribe sa pagkakaintindi ko. Yung letter C di ko pinili because siguro yung conclusion ni Mary hindi nag match sa reason niya pero yung kay Oliver parang malinaw naman ang statements niya. So ang parang sagot ko dito, yung kay Mary lang ang parang mali, parang ganon. Yung statement dito sa question na to, napakadaming negative words. Kaya lalo nagiging complex.*

*(This is a tough question. My answer is letter A. I did not choose B as my answer because Oliver does not misdescribe anything. That is how I understood it. I did not choose letter C because the conclusion of Mary does not match her given reason whereas the given statements of Oliver are seemingly clear. So it looks that it is only the given statements of Mary that are wrong. The statements in this question are loaded with negative words. That makes the question more complex.)*

Student J chose the unkeyed option by thinking poorly. He arrived at a conclusion that the answer is letter A without thorough analysis of the propositions given by two characters in the test. Students I and J must have explained that Oliver incorrectly attributed a conclusion which can serve as a position to the argument of Mary. The stated conclusion of Mary is "our country will become more progressive one" which is contrary to attributed conclusion of Oliver to argument of Mary. The wrongly attributed conclusion of the former is challenged by himself. This is a case of straw-person fallacy. Furthermore, the double negation in the given argument of Mary makes a positive. Both A and C are false.

In spite of the comment of student J that item 87 is difficult due to compounding of negative words,the said item was not revised for Ennis (1987) said that some sophistication in dealing with negation is necessary in the teaching and testing of critical thinking.

It is interesting to note that though correlation is moderate between performance and thinking scores does not necessarily mean that item 87 needs revision. If the mismatch of scores is due to the performance scores outweigh thinking scores then it may mean that the item can be retained as is because there is no basis for the revision of the item that can be inferred from the verbal reports of thinking. However, if the mismatch of scores is due to thinking scores outweigh the performance scores, then, the item needs revision or may even be basis for item rejection. In case of the latter, revisions can be made in the item stem or the keyed correct answer. Theinsights from the verbalized thoughts of the examinees are strong basis for item revision for they may lead test developers how the items should be revised.

Verbal reports of thinking are relevant in establishing the construct validity of critical thinking test items. Norris (1989) used this procedure in validating *Test on Appraising Observations*and recommended that verbal reports could be explored in determining the construct validity other than single-aspect critical thinking testin which a case in point is the*Test on Appraising Observations* which deals only with credibility judgment. Hence, his procedure was adapted in establishing the construct validity of all items of *The CEU-Lopez Critical Thinking Test* which is a multi-aspect general- knowledge critical thinking test that deals with five aspects of critical thinking: deduction, credibility judgment, assumption identification, induction, and meaning. These verbal reports provide direct evidence on the processes of students' thinking on how they arrived at their answers in a trial version of amulti-aspect general-knowledge critical thinking test using multiple-choice format. If students choose the unkeyed option but think critically then an item needs modification or replacement. If the students choose the keyed answer by mere guessing then the item can be retained for there is no clear basis from verbal reports that an item is defective.This is one notable disadvantage of using multiple-choice test format for a critical thinking test that an examinee may possibly select the answer keyed correct out of sheer guessing.

Furthermore, it is interesting to note that the correlation between performance and thinking scores together with the verbal reports provide clear guidance on whether the test item is valid or invalid. This is affirmed by Norris (1992) who said that direct evidence on the thinking processes that students used to answer items of multiple-choice critical thinking test is relevant to establish their construct validity.

Undoubtedly, there may be other relevant ways in determining the construct validity of a multiple-choice critical thinking test but verbal reports of thinking gathered from this study show empirically whether an item needs to be revised, retained, or discarded. The high correlation between performance and thinking scores may not mean that the item is free from error as exemplified in item 67. Alternatively, moderate correlation may not mean that the item is problematic and needs to be revised or discarded as shown in item 87. A test developer who considers toadapt this methodology should also look into the insights of verbal reports of thinking than just merely rely on the correlation of performance and thinking scores to establish construct validity of critical thinking test items.

In this fast-changing and highly-connected world, the challenge for educators is to come up with a test in which the focus is on the processes of thinking than just a type of test that calls for students to regurgitate information they just memorized from text. Verbal reports of thinking from this study show that students are capable to think well and it could be enhanced further by challenging our students to give them test that would stretch their thinking ability. Furthermore,although*The CEU-Lopez Critical Thinking Test* is a multiple-choice type of test, the verbal reports of thinking provide direct evidence that students think critically as they answer the questions of the said test that have direct bearing on the processes of critical thinking.

## References


Alderson, J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Anastasi, A. (1988). *Psychological testing*. New: Macmillan.

Ebel, R. L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9-26). New York: W. H. Freeman.

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

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

Ennis, R. H. (2009). Investigating and assessing multiple-choice critical thinking tests. In J. Sobocan & L. Groarke (Eds.), *Critical thinking education and assessment: Can higher order thinking be tested?* (pp. 75-97). Canada: The University of Western Ontario.

Ennis, R. H. (2011a). Critical thinking: Reflection and perspective Part I. *INQUIRY: Critical Thinking Across the Disciplines, 26*(1), 4-18.

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

Fisher, A., & Scriven, M.(1997). *Critical thinking: Its definition and assessment*. Edgepress and Centre for Research in Critical Thinking, University of East Anglia.

Lopez, M. Y. (2004). *Development and validation of critical thinking infusion lessons in communication skills for freshman college students*. Unpublished doctoral dissertation, Philippine Normal University, Manila.

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

Norris, S.P. (1988). Controlling for background beliefs when developing multiple-choice critical thinking tests. *Educational Measurement*, *7*(3), 5-11.

Norris, S.P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement*, *27*, 41-58.

Norris, S. P. (1991). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In J. F. Voss, D. N. Perkins, & J. Segal (Eds.), *Informal reasoning and education* (pp. 451-472). Hillsdale, NJ: Erlbaum.

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.

Norris, S. P., & King, R. (1983).*Test on appraising observations.* St. John's, NF: Institute for Educational Research and Development, Memorial University of Newfoundland.