Psychological Test and Assessment Modeling, Volume 56, 2014 (3), 275-286

Detection of Differential Item Functioning in the Cornell Critical Thinking Test across Korean and North American students

Brian F. French[1], Brian Hand[2], Jeonghee Nam[3], Hsiao-Ju Yen[4] & Juan Antonio Valdivia Vazquez[4]

Abstract

Critical thinking can be considered the conscious process a person does when he or she explores a situation or a problem from different perspectives. Accurate measurement of critical thinking skills, especially across cultures, in part, depends on the instrument's measurement properties being invariant or similar across those groups. The assessment of item level invariance is a critical component of building a validity argument to ensure scores on the Cornell Critical Thinking Test (CCTT) have similar meaning across cultures. Item Response Theory methods were used to examine differential item functioning by culture in the CCTT-Form X. Results suggest that the items do function similarly across Korean and United States students with only 14% of items displaying DIF. Additionally, the majority of these DIF items appeared on one of the subscales. Practitioners and researchers can have confidence that mean differences observed at the total score level or on 3 of 4 subscales are not a function of a lack of measurement invariance. Findings support the validity of the inferences drawn when comparing students from these two countries.

Keywords: Differential Item Functioning, Critical Thinking, Invariance, Culture, Korea

[1] *Correspondence concerning this article should be addressed to:* Brian F. French, Ph.D., Cleveland Hall, Rm 362, College of Education, Washington State University, Pullman, Washington 99163; email: frenchb@wsu.edu

[2] University of Iowa

[3] Pusan National University

[4] Washington State University

Critical thinking, defined as "reflective and reasonable thinking that is focused in on what to believe or do" (Ennis, 1993) allows for efficient and effective evaluations. Critical thinking receives much attention from a diverse group of research areas from student achievement (e.g., Garett & Wulf, 1978) to medical practice and education (e.g., Khosravani, Manoochehri & Memarian, 2005). A search of the Educational Resources Information Center database, for instance, between the years of 2000 and 2011 employing the keywords *critical thinking* resulted in over 4900 references. The previous 10 years resulted in identifying over 4800 references. This brief search does indicate a steady trend over the past 20 years to examine critical thinking. Critical thinking skills are important because they increase a person's capacity to assess a problem adequately (Ennis, 1993) while presenting better strategies to solve the problem (Glevey, 2006) and obtaining consequent information necessary to become a successful life-time learner (Halpern, 1998). An individual needs critical thinking (CT) skills to identify the logic and reasons of a problematic situation, to formulate inferences and calculate scenarios, and to draw conclusions while searching for the best solution (Halpern, 1998). Thus, CT skills are a desirable goal expected from the educational system and in the workforce. A constant challenge associated with developing CT skills, particularly with students, is accurately assessing a student's level of CT to be able to intervene with target strategies as appropriate.

Several tools have been developed to assess general or specific critical thinking skills including but not limited to the Watson-Glaser II Critical Thinking Appraisal (WGCTA; Watson & Glaser, 2009) and the Critical Thinking Assessment Test (CAT; Tennessee Technological University, 2010). This study focuses on the Cornell Critical Thinking Test (CCTT; Ennis, Millman & Tomko, 2005) as it is a widely used measure (e.g., King, Wood, & Mines, 1990; Nieto & Saiz, 2008; Solon, 2007). For instance, cross-cultural research practice encourages the continued assessment of a general thinking structure and skills (Mason, 2007) to determine if a set of universal CT skills exists or if such skills are culture dependent.

Cross-cultural assessment and Critical Thinking Skills

Reliability and validity evidence to support the inferences of the CCTT scores has been reported to be strong for general use. However, no studies have explored whether the items of the CCTT display differential item functioning (DIF) across cultures. Investigating DIF, or a lack of measurement invariance, is critical to ensure that the constructs across cultures have the same meaning and result in accurate mean comparisons (Cole, Maxwell, Avery, & Salas, 1993; Ferne & Rupp, 2007; Thissen et al., 1986). This is critical, as mean differences, or lack of, does not support the presence or absence of bias. Specifically, the current need for assessing construct equivalence has resulted in the combining of two American Psychological Association divisions (52 and 5 – International Psychology and Measurement). The focus, in part, is to be aware that there are three critical aspects to be examined carefully when engaging in cross-cultural assessment: construct validity, item bias, and the developing of local norms (Byrne, Oakland, Leong, van de Vijver, Hambleton, Cheung, & Bartram, 2009). We tackle the second need in this study.

There are three main reasons we are interested in how items function with Korean students compared to U.S. students. First, the traditional Korean educational system relies on methodologies that center on the teacher as the giver of knowledge and has a focus on memorizing practices or strategies. Second, critical thinking is in somewhat of a conflict with Korean culture because it promotes intellectual independency, student's autonomy, verbal persuasion and customary communication patterns that are not highlighted in this culture as they are in the U.S. educational system. Third, Korean collectiveness is another cultural aspect to consider when evaluating CT skills as this view is promoted unlike the US where individualism is promoted (McGuire, 2007). This difference in culture may influence how Korean students respond to items assessing CT that is quite different compared to their US counterparts. Moreover, these formats and cultural aspects can be found in other counties similar to the U.S. and Korea. Thus, findings here may be generalized to similar educational settings where these components overlap.

The importance of Differential Item Functioning detection

Validity is not considered a property of the assessment itself but rather based on the inferences drawn from test scores. Therefore, establishing evidence upon which appropriate inferences can be justified is critical. We specifically examine items, the foundation of an instrument, to determine if the items function differently across Korean and United States students. That is, we examine the CCTT for differential item functioning (DIF). DIF can be defined as the case where examinees from two groups who have equal levels on the measured ability have different probabilities of endorsing the same item response. The presence of DIF indicates that the item is not performing the same across groups. This also is referred to as a lack of item level measurement invariance. Under such conditions the item may have lower construct validity for one of the groups (Steinberg & Thissen, 2006).

Detection of DIF is important as it can influence the psychometric properties of an instrument and mean score comparisons (e.g., Church, Alvarez, Nhu, French, & Katigbak, 2011; French, Maller, & Zumbo, 2007). Church and colleagues, for example, in cross-cultural comparisons of personality traits, reported 40-50% of the items exhibited DIF and this DIF was shown to cascade to the subscale score level at which observed scores are compared. After accounting for DIF, some cultural observed mean differences were reduced or eliminated. Moreover, the Type I error rate of the *t*-test can be inflated in the presence of DIF (Li & Zumbo, 2009) as well as increased inequality of observed score variances across the groups in the presence of non-invariance (Finch & French, *in press*) resulting in inaccurate score comparisons. Thus, examining assessment for DIF before making group comparisons is critical in order to conclude that the comparisons are accurate (e.g., Galic, Schefer, LeBreton, 2014).

The CCTT measure was the focus of this DIF examination as it is the most widely recognized critical thinking test. The technical manual (Ennis et al., 2005) for the CCTT cites more than 34 studies using the CCTT-Form X until 2005. However, it appears this is the first study of Korean DIF for this assessment. Thus, the purpose here was to investigate the extent to which the CCTT-Form X items are invariant (i.e., show no DIF) across Korean and United States students via item response theory likelihood ratio test procedure for DIF detection.

Method

Participants

The United States students (n = 907) were asked to complete the CCTT-Form X in their respective classrooms as a component of a larger evaluation study. The locations the students were recruited from were sites in the Midwestern United States. These sites where involved in a randomized control trial aimed at increasing science literacy and reasoning skills with one targeted outcome being critical thinking skills. Thus, the CCTT was being used as an evaluation tool as part of a comprehensive assessment protocol. Sex was approximately equal with 47 % males and 53 % females. Ethnicity of the participants was only reported in two categories including Caucasian (97 %) and other (3 %). Age was not reported but students had a mean grade of 6.24 (SD = .99). Given the sample was from the United States, it is assumed the average age of the student was approximately 11 years old given the average grade of the student.

The Korean students (n = 811) were asked to complete the CCTT-Form X in their respective classrooms as a component of a larger evaluation study. The locations the students were recruited from schools in the largest and the second largest cities in Korea. Eight schools were involved in a randomized control trial aimed at increasing science literacy and reasoning skills with one targeted outcome being critical thinking skills. Thus, the CCTT was being used as an evaluation tool as part of a comprehensive assessment protocol. Sex of participants was 42 % males and 58 % females. Ethnicity of the participants was only Korean. Age range of the students was 12 to 14 years old and the average age was 13 years old.

Instrument

The Cornell Critical Thinking Tests (CCTT; Ennis, Millman, & Tomko, 2005) assesses general critical thinking ability including areas of induction, deduction, evaluation, observation, credibility of statements, assumption identification, and meaning. Form X is examined in this study as it is appropriate for grades 4 through the sophomore year of post-secondary education. The CCTT-Form X requires approximately 50 minutes responding to the 71 items. There are three response options per item in a multiple-choice format and the items are scored dichotomously (i.e., correct or incorrect). The items ask the students about the material they have read about in the test forms. These questions are mainly at the higher level of thinking (i.e., beyond factual recall). The technical manual does provide some score reliability and validity information reporting internal consistency reliability estimates ranging from 0.67 to 0.90 for Form X, the form under examination in this study. Internal consistency for the scores based on the US and Korean samples used in this study was 0.83 and 0.81, respectively.

The correlations between the CCTT and other critical thinking tests range from 0.60 with the Critical Reading in social studies assessment to 0.50 and 0.41 with the Logical Reasoning Test and the Watson-Glaser assessment, respectively (Ennis et al., 2005). The correlations between the CCTT and other constructs range from 0.74 with the Otis-Lemon assessment, 0.53 with the Houghton-Miffin Cognitive Abilities verbal assessment, and 0.52 with the SAT total score. These values together support validity of the inferences from the measure.

Analysis

Item response theory (IRT) provided the framework for DIF detection. The IRT framework is built upon the idea that a person's response to an item is determined by a latent ability on a single dimension of measurement that is independent of the administered test (Hambleton, Swaminathan, & Rogers, 1991). Given this assumption of unidimensionality, DIF assessment was conducted four times, one analysis for each domain or dimension assessed by the CCTT.

The 2-parameter logistic item response theory (2-PL IRT) model was selected for each subtest allowing for the assessment of both uniform and non-uniform DIF. We did not assume the *a*-parameter (i.e., item discrimination) was the same across the assessment. Additionally, we did not include the pseudo-guessing parameter, (c-parameter) as the sample size did not support accurate estimation of the 3PL. In fact, some parameters did not converge to a solution with the 2-PL model most likely due to sparseness of data. DIF detection followed a two-step process where a purified anchor set of items was determined and final DIF detection was completed. Purification is suggested for use with DIF detection regardless of method employed (Camilli & Shepard, 1994). To determine the purified anchor set of items we followed the suggested iterative procedures using logistic regression for DIF detection with a dual-criterion (French & Maller, 2007). We conducted purification by subtest to meet unidimensional assumptions.

The IRTLR procedure (Thissen et al., 1986), assesses for the presence of DIF (both uniform and nonuniform) by comparing the fit of an IRT model assuming equality of the parameter estimates for the item in question across the reference and focal groups (compact model) with the model fit when this constraint is relaxed (augmented model). IRTLR uses a sequential methodology, beginning with a comparison of the two groups on all item parameters simultaneously for the target item. The test statistic takes the form:

$$
\begin{aligned}
&LR = -2\ln L_c - \left(-2\ln L_A\right)\\
&\text{where}\\
&L_c = \text{log likelihood of the compact model}\\
&L_A = \text{log likelihood of the augmented model}
\end{aligned}
\tag{1}
$$

When the resulting χ^2 distributed statistic is significant, subsequent tests are computed to examine item parameter differences. Tests for individual parameters proceed in a sequential manner. In the most complex case for multiple choice items where pseudo-guessing can occur the pseudo-guessing values (in the 3PL context) for the two groups are compared by constraining them equal between groups for one model and comparing this L value with the log-likelihood for the model in which these parameters are not constrained equal. Next, the pseudo-guessing values are constrained, while the item discrimination values are allowed to vary between groups in the augmented model and the L statistics are used to construct a LR test for a assuming that c is equal between groups. Finally, the test comparing item difficulty is conducted in like manner, so that for the augmented model both c and a are constrained equal between groups while b is allowed to vary. Thus, the test for uniform DIF is conducted conditioned on group equality for both a and c. This process is conducted automatically in the IRTLRDIF software (Thissen, 2001). Recall, in this analysis, the 2PL was employed given data sparseness. However, we present the 3PL as differences in the pseudo-guessing can result in a loss of accuracy in DIF detection for differences on other parameters (Finch & French 2014).

To judge the magnitude of item functioning differences, the area between item characteristic curves (ICCs) was computed utilizing Raju's formula (1988), where small (0.40), medium (0.60), and large (0.80) differences between groups were identified. We recognize this categorization is not perfect. However, we employ it to capture DIF magnitude in the absence of a strong literature base on effect sizes in this area. In addition, many DIF simulations have used this categorization to develop DIF detection methods as well as compare DIF methods. In addition, using the dual-criteria can protect against inflation of Type I errors due to differences in ability differences (e.g., DeMars, 2010).

Results

IRTLRDIF detection showed very little culture DIF across the domains assessed with the expectation of the Induction domain. We do note that items 8, 13, and 14 on the Induction subscale resulted in standard errors that were not estimable due to sparseness of data. Thus, these items were marked as potential DIF items as they were identified as such yet there is some uncertainty of their status. However, the induction subtest did contain many DIF items. Table 1 displays the results for items identified as DIF items in the Induction section. Of the 23 items, six items (26%) exhibited DIF. Item 2 showed differences in item difficulty, whereas item 7 exhibited differences in both difficulty and discrimination. Items 17 and 22 exhibited differences only in item discrimination. The pattern of differences suggests that items 2 and 22 favor US students whereas items 7 and 17 favor Korean students. Examination of the associated effect sizes revealed large differences between US and Korean students only on item 2. Item 22 demonstrated a medium difference.

Table 1:
Differential Item Functioning Analysis for Korea and US Comparisons for the Cornell Critical Thinking Test by Subscale

Studied Item	US				Korea				
	a	(SE)	*b*	(SE)	*a*	(SE)	*b*	(SE)	area
Induction									
2 [b]	0.90	0.14	-2.15	0.21	0.90	0.15	-1.57	0.26	0.58
7 [a,b]	0.18	0.09	-0.73	0.54	0.50	0.11	-1.17	0.34	2.92
8 [a,b]	0.11	0.26	8.68	(****)	0.14	0.11	4.89	3.54	3.87
13 [a]	0.07	0.09	6.70	(****)	0.04	0.19	12.22	(****)	9.91
14 [b]	0.15	0.14	11.21	(****)	0.18	0.24	11.63	8.99	0.97
17 [a]	0.65	0.12	-0.06	0.20	0.41	0.11	-0.31	0.27	0.76
22 [a]	0.61	0.12	-1.38	0.19	1.09	0.15	-1.27	0.18	0.60
Observation & Credibility									
29 [c]	1.06	0.17	-1.44	0.13	1.52	0.21	-1.63	0.15	0.28
39 [a,b]	0.71	0.12	-1.21	0.17	0.42	0.11	0.99	0.34	2.22
Deduction									
52 [a,b]	0.58	0.09	0.23	0.19	1.11	0.15	-1.05	0.14	1.35
Assumption									
71 [a,b]	0.15	0.22	9.59	8.29	0.18	0.10	4.91	3.08	4.68

[a]Nonuniform DIF

[b]Uniform DIF

[c]The overall test was significant yet neither test for the individual parameters was significant.

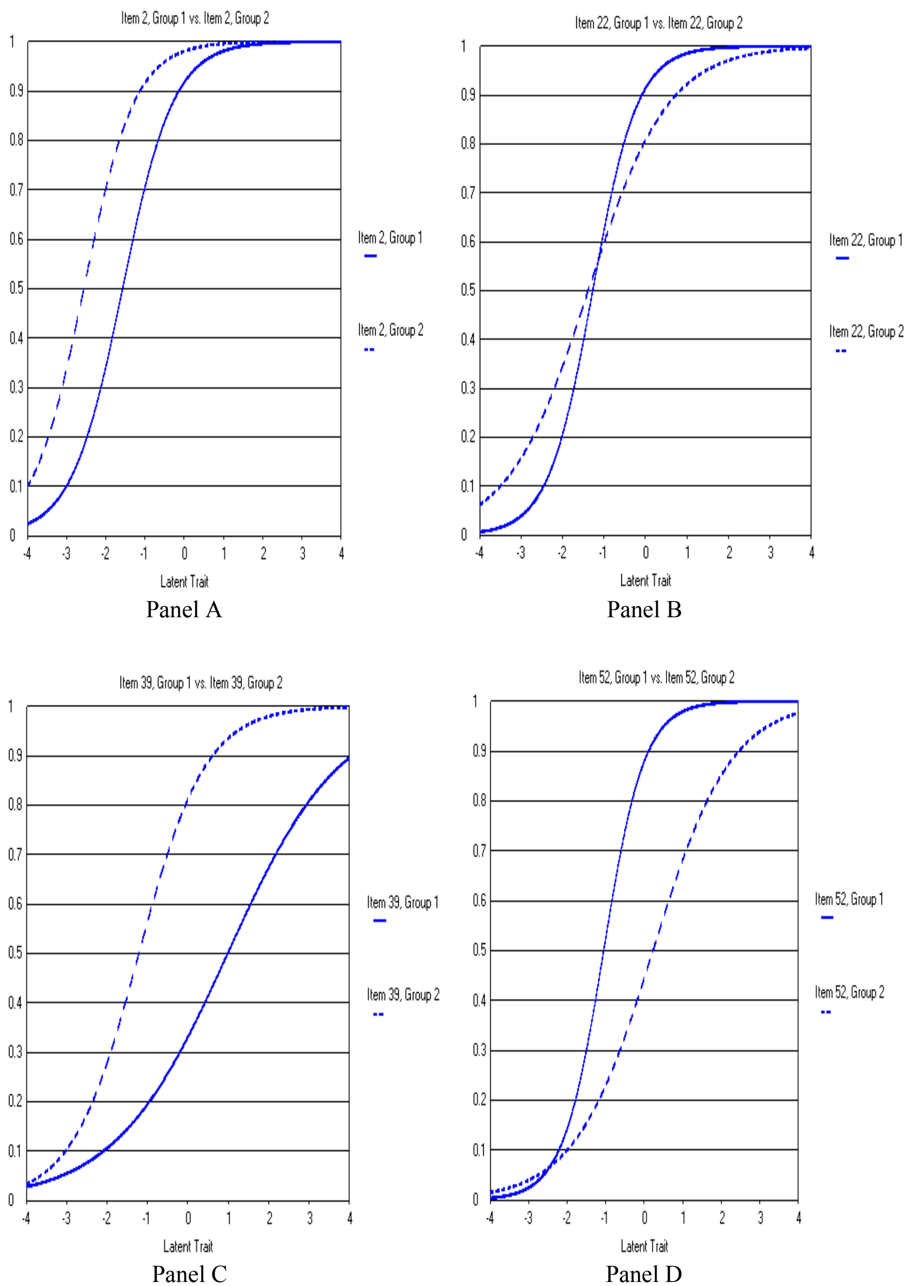

Figure 1:
Example DIF Items across Domains in Graphical Form

Figure 1 displays four panels each containing the item response curves across the two groups for an item identified as a DIF item. On the x-axis is the ability distribution whereas the y-axis contains the probability of a correct response ranging from 0 to 1.0. Item 2 in Panel A shows uniform DIF that favors US students (dotted line) over Korean students (solid line). Specifically, the US students have a higher probability of endorsing this item than Korean students. Item 22 in Panel B is a nonuniform DIF item with two item characteristic curves cross. This figure indicates that US students have higher probability endorsing this item than Korean students below -0.8 on the ability distribution; whereas Korean students have higher probability endorsing this item above -0.8.

As Table 1 indicated, there were only 2 items (8 %) out of the 24 items in observation and credibility skills that exhibited DIF. Items 39 favored US students whereas item 29 favored Korean students. The associated effect sizes revealed that item 39 exhibits a large effect size difference between Korean and US students whereas item 29 appears to have a small effect. The difference on item 39 can be seen graphically in Panel C in Figure 1. In addition, there was only 1 item (item 52) out of the 14 (7 %) that assesses deduction skills that exhibited DIF. This item favored the Korean students. The effect size revealed that item 52 exceeded the threshold of a large effect (0.80). This can be seen graphically in Panel D of Figure 1. Last, Table 1 revealed that only 1 item (10 %) out of the 10 assessing assumption skills was identified as a DIF item. This item appears to favor the Korean students. The effect size revealed that Korean and US students respond differently on item 71 with a large effect.

Discussion

The main goal of this study was to provide the first measurement invariance evidence for students from Korea and the United States at the item level for the Cornell Critical Thinking Test. Such validity evidence is an essential step in supporting the accuracy of mean differences between these groups. The results revealed that, in general, the items on the CCTT function similarly across Korean and US students with only 14 % of the items identified as DIF items. This is further supported by the associated effect sizes. Thus, it is likely that these items would not obscure mean comparisons at the total scale level given the number of items (Finch & French, *in press*). However, this should be investigated further as there is no clear answer in the DIF literature about how much is too much DIF to influence decisions about individuals. The assumption is likely true for all subscale score comparisons with the exception of the induction subscale scores. With 26 % of items in this section being identified as DIF items, careful content review of the items is suggested. Although the percentages are large, they are lower compared to what has been found with reasoning tests recently (e.g., 36% DIF; Galic et al., 2014). This review should assist in pinpointing why these items might be functioning differently and what type of revisions may be needed. That is, additional exploration into the reliability and validity of the test is warranted to continue to support the resulting scores in various environments, populations, and cultures. Specifically, there is a strong need to continue to compare item functioning across such different cultures to understand how larger grouping of cultures (e.g., individualistic vs. collective) is appropriate (e.g. Church et al., 2011) on many educational and psychological outcomes.

A few limitations of this study are worth mentioning. First, we were able to only conduct analysis on one of the two forms (Form X) of the CCTT. Results based on one form are not guaranteed to hold across forms for Korean and US students. That is, results cannot be generalized to Form Z. Efforts should be made to study this form as well. Second, we only examined differences related to one culture outside of the US. This does not guarantee that results generalize to other cultures. Additional work is needed to examine measurement invariance across other cultures to support the validity argument for the use of the scores and inferences based on these scores across these various groups. Last, failure to obtain standard error estimate for the three Induction items is questionable. We are unable to determine whether these three items demonstrate a real difference between the studied groups or other factors (e.g., data sparseness) that prevent such parameter estimate.

Based on the use of the CCTT as mentioned in the manual, the scale should provide valuable data to researchers and practitioners wanting to assess critical thinking skills. The CCTT may remain popular in such environments as the instrument is easy to administer, score, and interpret. The CCTT also is one measure that can begin to assess the transfer of critical thinking skills from one environment to another. The usefulness of the CCTT to do such tasks with meaning hinges on the validity evidence to support score-based inferences. Indeed, establishing such evidence for these types of measures is paramount, as there is a sustained effort and interest in the past two decades focused on CT skills. The evidence provided here does support the use of the CCTT scores for comparisons across US and Korean students with the exception of the induction scores if subscale score analysis is undertaken.

References

Byrne, B.M., Oakland. T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*, 94-105. doi: 10.1037/a0014516

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oakes, CA: Sage.

Church, T. A., Alvarez, J. M., T.Q. Nhu, French, B. F., & Katigbak, M. S. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the revised NEO personality inventory. *Journal of Personality and Social Psychology.*

Cole, D. A., Maxwell, S. E., Avery, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114*, 174-184.

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*, 961-972.

Ennis, R. H., Millman, J., & Tomko, T.N. (2005). *Cornell Critical Thinking Tests*. Seaside, CA: The Critical Thinking Co.

Ennis, R. (1993). Critical thinking assessment. *Theory into Practice, 32*, 179-186.

Ferne, T., & Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*, 113-148.

Finch, W. H., & French, B. F. (2012). The impact of factor noninvariance on observed composite score variances. *International Journal of Research and Reviews in Applied Sciences.*

Finch, W.H. & French, B.F. (2014). The impact of group pseudo-guessing parameter differences on the detection of uniform and nonuniform DIF. *Psychological Test and Assessment Modeling, 56,* 25-44.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for DIF detection. *Educational and Psychological Measurement, 67, 373-393.*

French, B. F., Maller, S. J., & Zumbo, B. (2007, April) *The influence of differential item functioning on multi-sample confirmatory factor analysis*. Paper presented at the National Council on Measurement in Education conference, Chicago, IL.

Galic, Z., Scherer, K. T., & LeBerton, (2014). Examining the measurement equivalence of the conditional reasoning test for aggression across U.S. and Croatian samples. *Psychological Test and Assessment Modeling, 56,* 195-216.

Garett, K. & Wulf, K. (1978). The relationship of a measure of critical thinking ability to personality variables and to indicators of academic achievement. *Educational and Psychological Measurement, 38,* 1181-1187.

Glevey, K.E. (2006). Promoting thinking skills in education. *London Review of Education, 4* (3), 291-302. DOI 10.1080/14748460601044005

Halpern, D.F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist, 53*, 449-455.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Khosravani, S., Manoochehri , H., & Memarian, R (2005). Developing Critical Thinking Skills in Nursing Students by Group Dynamics. *The Internet Journal of Advanced Nursing Practice, 7,* 1-12.

King, P.M., Wood, P.K., & Mines, R. A. (1990). Critical thinking among college and graduate students. *The Review of Higher Education, 13*, 167-186.

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica, 30, 343-370.*

Mason, M. (2007). Critical thinking and learning. *Educational Philosophy and Theory, 4,* 339-349. DOI: 10.1111/j.1469-5812.2007.00343.x.

McGuire, J.M. (2007). Why has the Critical Thinking movement not come to Korea?, *Asia Pacific Education Review, 8*(2), 224-232. doi: 10.1007/BF03029258

Nieto, A.M., & Saiz, C. (2008). Evaluation of Halpern's "Structural Component" for improving critical thinking. *The Spanish Journal of Psychology. 11*, 266-274.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Solon, T. (2007). Generic critical thinking infusion and course content learning in introductory psychology. *Journal of Instructional Psychology, 34*, 972–987.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods, 11,* 402-415.

Tennessee Technological University (2010). *CAT Instrument technical information.* Retrieved http://www.tntech.edu/images/stories/cp/cat/reports/CAT_Technical_Information_V7.pdf

Thissen, D. (June, 2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood ratio tests for differential item functioning.* Unpublished manual from the L. L. Thurstone Psychometric Laboratory: University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Watson, G. & Glaser, E.M. (2009). *Watson-Glaser II Critical Thinking Appraisal technical manual and user's guide.* San Antonio, TX: Pearson. Retrieved from http: http://www.talentlens.com/en/downloads/supportmaterials/WGII_Technical_Manual.pdf