Educational Sciences: Theory & Practice • 14(1) • 159-178
©2014 Educational Consultancy and Research Center
www.edam.com.tr/estp
DOI: 10.12738/estp.2014.1.1840

Cross-cultural Equivalency of the California Critical Thinking Disposition Inventory

Gökhan İSKİFOĞLU[a]
Cyprus International University

Abstract

This study describes the cross-cultural applicability of a multidimensional inventory of students' evaluation of critical thinking dispositions (California Critical Thinking Disposition Inventory). The goal was to assess the cross-cultural psychometric equivalency of the CCTDI through testing measurement invariance across American and Turkish linguistic populations. Based on the data from 583 Turkish students and 448 American students from different teacher education programs, the translated Turkish version and the original English version of the CCTDI displayed positive psychometric properties, thus supporting the applicability of the CCTDI in a Turkish educational context with alpha coefficients ranging from .81 to .90 for the sub-scales of the CCTDI Turkish and ranging from .85 to .91 for the sub-scales of the CCTDI English. Results also supported high content validity across cultural versions of the inventory with minimum content validity indices of .81 and .97 for both the Turkish and American versions of the inventory, respectively. However, a cross-cultural comparison of the factorial structure produced a poor fit of the hypothesized multidimensional model of CCTDI to the combined sample. Further analysis, based on the modification indices, supported the use of a four-factor model with reduced items for cross-cultural comparative research studies. Reasons for poor model fit and non-invariance across cultural groups were elaborated.

Key Words

CCTDI, Cross-cultural Equivalency, Cross-cultural Validation, Confirmatory Factor Analysis, Measurement Invariance.

The phrase 'critical thinking' (CT) has been very frequently uttered for the previous two decades in educational contexts across the United States, Europe, and Asia. During the previous five years, most researchers worldwide have (1) engaged themselves in trying to understand what critical thinking actually means, (2) why it is so important, (3) whether it is directly related to or affecting the education that they provide, (4) how they can embed critical thinking into the educational programs or help their students to engage in critical thinking, (5) whether critical thinking is a product or process laden issue, and most importantly, (6) how they can assess critical thinking or evaluate programs in terms of critical thinking. In addition to these questions, assessing critical thinking dispositions of pre-service teachers across different cultural groups with a cross-culturally validated instrument has been one of the most important concerns of a significant number of scholars worldwide (Grosser & Lombard, 2008; Johnson & Reiman, 2007; Lee, 2005; McBridge, Xiang,

[a] **Gökhan İSKİFOĞLU, PhD.**, is currently an assistant professor of Educational Psychology (Measurement and Evaluation in Education). His research interests include developing unique research instruments, experimenting the applicability of research instruments across different cultural groups, studies of SEM family, and experimental research in the field of teacher education. *Correspondence:* Cyprus International University, Faculty of Education, Department of Computer and Instructional Technology Teacher Education, Lefkoşa, Northern Cyprus, via Mersin 10 Turkey. Email: giskifoglu@ciu.edu.tr

& Wittenburg, 2002; Melnick & Zeichner, 1998; Metzler & Blankenship, 2008; Yeh, 2002). Cross-cultural assessment of the construct of 'critical thinking' has found itself to be the most frequently studied subject (Lee, 2005). For this reason, seeking and investigating reliable and valid method to assess critical thinking dispositions of pre-service teachers across multiple languages and cultures has become the major concern of scholars across the world. Many research instruments have been developed to gain deeper insight as to what extent prospective teachers possess both the abilities to use and the dispositions of critical thinking. Many scholars following these developments have preferred to use existing instruments by adapting them into their language and culture rather than developing new ones. According to several leading researchers in the field, one of the most important reasons that has accelerated such cross-cultural studies is related to understanding whether a proposed conceptualization regarding the construct of critical thinking and, in relation to this, hypothesized assessment model in one language and culture exist in a similar structure in different languages and cultures (Behling & Law, 2000; Hambleton, 2005; Sekaran, 1983; Sireci, Yang, Harter, & Ehrlich, 2006; Stansfield, 2003). The basic premise behind such efforts is to determine to what extent a measurement model designed for one culture is applicable for another one. Of-course, the root of this idea traces itself back to the curiosity for seeking of a universally accepted criteria for assessing critical thinking of pre-service teachers across the countries, cultures and languages of the world. For this very reason, the search for a means of ascertaining a reasonably informed opinion of pre-service teacher's critical thinking dispositions in Asia, Europe, and the US have led to the discovery of the California Critical Thinking Disposition Inventory (CCTDI), which is currently the only instrument found to be well conceptualized and translated into many languages, including Arabic, Chinese (Mandarin), Dutch, Farsi, Finnish, French (Canadian), Hebrew, Italian, Japanese, Korean, European Portuguese, Spanish (Mexico-Latin America), and Thai, in order to evaluate higher education programs in terms of critical thinking dispositions (Facione & Facione, 1992b). Since an *authorized* instrument measuring critical thinking dispositions and professional judgment in teacher education was not presently available in Turkish during the inception of this project, the current research endeavors to adapt the CCTDI from the English source language to the Turkish target language, in an effort to explore the cross-cultural validity of the suggested measurement model and to assess the applicability of the CCTDI across cultural groups. As suggested, when attempted to use an existing instrument to measure a phenomenon in another cultural group and/or language for which the instrument was not originally developed, it is required to assess the psychometric properties, examine the extend of the cross-cultural validity and obtain insight into whether the instrument possesses measurement invariance across cultures (Ægisdóttir, Gerstein, & Çinarbas, 2007; Chapman & Carter, 1979; Chen, Sousa, & West, 2005; Nunnally & Bernstein, 1994; Sireci et al., 2006; Stansfield, 2003).

Theoretical and Operational Construction of the CCTDI

Any psychological inventory designed for assessing a phenomenon for educational and research purposes needs to be based on a well-structured and well-conceptualized theoretical framework. For this reason, it is not always easy to develop a well-conceptualized assessment tool. However, for the last two decades, a considerable amount of effort has been given to conceptualize critical thinking and its components. A cross-disciplinary Delphi study, supported by the Committee on Pre-College Philosophy of the American Philosophical Association and which also included 46 international critical thinking experts, was conducted by Facione (1990); and although it continued for two years, it yielded a consensus definition of critical thinking. According to the Delphi report, critical thinking (CT) is composed of both a cognitive skills dimension and an affective dispositions dimension, and thus involves both a *willingness* and the *ability* to use one's cognitive powers of analysis, interpretation, inference, evaluation, explanation, and self-monitoring meta-cognition to make purposeful judgments about what to believe or what to do in a given context (Dewey, 1910; Ennis, 1993; Facione, 1990; Facione, Giancarlo, & Facione, 1995).

If we were to explicate this definition, we would understand that in order for a person to make purposeful judgments regarding what to believe or what to do in a given context, he/she needs not only to have cognitive skills, such as "interpretation," "analysis," "evaluation," "inference," "explanation," and "self-regulation," but also needs to be positively disposed to use these skills (Dewey, 1910; Facione, 1990; Lewin, 1935).

"Being disposed" refers to the affective dispositional dimension of critical thinking. In order to differentiate between a habit and skills, their application in daily life must be regarded; for instance, a person who is habituated to healthful living is more likely to engage in sport activities, eat healthy foods, follow magazines about health, and avoid risky activities (e.g. smoking, drugs, stress...). Another person, on the other hand, might possess the beliefs and skills needed to engage in the same practice but not habitually engage in them. In such a case, we would say that the latter individual is not positively disposed to engage in such practices. The same is valid for thinking. As Facione, Facione, and Giancarlo (1997) explain, people may have the necessary skill(s) to think well or deal with a given problem, and yet, unless some external force demands it, they may not apply their skills to solve the problem. In conceptualizing critical thinking disposition, this example shows that such individuals do not have a strong disposition toward critical thinking and are not internally motivated to use their cognitive skills to make purposeful judgments about what to believe or do in a given situation.

The analysis of the Delphi report reveals that both the definitions of critical thinking and of critical thinking disposition trace their way back to the documentations of John Dewey, Karl Popper, and Paulo Freire. For instance, Dewey describes the dispositional aspect of thinking as "personal attributes" (Dewey, 1910). According to Popper (1935) and Freire (1974) however, critical thinking *attributes* should primarily be considered as a reform strategy in education instead of critical thinking *skills*. Facione et al. (1995) further suggest that there is a "characterological profile, a constellation of attitudes, a set of intellectual virtues, and a group of habits of mind which we refer to as the overall disposition to think critically" (p. 2). In the Delphi study, these intellectual virtues and habits of mind have been characterized as "truth-seeking," "open-mindedness," "analyticity," "systematicity," "inquisitiveness," "critical thinking self-confidence," and "maturity of judgment." These virtues are considered as the *characteristics* of an "ideal critical thinker." Further effort, indeed, in defining ideal critical thinker has produced the following definition:

The ideal critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgments, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are precise as the subject and the circumstances of inquiry permit (Facione, 1990, p. 3).

The CCTDI, as the end product of the Delphi effort, aims at assessing the dispositional dimension of CT. According to the 46 experts on critical thinking participating in the Delphi study, the CCTDI represents a high degree of fit between the current conceptualization and measurement development of critical thinking dispositions. The seven affective dispositions that he CCTDI attempts to assess are shortly defined as follows:

1. Truth-seeking: is to "seek the truth, courageous about asking questions, and honest and objective about pursuing inquiry, even if the findings do not support one's interests or one's preconceived opinions"
2. Open-Mindedness: is to be "open-minded and tolerant of divergent views with sensitivity to the possibility of one's own bias."
3. Analyticity: is to be "alert to potentially problematic situations, anticipating possible results or consequences, and prizing the application of reason and the use of evidence even if the problem at hand turns out to be challenging or difficult."
4. Systematicity: is to be "organized, orderly, focused, and diligent inquiry in inquiry."
5. CT Self-Confidence: refers to "the level of trust one places in one's own reasoning processes."
6. Inquisitiveness: is to have "intellectual curiosity by means of valuing being well informed and learning, even if the immediate payoff is not directly evident."
7. Maturity of Judgment: is to make "reflective judgments based on cognitive maturity and epistemic development" (Facione & Facione, 1992a, pp. 11-12).

Assessing the dispositional dimension of CT has gained more importance than assessing the cognitive skills dimension. John Dewey, in *How We Think*, expresses, "If we were compelled to make a choice between these personal attributes and knowledge about the principles of logical reasoning together with some degree of technical skill in manipulating special logical processes, we should decide for the former" (1910, p. 34). Moreover, the motivational theory of Kurt Lewin presents the theoretical framework for the assumption that the disposition to value and employ CT would impel an individual to lead mastery over CT skills, being motivated to close the gap between what is valued and what is attained (Lewin, 1935).

As explained above, significant effort has been exerted in order to conceptualize critical thinking and to establish the theoretical foundation of the CCTDI. However, the development of the CCTDI has further continued by means of several other efforts. Development continued with generating measurement items from each of these 7 content domains that represent 7 dispositional aspects of critical thinking, which, in turn, have established the uni-dimensional assessment model after the necessary pilot tests and factor analyses were carried out within the mainstream of the Delphi effort.

When operationally evaluated, the CCTDI is composed of 75 items rated on a 6 point, forced choice scale (1 = totally disagree, 2 = disagree, 3 = partially disagree, 4 = partially agree, 5 = agree, 6 = totally agree) and intends to measure 7 dimensions of critical thinking dispositions with 7 sub-scales. The 6 point dichotomous forced choice scale intended to group respondents into two main categories, such as those who agree and who disagree, and also intended to measure the extent of agreement or disagreement within each category.

The Delphi study reported alphas for the sub-scales of the CCTDI: (1) Truth-seeking (12 items, α = .72); (2) Open Mindedness (12 items, α = .73); (3) Analyticity (11 items, α = .72); (4) Systematicity (11 items, α = .74); (5) Critical Thinking Self-Confidence (9 items, α = .78); (6) Inquisitiveness (10 items, α = .80); (7) Maturity of Judgment (10 items, α = .75); and overall scale (75 items, α = .90) (Facione, 1990).

The scale scores of the CCTDI range between 10 and 60 and are interpreted as follows: Scale scores in the 10 to 29 range indicate a low disposition; scores in the 30 to 39 range indicate an ambivalent disposition; scores in the 40 to 49 range indicate a positive disposition; and scores in the 50 to 60 indicate a high disposition (Facione & Facione, 1992a). The overall scores of the CCTDI range between 70 and 420 and are interpreted on the basis of the following standards: A total score falling in the 70 to 209 range signifies a negative disposition toward critical thinking; a total score falling in the 210 to 279 range signifies ambiguity or an ambivalence toward critical thinking; and a score falling in the 280 to 420 range signifies a positive disposition toward critical thinking (Facione & Facione, 1992a). As explained in the test manual, although the ranges defined for the interpretation of scores are considered to be universal, the ranges may also be arranged or adapted on the basis of what normative standards are held by any group to which the CCTDI will potentially be administered. The scoring procedures and particulars of score calculations have not been revealed due to international copyrights.

Following the development of the CCTDI, other researchers, especially psychologists have shown great interest in the CCTDI. Enthusiasm in understanding the interrelation of such conceptualization with pre-existing concepts has led to studies being conducted with the goal of seeking correlations between the CCTDI and other research instruments and constructs already made available, such as "openness to experience" (Costa & McCrae, 1985) and "ego-resiliency" (Block & Block, 1980). Sánchez (1993) found positive correlations between the scales of CCTDI and ego resiliency: Systematicity (r=.47, N=200, *p*<.001), Truth Seeking (r=.41, N=200, *p*<.001), and Inquisitiveness (r=.39, N=200, *p*<.001); as well as with the openness to experience construct: Truth-Seeking (r=.27, *p*<.001), Open-mindedness (r=.33, *p*<.001), CT Self-Confidence (r=.25, *p*<.004), Inquisitiveness (r=.37, *p*<.001), and Cognitive Maturity (r=.30, *p*<.001).

The US Department of Education (DOE) investigated the assessment measurers of student critical thinking dispositions and reviewed all of the inventories available in terms of several criteria. As a result of this study, US DOE released a national report (2000), entitled *Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing*, which included inventories and their specifications. The report also indicated that there are only a few instruments developed to measure critical thinking but that none of these instruments are designed to measure critical thinking disposition or professional judgment in teacher education, except for the CCTDI, which is well conceptualized and developed to measure the extent to which a person possesses the characteristics of the ideal critical thinker.

Authorized Arabic, Chinese (Mandarin), Dutch, Farsi, Finnish, French (Canadian), Hebrew, Italian, Japanese, Korean, European Portuguese, Spanish (Mexico-Latin America), and Thai language versions of the CCTDI are currently available. As mentioned in the former parts of this manuscript, the English version of the CCTDI displayed positive psychometric properties for use with English speaking American populations. However, the other language versions of the CCTDI were not tested in terms of their psychometric properties, except for the Chinese version by Yeh (2002). Yeh reported positive alpha coefficients for two of the sub-scales of the Chinese version of the CCTDI: (Inquisitiveness, 0.73 & self-confidence, 0.68). For the other sub-scales, the alpha coefficients ranged between 0.34 and 0.47, which, since they were below 0.50, were considered to be in need of refinement through further developmental actions. In addition, Yeh found that the measurement model hypothesized by Facione (1990) did not fit the Chinese data well. The results of his study suggested further adaptation and validation was needed in order to render useful all of the dispositional dimensions of the CCTDI with Chinese samples. Insight Assessment, a division of California Academic Press, contacted the researcher of this study and approved him as the authorized translator to produce the Turkish language version of the CCTDI and to analyze the psychometrics for Turkish and American samples for cross-cultural validation.

Method

The current study utilizes a descriptive design, supported by the measurement theory and psychometric theory (Nunnally & Bernstein, 1994) to cross-culturally validate the CCTDI. The main premise behind employing the cross-cultural psychometric methodology is to provide evidence regarding whether the results obtained from a target language version of the inventory is due to errors in translation or due to true differences in the people or the variables being measured (Hambleton, 2005). Chapman and Charter (1979) stated that psychometric equivalency could be investigated by examining the measurement invariance across cultural groups. As such, in order to obtain an informed opinion of the psychometric properties of both tests, the reliability and validity of both the agreed upon Turkish version as well as the original English version were studied.

Purposes of the Study and Research Questions

The purposes of the current study were: (1) to translate the CCTDI from its English source language into the target language (Turkish); (2) to study the psychometric properties of both the Turkish and English versions of the CCTDI; (3) to assess the factorial validity via Confirmatory Factor Analysis (CFA); and (4) to assess the measurement invariance (MI) across the Turkish and American populations. The following research questions were addressed in order to achieve the purposes specified for the research:

1. Given findings regarding the necessary statistical analysis, what do both the translated Turkish and the original English versions of the CCTDI demonstrate in terms of their psychometric properties?
2. Given findings regarding the confirmatory factor analysis (CFA), what is the extent to which the data derived from both the Turkish sample and the American sample explain the hypothesized 7-factor measurement model of the CCTDI?
3. Given findings regarding the measurement invariance tests, what is the extent to which the translated Turkish and the original English versions of the CCTDI allow for cross-cultural mean comparison of the construct?

Translation and Back-translation Process

Prior to any translation attempt, all the necessary permissions to translate the CCTDI had been obtained from Insight Assessment, a division of California Academic Press who is the current copyright holder of the instrument, and the author of this article was commissioned as the *authorized* translator for the Turkish version of the CCTDI.

The first attempt in undertaking the process of translation was to clarify the author's intended meaning for each item in the original CCTDI. Having an extensively informed opinion regarding the intention behind each item prior translation was just as important for maximizing the semantic, conceptual, and normative equivalencies as it was for minimizing item bias across the various language versions (Ægisdóttir et al., 2007; Behling & Law, 2000). For that matter, each individual item in the inventory was negotiated with Peter Facione, author of the inventory; after which detailed, informed opinions regarding the intended meaning of each item in the original CCTDI were considered thereby leading to an ultimate, definitive decision being agreed upon.

Following meaning clarification, initial translation and back-translation process took place. For the current study, the translation and back-translation process, as suggested by Brislin (1970), was embedded into an interactive adaptation process in order to maximize translation equivalency, in which each cycle involved three important steps: (1) Initial translation, (2) Back-translation, and (3) Comparison of the original and back-translated versions for any modification and adaptation. Therefore, the following multiple interactive translation process was used to produce a linguistically valid and succinct Turkish version of the CCTDI (see Figure 1).

Cycle 1 – Step 1: First, the author of this article translated the CCTDI from its English source language into the Turkish target language (English Version 1 into Turkish Version 1).

Cycle 1 – Step 2: Second, the translator, who is bilingual and has a background in regard to the field of study, back translated Turkish Version 1 into English Version 2 without any knowledge of English Version 1.

Cycle 1 – Step 3: The English Version 2 back-translation was then compared to English Version 1 by Peter A. Facione and a panel selected by the California Academic Press (Developers and Copyright Holders of the Instrument).

1st Interactive Response: As a result of the 1st comparison stage, reviewers sent a list of items that need to be changed, revised, or responded to. According to the response, 17 items were found to be problematic in terms of linguistic equivalency, thus entailing that the intention and/or actual message was not found in those specified 17 items in the English version 2 back-translation. On the basis of the 1st response received from the California Academic Press (CAP), each of these 17 items were negotiated online in order to detect the root of the problem. As a result of this attempt, we concluded that some items in the Turkish version did not include the intended message because of lexical preference, and that the actual intentions of several items were unable to be merged into Turkish target language because of normative, conceptual, and semantic problems. In addition, some of the original English items included proverbs that did not exist in Turkish culture. Therefore, these items needed to be adapted to fit Turkish culture while also protecting their original intentioned meanings (Weeks, Swerissen, & Belfrage, 2007).

Cycle 2 – Step 1: Regarding the decisions drawn from the first interactive response, the first author made the necessary adaptations and then rewrote the specified 17 items by considering the nuances in Turkish culture while giving special attention to preserve the original intention, which led to Turkish Version 2.

Cycle 2 – Step 2: Regarding the suggestions of the related literature (Herrera, DelCampo, & Ames, 1993; Weeks et al., 2007), each time a new version was to be made, a new, independent translator was attained to proceed with a new translation of the instrument. For this reason, a third independent translator, who was also bilingual and who studied in the field of higher order thinking, back translated the Turkish Version 2 into English Version 3 without any knowledge of either English Version 1 or Turkish Version 1.

Cycle 2 – Step 3: The English Version 3 back-translation was sent to the panel of experts in CAP to be compared with English Version 1 for any instances of non-equivalency.

2nd Interactive Response: The second translation attempt was deemed successful in terms of reducing the number of problematic items from 17 to 3. The remaining 3 items were English proverbs which for which modifications were insufficient to provide Turkish equivalents that conveyed the original intentions of the English items. After further negotiation with Peter A. Facione and the expert panel, we decided to completely change these items in such a way that would both preserve the original message while also being expressed as a Turkish proverb.

Cycle 3 – Step 1: On the basis of the decision made during the second interactive response stage, attention was directed to produce 3 new Turkish items that could be considered identical with the original English items in terms of the construct being measured and which were expressed in the form of proverbs. However, this attempt required more time and effort in order to meet the conditions. We knew that each of these three items intended to measure a different facet of critical thinking disposition; therefore, the construct that each item intended to measure was different in terms of the message conveyed. For this reason, we generated at least 15 new optional items in the form of proverbs that might be considered as alternatives. In doing so, several factors, such as the original items, the meaning clarification report sent by CAP, the experience gathered from the interactive response stages, and the theory based sources regarding the phenomenon being translated were jointly considered (Behling & Law, 2000). Prior to composing a third Turkish version, selected the best 3 items among the many items generated for each single construct. As a result of the experiences drawn from this multiple interactive translation process, the 3rd Turkish version of CCTDI emerged.

Cycle 3 – Step 2: As the cross-control check step, the latest Turkish Version 3 was then back translated into English Version 4 by a different independent

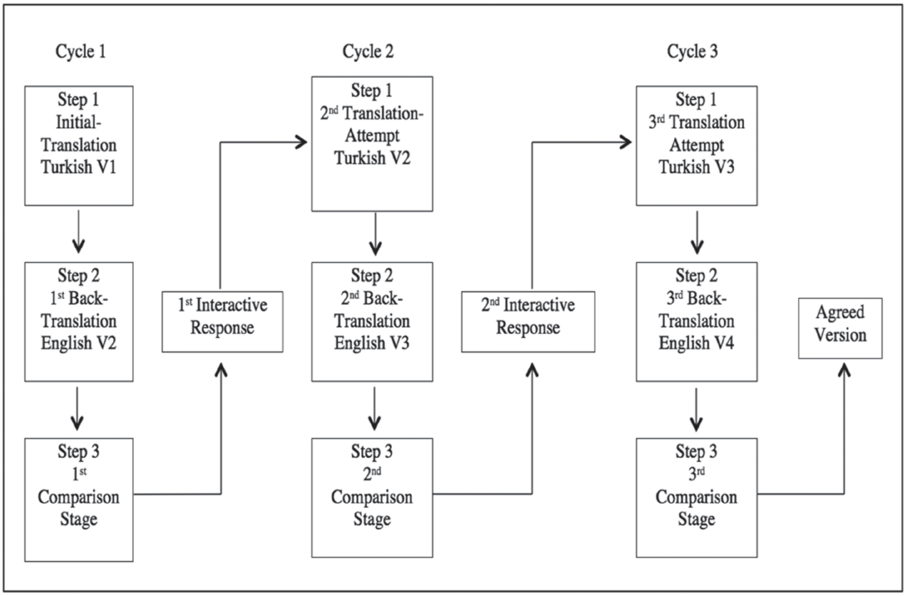
Cycle 1
Cycle 2
Cycle 3
Step 1 Initial-Translation Turkish V1
Step 2 1st Back-Translation English V2
Step 3 1st Comparison Stage
1st Interactive Response
Step 1 2nd Translation-Attempt Turkish V2
Step 2 2nd Back-Translation English V3
Step 3 2nd Comparison Stage
2nd Interactive Response
Step 1 3rd Translation Attempt Turkish V3
Step 2 3rd Back-Translation English V4
Step 3 3rd Comparison Stage
Agreed Version

Figure 1.
Multiple Interactive Translation Back-Translation Process

bilingual translator, who is an expert in translating proverbs, without any knowledge of the original text or any of the previous Turkish versions.

Cycle 3 – Step 3: The English Version 4 back-translation was sent to CAP for further evaluation of the three items in question. This was the last interactive comparison stage run with CAP; and as a result of their evaluation of those three items, they reached a consensus that the Turkish Version 3 was the Turkish version identical to the CCTDI.

For the current research however, finalizing this translation and the adaptation phase were not enough to prove that the agreed Turkish version of the CCTDI was cross-culturally valid. Any translation of inventories for cross-cultural use needs to be followed by a cross-cultural equivalency study in order to assure what is being measured exists and is functionally equivalent across cultures (Brislin, 1970; DiStefano & Hess, 2005; Nunnally & Bernstein, 1994).

Sampling and Data Collection Procedures

For the purpose of the current research study, the original English version and the translated Turkish version of the CCTDI needed to be administered to a representative number of American and Turkish participants in order to study the psychometric properties, to check for functional equivalency, and to explore for comparability of the scores of the CCTDI across Turkish and American samples. Therefore, it was decided that the Turkish CCTDI should be administered to Turkish participants in Turkey and that the original English version of CCTDI should be administered to American participants in the United States. Hambleton's research (2005) holds that the administration of dual language versions of measurement instruments should take place in their natural settings by native speakers of the language of the tests.

For the American sample, the study was presented to department chairs of 15 Universities in the United States. Three universities gave permission to collaborate for the current research on the condition that they would participate in and earn passing scores on an online course offered to international researchers. This online course offered by the Collaborative International Training Initiative (CITI) over a three month period that discussed the ethics and politics of carrying out research in the United Sates which included modules on vulnerable human resources and historical perspectives. The researcher took an online exam after completing each module. The condition was to earn a score of at least 80 out of 100 in order to pass each module of the online course, which was composed of 11 distinct modules. After completing all of the modules successfully and receiving the Human Subjects Research Curriculum Report, research was presented to the potential participants in those 3 universities who were informed about the confidentiality of their answers and whose permission was sought via a consent form to participate in the study. The names of the universities and participants were not mentioned anywhere in this study as they were confidential and under obligation not to be mentioned. The same consent form was translated into a Turkish version and the study was introduced to potential participants in a university in Turkey.

Regarding the purposes of the current adaptation study, a judgmental participant selection procedure was employed. The concern in regard to cross-cultural judgmental sampling designs is based on placing attention on only controlling key variables (Hamleton, 2005; Sekaran, 1983). Such key variables represent major characteristics of the pilot and/or validation sample(s) (Sekaran, 1983). The key variables for both samples were native language, culture, and field of study (major). In other words, being enrolled in a teacher education program, being a native speaker of the language of the inventory, and identifying oneself as Turkish or American were enough to be a participant in this study. However, in order to determine which subjects were representative of the central tendencies of the country and culture, extra effort was exerted in order to judge the subjects in terms of several criteria, including place of birth, amount of time spent in the country, languages they speak, and reasons for identifying themselves as Turkish or American.

The corresponding literature suggests the ideal of reaching at least 400 (n ≥ 400) participants for each language and culture versions in order to be validated (Hu & Bentler, 1999; Tucker, Ozer, Lyubomirsky, & Boehm, 2006). Specifically, Tucker et al. (2006) stated that some tests, such as the chi-square test, confirmatory factor analysis tests, and invariance tests, are sensitive to sample size and therefore suggested a sample group of at least 400 (n ≥ 400) participants for a successful cross-cultural comparison of hypothetical assessment structures and adaptation studies.

Based on the standards and guidelines suggested above, a Turkish sample group composed of 583 (n=583) freshmen to senior undergraduate students (51.3% female) from 5 different teacher education programs was obtained in Turkey. Another sample of 448 (n=448) freshmen to senior undergraduate students (54.7% female) was obtained from 5 different teacher education programs in the United States. The Turkish sample included 231 (39.6%) freshman, 179 (30.7%) sophomore, 124 (21.3%) junior, and 49 (8.4%) senior students, whose ages ranged from 18 to 29, with a mean age of 19 (SD = 1,72). The American sample, which was diverse in terms of ethnicity but relatively equal in terms of program status (see Table 1), which included 121 (27.0%) freshmen, 116 (25.9%) sophomore, 106 (23.7%) junior, and 105 (23.4%) senior students, whose ages ranged from 19 to 56, with a mean age of 21 (SD = 4,04).

Table 1.
Demographic Characteristics of Participants

Characteristics	American Sample		Turkish Sample	
	n	(%)	n	(%)
Ethnicity				
African American	106	(23.7)	0	(0.0)
Anglo American, Caucasian	161	(35.9)	0	(0.0)
Asian American	87	(19.4)	0	(0.0)
Hispanic, Latino, Mexican	61	(13.6)	0	(0.0)
Native American	33	(7.4)	0	(0.0)
Turkish	0	(0.0)	583	(100)
Gender				
Female	245	(54.7)	299	(51.3)
Male	203	(45.3)	284	(48.7)
Age				
18	0	(0.0)	202	(34.6)
19	147	(32.8)	178	(30.5)
20	95	(21.2)	111	(19.0)
21	89	(19.9)	43	(7.4)
22	64	(14.3)	21	(3.6)
23	11	(2.5)	18	(3.1)
24	10	(2.2)	2	(0.3)
25	3	(0.7)	0	(0.0)
26-35	21	(4.7)	8	(1.4)
36-56	8	(1.8)	0	(0.0)
Major				
Science Education	84	(18.8)	0	(0.0)
Adult Education	96	(21.4)	0	(0.0)
Special Education	87	(19.4)	0	(0.0)
Math Education	81	(18.1)	0	(0.0)
History Education	100	(22.3)	0	(0.0)
English Language Education	0	(0.0)	75	(12.9)
Computer Education	0	(0.0)	92	(15.8)
Elementary Education	0	(0.0)	114	(19.6)
Preschool Education	0	(0.0)	186	(31.9)
Sociology Education	0	(0.0)	116	(19.9)
Program Status				
Freshmen	121	(27.0)	231	(39.6)
Sophomore	116	(25.9)	179	(30.7)
Junior	106	(23.7)	124	(21.3)
Senior	105	(23.4)	49	(8.4)

Administering the CCTDI to Turkish participants required 15 separate sessions lasting a total of 2 months whereas administering the CCTDI to American participants at 3 different Universities in the US required 46 sessions requiring 8 months to complete. Nearly one year of time was allotted to administer the CCTDI to both samples. The administration and completion of the CCTDI for each session took approximately 20 minutes.

The mode of administration and standards had already been identified by Facione and Facione (1992a) in the test manual of the CCTDI. Therefore, this test manual was considered to be the standardized guideline for test administration. For the purpose of administration, these guidelines were shared with the research partners in the research sites who were also informed that the way the CCTDI was to be administered to different participants should depend on the same standards for each session. As Anastasi and Urbina (1997) argued, the same test was to be administered to different participants at different times and different places should be according to the same standards so as to conclude whether the differences observed between variables were due to true differences between the participants as opposed to the mode of administration. Therefore, prior to administration, the professional personal hired for administering the CCTDI were provided with an online training about the standardized mode of administration to be employed.

Procedures for Cross-cultural Data Analysis

The first attempt in undertaking the issue was to assess the content validity of both the Turkish and English versions of the CCTDI. This was required to ensure whether the items adequately measured the content domains that were hypothetically desired to measure (Grant & Davis, 1997; Nunnally & Bernstein, 1994). The actual process required content experts to judge each item against their definitions and was composed of a set of events that needed to be followed in sequential order. For that matter, the first issue was to select the content experts for whom experience in their fields of study, history of publications, research on the phenomenon of critical thinking, and qualifications were considered as selection criteria (Davis, 1992; Drasgow & Probst, 2005; Grant & Davis, 1997; Hambleton, 2005; Polit & Beck, 2006). After three months of negotiations with several universities in Turkey and in the United States, five Turkish experts (n = 5) from Turkey and five American experts (n = 5) from the US were selected from the fields of educational psychology, linguistics, and critical pedagogy to serve as content validators for the translated Turkish version and the original English version of the CCTDI, respectively. Each expert was provided with a set of four different documents (Doc 1: Cover Letter, Doc 2: Content Domains, Doc 3: CCTDI, and Doc 4: Content Validity Estimation Scale). The first document, the 'cover letter,' informed the content validators about the study, the confidentiality of their answers, their roles as content validators, the measurement model of the CCTDI, and detailed information of the other three attached documents. The second document, which included the content domains and element definitions from which all 75 items of the CCTDI were obtained, served as a standard for content validators to compare each item against the definition. In order for experts to see the complete inventory, the translated Turkish version of the CCTDI was given to Turkish experts and the original English version was given to American experts. Experts were also given the content validity index (CVI), developed by Waltz, Strickland, and Lenz (1991), in order to rate each item of the CCTDI for their relevance to content domains (ranging from 1 = not relevant to 4 = very relevant), clarity (ranging from 1 = not clear to 4 = very clear), simplicity (ranging from 1 = not simple to 4 = very simple), and ambiguity (ranging from 1 = doubtful to 4 = meaning is clear) on a four-point scale. The index for accepting a sub-scale or a total instrument as being valid for the specified content was the percentage of sub-scale items or overall scale items by receiving a score of 3 or 4 from each category (Drasgow & Probst, 2005; & Waltz et al., 1991) with a minimum content validity index of .90 (Davis, 1992; Wynd, Schmidt, & Schaefer, 2003; Yaghmaie, 2003).

After examining the CVIs of both versions via expert agreements, the alpha coefficients for the seven sub-scales of the CCTDI were computed with both the American and Turkish samples in order to assess the internal consistency reliability prior to performing any confirmatory factor analysis (CFA) and invariance analysis (IA) attempts. Computing the coefficient alphas for the CCTDI scales before CFA and after CFA with modified items was necessary to verify whether any item trimming led to an inacceptable decrement in alpha coefficients (Nunnally & Bernstein, 1994; Yang & Green, 2011). Since this was the first attempt to translate the CCTDI from its English source language into the Turkish target language, the cutoff point, as recommended, for internal consistency reliability was considered to be greater than .70 (George & Mallery, 2003; Gliem & Gliem, 2003; & Yang & Green, 2011).

A group of participants (n = 53) from the actual Turkish sample (n = 583) and a another group of participants (n = 38) from the actual American sample (n = 448) were asked to take the test two months apart in order to examine time interval test-retest reliability. Pearson's correlation coefficient *r* was calculated to determine the extent to which the two sets of scores (Time 1 Scores and Time 2 Scores) were correlated for both Turkish (n = 53) and American (n = 38) data sets. In addition, paired sample T-Test analysis for comparing the mean scores and F tests for testing the equivalency of variance were used. The most important part in this phase of psychometric analysis was the interval between the first and second administration of the test on the same participants. To be effective, the time interval between the occasions was not allowed to be either too long, so as not to allow the construct to change in participants, as it might naturally, or too short, so that they might remember their previous responses to the items in the CCTDI (Fraenkel & Wallen, 2006). In order not to violate this rule, two months of time were allocated for the time interval as two months is not long enough of a time to cause a significant difference in critical thinking dispositions but is more than enough time for them to forget the responses they had made for the items on the CCTDI. As Metzler and Blankenship (2008) hypothesized, for a construct such as critical thinking disposition to be changed in an individual from one direction to another (from positive to negative or vise versa), one needs to be exposed to a significant change in his/her life spaces, life conditions and standards, and perceptions for a considerably long time. Disposition is defined as a characterological profile or a habit of mind, and for this very reason, scholars do not expect a dramatic change in the intellectual functioning of individuals unless they are affected by an external force, which could also be considered as a series of planned actions (Benesch, 1993; Ernst & Monroe, 2006; Facione, Facione, & Giancarlo, 2000).

Remembering again, the issue of adapting psychological tests into a different language and culture require a decentering approach, meaning that both versions in question for this study were subject to all necessary modifications and refinements (Hambleton, 2005). For that matter, the factorial validity of both the English and Turkish versions of the CCTDI was examined.

Since Confirmatory Factor Analysis (CFA) is considered to be the most advanced technique for testing hypotheses concerning measurement models (Kline, 2005), it was advocated to be the most advanced and strongest way to provide evidence for cross-cultural construct validity of dual language versions of an inventory (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Hambleton, 2005; Kline, 2005; Sireci et al., 2006). In this respect, in order to test the degree of existence of the hypothesized 7 factor uni-dimensional measurement model of the CCTDI across Turkish and American samples, CFA was employed. According to Kline, the primer requirement to utilize CFA for any adaptation and validation research is to ensure whether the development of inventory has a strong theoretical basis. Thus, CFA was found to be the most appropriate analysis technique for the current dissertation study since the current study aimed to adapt and validate the CCTDI, which was originally developed in the English language for use in an American cultural setting and which has a strong theoretical ground hypothesizing the 7 factor measurement model with empirical evidence supporting this structure. In addition, as it is clear by the name of the analysis, the aim is to ascertain to what extent the hypothesized measurement model exists in the targeted samples of corresponding cultures and languages. Since there is more than one sample, the sort of analysis is also labeled as "multiple sample confirmatory factor analysis" (Kline, 2005; Nunnally & Bernstein, 1994; Schreiber, Nora, Stage, Barlow, & King, 2006). To better comprehend the basic assumptions of the CFA, the following guideline has been suggested by Kline (2005): If the hypothesized 7 factor measurement model is correct for both samples, the analysis of results should yield less than .95 estimated correlations between the indicators of each factor. Apart from this, if the model is correct, then, each indicator, which intends to measure an underlying factor, should possess standardized loadings preferably higher than .30. The former refers to the discriminant validity if the correlations between factors are smaller than .95, and the later refers to convergent validity. To conduct CFA, Pearson Product Moment Correlations were preferred instead of polychoric correlations because polychoric correlations were found to be non-practical in improving the model fit during the nested model analysis stage (Chen, 2007; Sun, 2005). The current researcher preferred to use the following indices and cutoff points to evaluate the model fit of the 7-factor model of the CCTDI on the basis of the related literature (Chen, 2007;

Hu & Bentler, 1999; Milfont & Fisher, 2010; Sass, 2011): chi-square (χ^2), degrees of freedom (df), the ratio of chi-square to degrees of freedom (χ^2/df < 4.0), the root mean square error of approximation (RMSEA < .06 = good fit; values between .06 and .08 as adequate fit; and values between .08 and .10 as mediocre fit), the standardized root mean square residual (SRMR < .06 = good fit; values between .06 and .08 as adequate fit; and values between .08 and .10 as mediocre fit), and comparative fit index (CFI ≥ .90 = adequate fit; and values greater than .95 as good fit).

The evaluation of the existence of the 7-factor structure of the CCTDI via an initial confirmatory factor analysis did not provide evidence for any extent of comparability of mean scores across cultural groups (multiple sample analysis). The actual concern here was to test the measurement invariance, defined by Kline (2005) as "whether a set of indicators assess the same construct in a same way across different groups" (p. 295). In order for a researcher to make use of the different language versions for the same instrument in order to perform reliable, valid cross-cultural mean comparisons of the related construct, the measurement invariance of the instrument should be tested (Behling & Law, 2000; Chen et al., 2005; Cheung & Rensvold, 1999; Hambleton, 2005; Kline, 2005; Sass & Schmitt, 2010; Tucker et al., 2006). Kline clearly explains the typical practice of measurement invariance analysis and, as he puts it, the analysis involves "the comparison of the relative fits with the χ^2 difference statistic of models, one with cross-group equality constraints imposed on some of its parameters and the others without constraints" (p. 295).

For testing measurement invariance across cultural groups, the criteria suggested by Chen (2007) were adopted. According to Sass (2011), to develop an accurate judgment regarding the invariance model fit, ΔRMSEA, ΔSRMR, and ΔCFI should be considered in addition to the ($\Delta\chi^2$) statistic because the use of the ($\Delta\chi^2$) statistic alone may be misleading and is also very sensitive to large sample sizes with complex measurement models. Therefore, more practical criteria for accepting an invariance model fit involved: ΔRMSEA ≤ .015, ΔSRMR ≤ .03, and ΔCFI ≤ .01 for interpreting the results for tests of factor loading invariance and ΔRMSEA ≤ .015, ΔSRMR ≤ .01, and ΔCFI ≤ .01 for interpreting the results for tests of intercept invariance and residual invariance (Chen, 2007; Sass, 2011). Both SPSS version 18 and IBM AMOS version 20 were used to run the required statistical analyses throughout the study.

Results

Psychometric Properties

Following the translation and back translation process, the second phase of the adaptation process was to assess the psychometric properties of the Turkish and English versions of the CCTDI. Since the decentering approach was utilized, both versions of the inventory were subject to the necessary adaptations. In order to test the psychometrics of the CCTDI for both samples, the following research question was addressed:

Research Question 1: Given the findings regarding the necessary statistical analysis, what do both the translated Turkish and the original English versions of the CCTDI demonstrate in terms of their psychometric properties?

The first psychometric check was carried out to gauge the content validity of both the Turkish and English language versions. The goal was to comprehend to what extent each item represented the corresponding latent factor's content domains. The CVIs ranged from 0.83 to 0.99 for the subscales of the Turkish CCTDI and ranged from 0.97 to 1 for the subscales of the English CCTDI. Evidence for content validity in the sub-scales existed across versions. Although the raters agreed that the items were relevant to the content domains and definitions specified for each corresponding latent factor, a lower level of agreement was found for the open mindedness scale in the Turkish version (see Table 2).

Table 2.
Content Validity Indices and Alpha Coefficients for Sub-Scales across Turkish and English Versions of the CCTDI

Sub-Scales (number of items)	Content Validity Indices		Alpha Coefficients	
	American (n=5)	Turkish (n=5)	American (n=448)	Turkish (n=583)
Truth-Seeking (12)	1	.93	.88	.85
Open Mindedness (12)	.97	.83	.87	.82
Analyticity (11)	.98	.90	.91	.90
Systematicity (11)	.98	.93	.89	.86
Inquisitiveness (10)	1	.96	.86	.86
CT Self Confidence (9)	.99	.99	.88	.88
Maturity of Judgment (10)	.98	.96	.85	.81
Overall (75)	.99	.93	.87	.87

The standard deviations and means for each item, as well as for sub-scales, were computed to determine the central tendencies prior to performing a reliability analysis. When the alpha coefficients were studied with 75 items prior to a CFA attempt, the alphas for the sub-scales ranged from .81 to .90 for the Turkish CCTDI and ranged from .85 to .91 for the English CCTDI (see Table 2). These values satisfied the minimum expected criteria of .70 for the first attempt at translating and adapting the scales (George & Mallery, 2003). It should, however, be noted that the coefficient alphas were recalculated with the latest versions of these sub-scales after a subsequent CFA for cross-cultural comparability.

The third type of reliability analyzed for both cultural samples was the time interval test-retest reliability. When the mean score results, that is, the Pearson correlation coefficients, the results of the *t*-tests, and the variance analysis, were evaluated for the Turkish sample (N = 53), evidence supporting test-retest reliability the Turkish CCTDI was found. As can be gathered from the table (see Table 3), all the Pearson's r statistics were statistically significant at a 0.01 significance level, ranging from a high of 0.57 for the analyticity sub-scale to an even higher score of 0.73 for the critical thinking self-confidence sub-scale. For a more restricted analysis, the significance level was then adjusted to 0.001. In the second run however, the Pearson correlation coefficients showed no significant difference in the correlations across the time-1 and time-2 scores. The *t*-test result also showed no significant difference in the mean scores across the specified time interval. In addition to this analysis, since there was no significant difference detected in the variances of related samples, the F test results revealed that there was an equality of variances. In other words, the differences observed in variances were non-significant at a significance level of 0.001. Thus, a considerable degree of stability among scores was evident between the related Turkish samples.

When the same procedures were repeated for the American sample (N = 38), evidence for test-retest reliability existed for the American CCTDI as well. As can be observed from the table (see Table 4), all the Pearson's r statistics were statistically significant at the 0.01 significance level, ranging from a high of 0.52 for analyticity sub-scale to an even higher score of 0.79 for critical thinking self-confidence sub-scale. Only a lower score of 0.42 in the Pearson's correlation coefficient for systematicity sub-scale was found in the American sample. Overall however, a significant level of test retest reliability existed. Thus, there was no significant difference in the mean scores across time during the two different occasions. Likewise, since there was no significant difference detected in the variances of related American samples, the F test results showed that there was an equality of variances. That is to say, the differences observed in variances were non-significant at a significance level of 0.001. Therefore, a considerable degree of variability and stability among the scores were evident between the related American samples as well.

CFA: Factorial Validity of the CCTDI across Turkish and American Samples

In order to conduct a further analysis regarding the psychometric properties of the CCTDI, the

Table 3.
Correlations, Paired t Test, and Variances for Sub-Scale Scores of the Turkish CCTDI across Time

Subscale	Mean (SD)	r	t-test (df)	variance	f-test (df)
Truth-Seeking					
Time 1	45.36(7.50)	0.60[a]	1.60(52)	56.31	2.58(51)
Time 2	43.85(7.64)			58.36	
Open-Mindedness					
Time 1	42.53(6.96)	0.70[a]	-0.62(52)	48.48	0.38(51)
Time 2	43.00(7.39)			54.65	
Analyticity					
Time 1	44.89(4.15)	0.57[a]	1.54(52)	17.27	2.37(51)
Time 2	44.04(4.49)			20.19	
Systematicity					
Time 1	39.77(5.44)	0.59[a]	1.62(52)	29.64	2.63(51)
Time 2	40.85(5.21)			27.13	
Inquisitiveness					
Time 1	41.23(4.50)	0.59[a]	0.60(52)	20.22	0.36(51)
Time 2	40.87(4.98)			24.81	
CT-Self-Confidence					
Time 1	40.28(5.99)	0.73[a]	0.29(52)	35.94	0.085(51)
Time 2	40.11(5.55)			30.83	
Maturity					
Time 1	31.64(5.30)	0.59[a]	-1.30(52)	28.08	1.71(51)
Time 2	32.58(6.21)			38.56	
Overall					
Time 1	285.70(23.29)	0.60[a]	0.13(52)	542.37	0.017(51)
Time 2	285.30(25.08)			629.18	

[a] $p < 0.01$. (N=53)

Table 4.
Correlations, Paired t Test, and Variances for Sub-Scale Scores of the English CCTDI across Time

Subscale	Mean (SD)	r	t-test (df)	variance	f-test (df)

Truth-Seeking
Time 1
Time 2
	29.81(6.49)
	31.17(6.74)
		0.68[a]	0.69(37)	34.71
				18.96
					2.78(36)

Open-Mindedness
Time 1
Time 2
	39.00(3.82)
	38.02(4.02)
		0.71[a]	-0.65(37)	22.32
				24.30
					0.36(36)

Analyticity
Time 1
Time 2
	41.54(3.43)
	40.31(3.83)
		0.52[a]	1.66(37)	17.27
				25.19
					1.62(36)

Systematicity
Time 1
Time 2
	37.53(5.56)
	35.87(4.13)
		0.42[a]	2.67(37)	20.60
				23.36
					0.67(36)

Inquisitiveness
Time 1
Time 2
	47.66(6.13)
	40.87(4.89)
		0.50[a]	2.40(37)	25.22
				33.76
					1.24(36)

CT-Self-Confidence
Time 1
Time 2
	41.39(5.66)
	40.52(5.94)
		0.79[a]	0.99(37)	44.78
				47.33
					0.33(36)

Maturity
Time 1
Time 2
	37.53(3.08)
	34.11(4.02)
		0.65[a]	- 2.66(37)	23.38
				23.21
					0.03(36)

Overall
Time 1
Time 2
	274.46(18.56)
	260.87(24.43)
		0.84[a]	3.76(37)	438.60
				546.09
					1.88(36)

[a]$p < 0.01$. (N=38)

factorial validity of the CCTDI across the Turkish and American samples was examined. To determine the fit of the hypothesized measurement model of the CCTDI for both cultural groups, the following research question was addressed:

Research Question 3: Given the findings regarding the confirmatory factor analysis (CFA), to what extent do the data derived from both the Turkish and American samples explain the hypothesized 7-factor measurement model of the CCTDI?

When the standardized estimates were taken into consideration to examine factorial validity, the hypothesized 7-factor measurement model produced quite a poor fit for the Turkish sample, χ^2(df = 2679) = 10090.724, $p< .0001$, χ^2/df = 3.767, RMSEA = .069, SRMR = .096, CFI = .66, as well as for the American sample, χ^2(df = 2679) = 10566.346, $p< .0001$, χ^2/df = 3.944, RMSEA = .081, SRMR = .100, CFI = .61.

When the regression slopes and the correlation matrix were examined to ascertain the reason behind this poor model fit, the modification index suggested the need to exclude three factors from the measurement model, namely "Open-Mindedness," "Analyticity," and "Inquisitiveness" because of their exceedingly low estimation effects on parameter estimates, specifically on factor loadings and factor pattern coefficients. When those selected factors were removed from the model, although the adapted four-factor model produced better results, the level of improvement was insufficient for both the Turkish sample, χ^2(df = 813) = 3019.200, $p < .0001$, χ^2/df = 3.714, RMSEA = .068, SRMR = .083, CFI = .76, and the American sample, χ^2(df = 813) = 3279.212, $p < .0001$, χ^2/df = 4.033, RMSEA = .082, SRMR = .093, CFI = .72. The fit of the four-factor model to the Turkish sample was slightly better in comparison to the American sample; however, it still did not meet the criteria to be considered a good model fit.

Further consideration of the standardized estimates for both samples revealed several items with exceeding low standardized factor loadings smaller than .30, indicating that these items might not belong to the corresponding hypothesized latent factors. For this reason, with the intention of increasing the factorial validity for both versions, five items from the truth seeking sub-scale (item12, item19, item23, item50, & item62), five items from the systematicity sub-scale (item4, item29, item37, item58, & item68), four items from the critical-thinking self confidence sub-scale (item10, item16, item18, item56), and six items from the maturity of judgment sub-scale (item3, item7, item11, item14, item53, & item71) with factor loadings smaller than .30 were deleted. After deleting these items, the modified hypothesized model displayed a significant, albeit still insufficient, improvement for both the American sample, χ^2(df = 203) = 682.324, $p < .0001$, χ^2/df = 3.361, RMSEA = .073, SRMR = .072, CFI = .91 and the Turkish sample, χ^2(df = 203) = 730.348, $p < .0001$, χ^2/df = 3.598, RMSEA = .067, SRMR = .060, CFI = .92. When the modification indices were carefully evaluated, it was discovered that correlating 3 residuals, along with their pairs

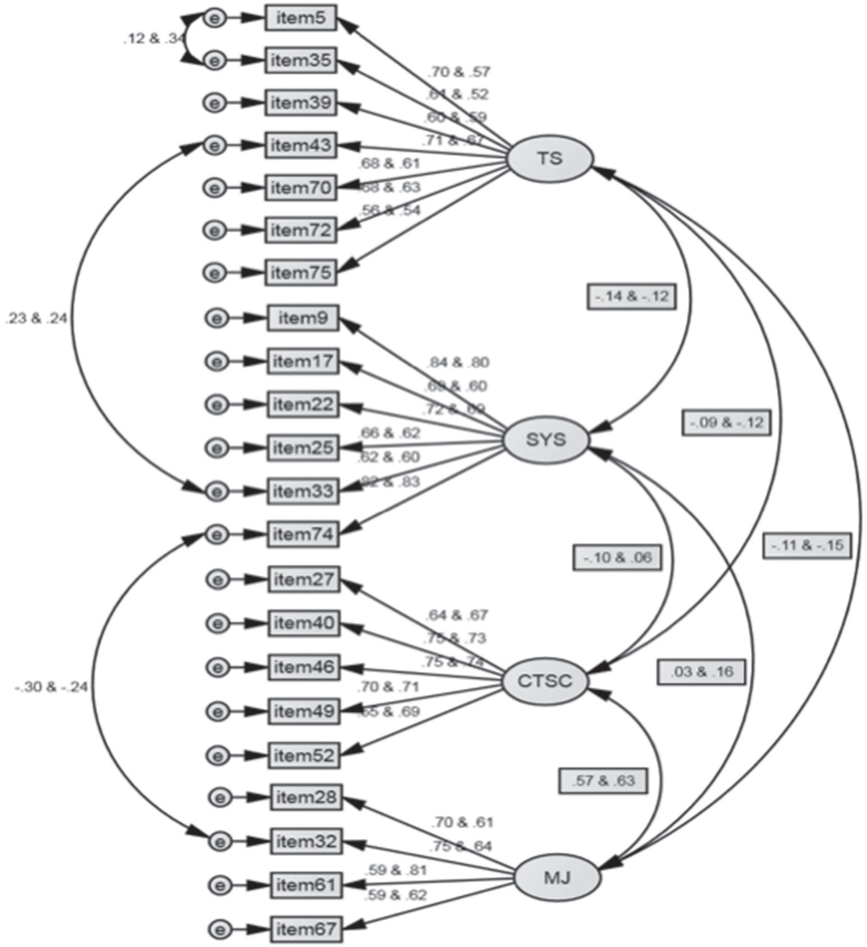

Figure 2.

Modified Four-Factor Model of California Critical Thinking Disposition Inventory with Parameter Estimates for Each Cultural Group Separately.

Note: The first numbers always refer to the American sample. TS: Truth-Seeking Scale; SYS: Systematicity Scale;
CTSC: Critical-Thinking Self Confidence Scale; MJ: Maturity of Judgment Scale.

(see Figure 2), produced an acceptable, adequate fit for the American sample and a good fit for the Turkish sample (see Table 5), which, in turn, produced a better baseline to establish a better Configural model for subsequent nested model comparison and further invariance analysis. The Pearson Moment Product Correlations, means, and standard deviations for both cultural groups supported the suggested four-factor model configuration.

When the alpha coefficients were re-computed after CFA with the four-factor model, the alphas for the sub-scales of the Turkish and English versions of the CCTDI were as follows, respectively: (Truth-seeking, 7 items, α = .80; Systematicity, 6 items, α = .75; CT Self-confidence, 5 items, α = .83; Maturity of judgment, 4 items, α = .77) and (Truth-seeking, 7 items, α = .84; Systematicity, 6 items, α = .87; CT Self-confidence, 5 items, α = .81; Maturity of judgment, 4 items, α = .75). The modification suggested that CFA produced decrements in the alphas of all scales across cultural groups. The highest amount of decrement among the scales was recorded in the maturity of judgment scale of the English CCTDI and in both the systematicity scale and the maturity of judgment scale of the Turkish CCTDI. All of the alphas for the sub-scales across both cultural groups were above the critical point of .70 and displayed evidence of stability for internal consistency reliability.

Cross-cultural Validity: Measurement Invariance of the CCTDI across Cultures

As the final stage of the cross-cultural adaptation process, the degree of measurement invariance across cultural groups was to be tested. Although this was to be one of the most complicated statistical analysis, it was highly necessary in order to provide evidence regarding the comparability of the possible results by means of obtaining scores from both cultures using the CCTDI for further inferential statistical analysis. Since this was one of the purposes of the current dissertation study, the following fourth research question was addressed:

Research Question 4: Given the findings regarding the measurement invariance tests, to what extent do the translated Turkish and the original English versions of the CCTDI allow for a cross-cultural mean comparison of the construct?

The first model tested was the configural invariance model (CIM) (see Table 5), which produced a good fit. A separate examination of the modification indices for each cultural group showed that there to be no residuals for any items with a large modification index, indicating that correlating residuals do not result in a significant improvement in model fit. Therefore, the initial CIM served as a baseline model against which the imposition of more restrictive models could then be tested. The next step was to test for the metric invariance model (MIM) by forcing factor pattern coefficients to be equal across cultural groups. The standard sequence for identification of non-invariant items was based on covariance matrices. In this regard, results regarding the comparison of Model 2 = MIM to Model 1 = CIM indicated that constraining the factor loadings across the groups achieved metric invariance from both the statistical $\Delta\chi^2$ perspective and the practical ΔRMSEA ≤ .015, ΔSRMR ≤ .03, and ΔCFI ≤ .03 perspectives (see Model 2).

Once the metric invariance model (MIM) was supported, the scalar invariance model (SIM) was then tested. Here, I set not only the factor loadings, but also the item intercepts, to be equal across groups. The comparison of Model 3 = SIM to Model 2 = MIM produced a statistically significant $\Delta\chi^2$ (see Table 5). Although Model 3 = SIM seemed to be an acceptable model from the practical perspective, a statistically significant $\Delta\chi^2$ meant that item parameters were unequal across groups, leading us to consider to conduct the partial scalar invariance model (PSIM). Vandenberg and Lance (2000) advocated that configural invariance and metric invariance should be satisfied in order to proceed with any further partial invariance models. Once this requirement was met, to identify the items which might cause the model misfit, modification indices were carefully evaluated with the intention of producing a non-significant $\Delta\chi^2$ statistic when compared to Model 2 = MIM in order to produce an acceptable PSIM. After considering the modification indices, in order to produce a non-significant $\Delta\chi^2$ for the Model 4 = PSIM$_{(i75)}$, it was suggested to relax the constraints placed on factor loadings and item intercepts only for item 75. Although the results regarding ΔRMSEA ≤ .015, ΔSRMR ≤ .01, and ΔCFI ≤ .01 displayed evidence for invariance across groups, relaxing item 75 did not reveal a non-significant $\Delta\chi^2$ for Model 4, thereby indicating that the null

Table 5.
Model Fit Statistics across Cultural Groups.

Model	χ^2	df	χ^2/df	$\Delta\chi^2$	Δdf	RMSEA	ΔRMSEA	SRMR	ΔSRMR	CFI	ΔCFI
American	636.592	200	3.183			.070		.070		.949	
Turkish	626.811	200	3.134			.061		.057		.951	
M1: CIM	1263.403	400	3.159			.046		.057		.950	
M2: MIM	1285.525	418	3.075	22.12	18	.045	.001	.058	.001	.950	.000
M3: SIM	1381.119	440	3.139	95.59*	22	.046	.001	.058	.000	.941	-.009
M4: PSIM$_{(i75)}$	1344.265	438	3.069	58.74*	20	.045	.000	.057	-.001	.951	.001
M5: PSIM$_{(i75 \& i17)}$	1318.705	436	3.024	33.18*	18	.045	.000	.058	.000	.951	.001
M6: PSIM$_{(i75 \& i17 \& i25)}$	1311.505	434	3.021	25.98	16	.046	.001	.058	.000	.950	.000
M7: RIM	1344.745	462	3.250	33.24	22	.047	.001	.059	.001	.929	-.021
M8: FVIM	1346.325	466	3.226	1.58	4	.047	.000	.060	.001	.929	.000

Note: Statistically significant $\Delta\chi^2$ at the .050 p level were marked with an *. RMSEA = Root Mean Square Error Approximation; SRMR = Standardized Root Mean Square Residual; CFI = Comparative Fit Index; CIM = Configural Invariance Model; MIM = Metric Invariance Model; SIM = Scalar Invariance Model; PSIM = Partial Scalar Invariance Model; RIM = Residual Invariance Model; FVIM = Factor Variance Invariance Model.

hypothesis of "there being no significant differences across cultural groups" could be rejected. According to modification index, the other two items that seemed to be non-invariant were items 17 and 25. I first preferred to relax the constraints on both items 17 and75, thereby observing a significant, albeit still insufficient, decrement in the $\Delta\chi^2$ for the Model 5 = $\text{PSIM}_{(\text{i75 \& i17})}$. The factor loadings and item intercepts for items 75, 17, and 25 were then relaxed, after which the Model 6 = $\text{PSIM}_{(\text{i75 \& i17 \& i25})}$ revealed a good model fit with a non-significant $\Delta\chi^2$ statistic when compared to Model 2 = MIM (see Table 5). This indicated that partial scalar invariance was achieved across cultural groups when those specified three non-invariant items were relaxed. Even though achieving metric and scalar invariance was considered sufficient for supporting the cross-cultural comparability of scores for inferential statistics (Milfont & Fisher, 2010; Tucker et al., 2006), I also tested for more restrictive invariance models. As can be seen from Table 5, I constrained error variances in order to produce a residual invariance model (RIM) and constrained factor variances in order to produce a factor variance invariance model (FVIM), respectively in addition to factor loadings and item intercepts. Model 7 = RIM revealed statistically a non-significant $\Delta\chi^2$ in comparison to a less restrictive partial invariance model, the Model 6 = $\text{PSIM}_{(\text{i75 \& i17 \& i25})}$ with accepted $\Delta\text{RMSEA} \leq .015$, $\Delta\text{SRMR} \leq .01$, and $\Delta\text{CFI} \leq .01$. As expected, Model 8 = FVIM also resulted in evidence for factorial invariance across cultural groups when compared to Model 7 = RIM, thus indicating that the range of scores on the latent factors do not vary across cultural groups.

Discussion and Conclusion

Given the fact that a sound English language instrument from the United States, the CCTDI, was both identified and purported to be available in several non-English versions, it was deemed logical that the CCTDI should be adapted and subjected to the translation and cross-cultural validation process outlined in this dissertation. The first remark regarding the findings of this study is that the translation-back translation process yielded a linguistically equivalent Turkish version of the CCTDI. Yet, some intriguing points remained. For instance, it is detected that when the items possessed semantic problems, they did not possess conceptual or normative problems. If they displayed conceptual or normative problems, then the source of the problem was not semantic. The existing literature, however, does not specify any explanation regarding this finding. With this being the case, the significance of such a difference should be sought since the difference may provide international test translators with general information concerning the translation-back translation process as well as recommend techniques on how to maintain equivalency when subjecting works to a translation-back translation process.

The findings of this study supported that obtaining an identical target language version of the instrument in terms of linguistic equivalency with high values of Cronbach's alpha and CVIs for the sub-scales of the CCTDI did not necessarily entail that the target language version possessed a good factorial validity and/or measurement invariance across cultures. Thus, evidence retrieved from the results of the initial run consisting of 75 items using a 7-factor model to gauge factorial validity produced a poor model fit for both cultural groups with items from each sub-scale with low parameter estimates. A similar instrument adaptation study carried out by Herrera et al. (1993) indicated similar supporting results, which in turn highlighted the necessity to conduct two separate analyses for the cross-validation and measurement invariance across the Turkish and American populations. Interestingly, those items manifested themselves during the interactive response stage of the translation back-translation process and were considered to be emic-unique to the source culture. Chomsky (2011) is a unique author who mentioned the necessity of studying emic-etic situations for cross-cultural research. For that matter, based on the modification indices, those items were considered to be non-invariant and removed from the study so as not to be further analyzed. Removing these items from the model displayed significant, albeit insufficient, improvement in the model fit for cultural groups. By looking back to the modification indices, it was suggested to correlate the three residuals with their pairs, and this time the modified 22 item four-factor model revealed an acceptable model fit for the American sample and a good fit for the Turkish sample. Here, it is necessary to indicate that the modified four-factor structure enhanced the model fit without distorting the structure for the remaining four-factor since the alpha coefficients and the number of items allocated for each latent factor remained sufficient according to the norms specified by the literature (Cheung & Rensvold, 2000; DiStefano & Hess, 2005). In other words, it was made evident that the remaining items continued to measure what they were supposed to measure.

The other intriguing finding was that the model fit statistics showed a better fit for the Turkish target sample instead of the American source sample. Usually, the source language version is expected to show a better fit to the observed data (Dimitrov, 2010). There are several reasons which might explain this situation; one of them being that the English version of the instrument was developed by Facione (1990) and no subsequent update was considered for the English source version of the CCTDI. As Chomsky (2011) argues in *Hopes and Prospects*, language, like societies, has evolved as a result of the circumstances of the era. As social realities change as a result of globalization and other associated factors, culture also changes, thereby influencing language and perception (Chomsky, 2011). In other words, globalization influences language through the screen of society and culture. Therefore, the way people perceive the phenomenon of critical thinking is different today in comparison to their perceptions thereof two decades ago. The disposition aspect of critical thinking was influenced by the evolution in language, culture, and perceptions of people. With respect to this finding, the present research underlines the importance of considering linguistic theories when assessing equivalency between the linguistic versions of measurement instruments, suggesting further studies to modify the items to update the original version of the CCTDI by considering the perceptions of individuals living in the source culture and by performing a reassessment of the factorial validity with a similar study. Thus, one of the reasons explaining better fit with the Turkish sample can be considered to be the three-cycle multiple interactive translation procedure, which updated the items of the Turkish version of the instrument in terms of language use and proverb preference. The narrative and ultimate resolution of the proverb dissonance between language translations is just as much of a vivid example of the importance to ensure cultural validation as it is a crucial contribution to the growing literature on the international applicability of assessment instruments.

Following factorial validity check, the current research project examined the measurement invariance of the CCTDI across cultural groups to gauge whether researchers could proceed with a meaningful cross-cultural mean comparison. Each model was tested against more restrictive models whose results were derived from the comparison and which achieved a full metric invariance from both the statistical $\Delta\chi^2$ and the practical ΔRMSEA = .001, ΔSRMR = .001, and ΔCFI = .000 perspectives. However, the results did not support a full scalar invariance and instead showed that relaxing constraints placed on items 75, 17, and 25 exhibited partial scalar invariance. Regardless of the messages each of these three items intended to provide the test taker, the way the items were written might have caused potential non-invariance across the groups since they were constructed in the form of proverbs. It may be said that the use of proverbs in the construction of such psychological tests might be the cause of non-invariance across cultural groups.

The issue regarding the use of proverbs in psychological tests has two dimensions. The first dimension includes the argument of whether or not to use proverbs in psychological tests and the second one includes arguments about whether or not researchers should consider translating and using these proverbs for cross-cultural comparison of constructs (Behling & Law; 2000). A group of researchers have accumulated who hold that the use of proverbs may be allowed to a certain extent, specifically if the construct being measured is emic-unique to the source culture (Behling & Law; 2000; Berry, 1969; Yang, 1997). However, this is not valid for etic constructs since etic constructs posses the same components in terms of definition regardless of culture. When the content domains and definitions for latent factors of the four-factor model of the CCTDI were evaluated, it became necessary to state that the constructs being measured were etic. This may therefore be considered as one of the evidences explaining the potential cause of non-invariance across cultural groups. In other words, an etic construct that intends to measure the phenomenon under investigation with emic items cannot be merged successfully into the target language and culture for cross-cultural mean comparison. However, the current research does not provide strong evidence regarding as to why items constructed in the form of proverbs were found to be non-invariant. Therefore, this can also be considered as a concern of future research, which should be conducted to investigate the possible reasons behind such non-invariance. Future research is also needed to deepen the understanding of other possible causes of differences, since differences may be due to the administration of tests, translation errors, participants' perceptions of the items, culture specific emic constructs, and different conceptions of what critical thinking disposition consists of. The analysis of the current study's results continue to state that when additional constraints are employed to test more restrictive models, evidence for residual invariance and factorial invariance exists across cultural groups. Overall, support for partial scalar invariance indicated that latent means could be meaningfully compared across cultural groups without any measurement bias. However, the differences might be due to a reason still unknown to the researchers. A differentiated item functioning (DIF) study may also be suggested for further researchers to obtain deeper insight as to why several items functioned for Turkish culture in a different way (Ercikan, 2002). Specifically, the cultural conventions of Turkish society, its construct of critical thinking, and the items' relatedness to Turkish culture need to be considered together when studying DIF for those items.

A consideration of the results has produced the following suggestions for researchers who would like to use the CCTDI to measure one's dispositions toward critical thinking as well as those who would like to utilize cross-cultural mean comparison (Cheung & Rensvold, 1999; Milfont & Fisher, 2010). First, despite the fact that the initial 75 item 7-factor model of the CCTDI did not achieve a good fit with the observed data, researchers can use this model to collect data in order to assess the critical thinking dispositions of pre-service teachers in Turkish higher education institutions by relying on the high values of Cronbach's alpha for internal consistency reliability, the high values of test-retest reliability, and the high values of CVIs for evidence for content validity (Cheung & Rensvold, 1999). Second, researchers may omit the three specified latent factors that were considered to be non-invariant, delete the items with low parameter estimates, and use the modified 22 item four-factor model to assess the disposition dimension of critical thinking across Turkish and American higher education institutions. Third, researchers may either prefer to use the partial scalar invariance model or assume that the differences between cultural groups are not large enough to influence the results, and therefore use all the items to proceed with a cross-cultural mean comparison. Fourth, researchers may simply use the scales for within culture analysis or use the scales while avoiding cross-cultural mean comparisons.

In conclusion, assessing critical thinking dispositions has become a wider issue and no longer remains an issue of local concern. With an increasing interest in international research, researchers seek to find reliable and valid instruments to make cross-cultural comparisons. Using existing instruments to measure a construct in another cultural group requires performing a cross-cultural validation study rather than a simple translation procedure. In addition to this, this research project will lead to question the Arabic, Chinese (Mandarin), Dutch, Farsi, Finnish, French (Canadian), Hebrew, Italian, Japanese, Korean, Portuguese, Spanish (Mexico-Latin America), and Thai language versions of the CCTDI since no study has been performed examining the factorial validity or empirical evidence for the cross-cultural applicability of those language versions except for the Chinese (Mandarin) and Turkish language versions. This study also underlines an important fact: that although all the procedures may be employed to translate and back-translate psychological instruments for cross-cultural use, this does not ensure that the translated version will display factorial equivalency between the linguistic versions of the instruments. Therefore, researchers should provide strong evidence supporting whether the results regarding cross-cultural comparisons are due to real differences in the people and the variables being measured rather than due to errors in translation or any other reason. One of the contemporary ways of providing evidence regarding the issue is to make use of both Confirmatory Factor Analysis as well as a measurement invariance analysis across cultural groups.

References

Ægisdóttir, S., Gerstein, L. H., & Çinarbas, D. C. (2007). Methodological issues in cross-cultural counseling research: Equivalence, bias, and translations. *The Counseling Psychologist, 36*(2), 188-219.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). London: Prentice Hall.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. California: Sage Publications.

Benesch, S. (1993). Critical thinking: A learning process for democracy. *TESOL Quarterly, 27*(3), 545-548.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4*, 119-128.

Block, J.H. & Block, J. (1980). The role of ego-control and ego-resiliency in the organization of behavior. In W. A. Collins (Ed.), *Development of Cognition, Affect, and Social Relation: The Minnesota Symposia on Child Psychology*, Vol. 13. Hillsdale, NJ: Lawrence Earlbaum.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185-216

Chapman, D. W., & Carter, J. F. (1979). Translation procedures for the cross-cultural use of measurement instruments. *Educational Evaluation and Policy Analysis, 1*(3), 71-76.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14*, 464-504.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, 12, 471-492.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*(1), 1-27

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equation modeling. *Journal of Cross-Cultural Psychology, 31*, 188-213.

Chomsky, N. (2011). *Hopes and prospects*. Chicago: Haymarket Books.

Costa, P.T., Jr., & McCrae. R.R. (1985). *The NEO Personality Inventory Manual.* Odessa, FL: Psychological Assessment Resources.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5, 194-197.

Dewey, J. (1910). *How we think*. Boston: D. C. Heath.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2) 121–149

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment, 23*, 225-241.

Drasgow, F., & Probst, T. M. (2005). The psychometrics of adaptation: Evaluating measurement equivalence across languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). New Jersey: Lawrence Erlbaum Associates.

Ennis, R. H. (1993). Critical thinking assessment. *Theory into Practice, 32*(3), 179-186.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multi language assessments. *International Journal of Testing, 2*(3), 199-215.

Ernst, J. A., & Monroe, M. (2006). The effects of environment-based education on students' critical thinking skills and disposition toward critical thinking. *Environmental Education Research, 12*, 429-443.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Millbrae, CA: California Academic Press.

Facione, P. A., & Facione, N. C. (1992). *California critical thinking disposition inventory test manual*. Millbrae, CA: California Academic Press.

Facione, P. A., & Facione, N. C. (1992). *The California critical thinking disposition inventory*. Millbrae, CA: California Academic Press.

Facione, P. A., Facione, N. C., & Giancarlo, C. A. (1997). The motivation to think in working and learning. In E. Jones (Ed.), *Preparing competent college graduates: Setting new and higher expectations for student learning* (pp. 67-79). San Francisco, CA: Jossey-Bass Publishers.

Facione, P. A., Facione, N. C., & Giancarlo, C. A. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic, 20*(1), 61-84

Facione, P. A., Giancarlo, C. A., & Facione, N. C. (1995). The disposition toward critical thinking. *Journal of General Education, 44*(1), 1-25.

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research* (6th ed.). New York: McGraw-Hill.

Freire, P. (1974). *Education for critical consciousness*. New York: Continuum.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Boston: Allyn & Bacon.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's Alpha reliability coefficients for likert-type scales. Paper presented at Midwest Research to Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, October 8-10, 2003.

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health, 20*, 269-274.

Grosser, M. M., & Lombard, B. J. J. (2008). The relationship between culture and the development of critical thinking abilities of prospective teachers. *Teaching and Teacher Education, 20*, 1364-1375

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). New Jersey: Lawrence Erlbaum Associates.

Herrera, R. S., DelCampo, R. L., & Ames, M. H. (1993). A serial approach for translating family science instrumentation. *Family Relations, 42*(3), 357-360.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-15.

Johnson, L. E., & Reiman, A. J. (2007). Beginning teacher disposition: Examining the moral/ethical domain. *Teaching and Teacher Education, 23,* 676-687.

Kline, R. B. (2005). Principles and practices of structural equation modeling. New York: The Guilford Press.

Lee, H. J. (2005). Understanding and assessing preservice teachers' reflective thinking. *Teaching and Teacher Education, 21,* 699-715.

Lewin, K. (1935). *A Dynamic Theory of Personality: Selected Papers*. New York: McGraw-Hill.

McBridge, R. E., Xiang, P., & Wittenburg, D. (2002). Dispositions toward critical thinking: The preservice teachers' perspective. *Teachers and Teaching: Theory and Practice, 8*(1), 29-40

Melnick, S. L., & Zeichner, K. M. (1998). Teacher education's responsibility to address diversity issues: Enhancing institutional capacity. *Theory into Practice, 37*(2), 88-95

Metzler, M. W., & Blankenship, B. T. (2008). Taking the next step: Connecting teacher education, research on teaching, and programme assessment. *Teaching and Teacher Education, 24,* 1098-1111

Milfont, T. L., & Fisher, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-121.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing and Health, 29*, 489-497

Popper, K. (1935). *The logic of scientific discovery* (J. Freed & L. Freed, Trans.). London: Routledge.

Sánchez, C. A. (1993). An exploration of cognitive strategies and dispositions in relation to ego resiliency. Unpublished manuscript, University of California, Riverside.

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347-363.

Sass, D. A., & Schmitt, T. A. (2010). A Comparative investigation of rotation criteria within exploratory factor analysis. Multivariate Behavioral Research, 45, 1-33.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323-337

Sekaran, U. (1983). Methodological and theoretical issues and advancements in cross-cultural research. *Journal of International Business Studies, 14*(2), 61-73.

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating guidelines for test adaptations: A methodological analysis of translation quality. *Journal of Cross-Cultural Psychology, 37*(5), 557-567.

Stansfield, C. W. (2003). Test translation and adaptation in public education in the USA. *Language Testing, 20*(2), 189-207.

Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development, 37*(4), 240-257.

Tucker, K. L., Ozer, D. J., Lyubomirsky, S., & Boehm, J. K. (2006). Testing for measurement invariance in the satisfaction with life scale: A comparison of Russians and North Americans. *Social Indicators Research, 78*, 341-360.

U.S. Department of Education, National Center for Education Statistics. (2000). *The NPEC sourcebook on assessment, volume 1: Definitions and assessment methods for critical thinking, problem solving, and writing* (NCES Publication No. 2000-172). Washington, DC: U.S. Government Printing Office.

Vandenberg, R. J., & Lance, C. E. (2000). A review of synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Waltz, C.F., Strickland, O., & Lenz, E. (1991). *Measurement in nursing research* (2nd ed.). Philadelphia: F.A. Davis.

Weeks, A., Swerissen, H., & Belfrage, J. (2007). Issues, challenges, and solutions in translating study instruments. *Evaluation Review, 31*(2), 153-165

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research, 25*(5), 508-518

Yaghmaie, F. (2003). Content validity and its estimation. *Journal of Medical Education, 3*(1), 25-27

Yang, K. S. (1997). Indigenizing westernized Chinese psychology. In M. H. Bond (Ed.), *Working at the interface of cultures: Eighteen lives in social science* (pp. 62-76). London: Routledge.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? Journal of Psychoeducational Assessment, 29, 377-392.

Yeh, M. L. (2002). Assessing the reliability and validity of the Chinese version of the California Critical Thinking Disposition Inventory. *International Journal of Nursing Studies, 39,* 123-132.