Thinking Skills and Creativity 17 (2015) 33–44

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: http://www.elsevier.com/locate/tsc

The Halpern Critical Thinking Assessment: Toward a Dutch appraisal of critical thinking

Hannie de Bie*, Pascal Wilhelm*, Hans van der Meij*

Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 18 July 2014
Received in revised form 27 February 2015
Accepted 13 April 2015
Available online 24 April 2015

Keywords:
Halpern Critical Thinking Assessment
Validity
Reliability

ABSTRACT

Critical thinking is a vital component of 21st century skills. To assess this skill, a valid and reliable instrument is needed. This study focuses on the psychometric properties of the Dutch version of the Halpern Critical Thinking Assessment (HCTA). The HCTA was administered to university students in communications and psychology ($N = 240$), together with a Real-World Outcomes inventory (RWO-NL) used to measure negative life events. The number of negative life events was hypothesized to be inversely related to critical thinking ability. Reliability of the HCTA appeared adequate ($\alpha = .75$; $\lambda_2 = .77$), and factor analysis indicated that use of the constructed-response and forced-choice formats, with the five critical thinking subscales emerging for each, is an appropriate method for assessing critical thinking ability. Total HCTA and RWO-NL scores were not significantly related, $r = -.11$, *ns*. Recommendations for improving the Dutch HCTA are discussed.

1. Introduction

There are many examples of people being taken in by superstition and commercialism. One example is the sale of so-called Power Balance bracelets in the Netherlands during the 2010 World Cup. Well-known football players such as Wesley Sneijder and Arjen Robben swore to their effectiveness, and said that the special balance bracelets would increase strength, balance, and agility. The bracelets sold for €30 and, according to the importer, thousands of bracelets were sold every week in the months before the World Cup. The VU University in Amsterdam investigated the effect of wearing the bracelets and, not surprisingly, found that wearing a balance bracelet had no effect on balance, strength, and agility; there was not even a placebo effect (Vrije Universiteit Amsterdam, 2010, September 13). How is it possible that so many people fell for this hype? It is a clear example of poor critical thinking.

Critical thinking is becoming increasingly important, together with other '21st century skills' on which educators are starting to focus (Ananiadou & Claro, 2009). Other 21st century skills include communications, ICT literacy, social and/or cultural skills, creativity, collaboration and problem-solving skills (Voogt & Roblin, 2010). While these skills are not new (Rotherham & Willingham, 2010), becoming successful in our current society is considered to depend strongly on possession of them, meaning that students should be taught these skills starting at an early age (Voogt & Roblin, 2010, 2012).

Despite the existence of many conflicting definitions of critical thinking (e.g. Black, 2012; Ennis, 1996; Facione, 1998; Halpern, 1998; Moseley et al., 2005; Sternberg, 1986), Butler (2012) concludes that most researchers agree that critical

* Corresponding authors at: University of Twente, Faculty of Behavioural, Management and Social sciences, De Zul 10, 7522NJ Enschede, The Netherlands. Tel.: +31 53 489 3562.
E-mail addresses: h.debie@alumnus.utwente.nl (H. de Bie), p.wilhelm@utwente.nl (P. Wilhelm), h.vandermeij@utwente.nl (H. van der Meij).

http://dx.doi.org/10.1016/j.tsc.2015.04.001
1871-1871/

thinking "involves attempting to achieve a desired outcome by thinking rationally in a goal-oriented fashion" (p. 721). In line with that observation, this paper uses the definition offered by Halpern (1998):

> The term critical thinking refers to the use of those cognitive skills or strategies that increase the probability of a desirable outcome. [...] Critical thinking is purposeful, reasoned, and goal-directed. It is the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions. Critical thinkers use these skills appropriately, without prompting, and usually with conscious intent in a variety of settings (pp. 450–451).

If critical thinking does become (more important as) an educational objective, appropriate assessment of students' level of critical thinking is needed. This study focuses on the Halpern Critical Thinking Assessment (HCTA). The HCTA presents 25 everyday scenarios. These scenarios are derived from various real-life domains (e.g., health, education, and personal relationships) and address issues within these domains, such as the lottery, the death penalty, and slimming programs. The HCTA measures critical thinking skills for five skill types: (a) verbal reasoning (e.g., the ability to detect and defend against persuasive or deceptive language usage); (b) argument analysis (e.g., the ability to assess the strength of an argument); (c) thinking as hypothesis testing (e.g., the ability to reason scientifically, to determine whether or not the given information confirms a hypothesis); (d) using likelihood and uncertainty (e.g., using the correct estimate of a probability); and (e) decision making and problem solving (e.g., the ability to define a problem, identify goals and consider both positive and negative results) (Halpern, 2012).

For each scenario that is presented in the HCTA, participants are first asked to respond to an open-ended question, and then to answer from one to ten forced-choice questions. The following situation is an example addressing the skill of using likelihood and uncertainty: "Svetlana always plays the lottery. She chooses numbers that seem random, because she thinks she is more likely to win with random numbers than with numbers that are all even (e.g., 10, 9, 12) or numbers that are in ascending order (e.g., 7, 8, 9)." The respondent is asked to indicate whether Svetlana will have a better chance of winning by choosing numbers that seem random. As part of the open-ended constructed-response format, the respondent is asked to reason about the answer to this question. The respondent continues to the next page with forced-choice questions and is asked to select the statements that are true about the chance of winning with a particular set of numbers: (a) "If Svetlana truly believes that the numbers she chooses bring good fortune, she can increase her chance of winning", (b) "If the lottery is fair, than a random selection of six numbers has a better chance of winning than an ordered combination of numbers", (c) "You have as much chance of winning with random numbers as with an ordered combination", (d) "If Svetlana chooses the same random number combination every week, it will increase her chances of winning", (e) "If Svetlana chooses numbers that were not profitable last week, it will increase her chances of winning for next week". The answers for both question formats should show that the respondent recognizes that all combinations of numbers have an equal chance (correct answer: c).

As can be seen in the above example, the respondent's evaluation of the everyday scenario starts with a open-ended question format, and is then followed by forced-choice questions. The presence of the two formats in one assessment makes the HCTA unique. Other unique features of the HCTA are the use of a computerized grading system and the possibility of both offline and online administration. The next section discusses these features and other characteristics of the HCTA in more detail.

The use of the two question types in combination distinguishes the HCTA from other instruments to measure critical thinking skills (Halpern, 2012). Critical thinking assessments that rely solely on a constructed-response format include the Ennis-Weir Critical Thinking Essay Test (Ennis & Weir, 1985) and the ICAT Critical Thinking Essay Examination (Sonoma State University, 1996). Critical thinking assessments that rely solely on a forced-choice format include the California Critical Thinking Skills Test (Facione, 1990) and the Cornell Critical Thinking Test (Ennis, Millman, & Tomko, 1985).

The open-ended questions involve free recall. Respondents must search their memory for the correct response and construct an answer (Bridgeman & Morgan, 1996; Butler, 2012; Ku, 2009). Giving a constructed response takes more cognitive effort, but also reveals more of the respondents' disposition to engage in critical thinking. The open-ended format is more likely to reveal the motivational and intentional aspects of the respondents' critical thinking because this format shows the extent to which respondents are willing and able to engage into critical thinking at the right moment (Ku, 2009). Drawbacks of this format are that it is time-consuming to administer and that concerns about the subjectivity of scoring can be raised.

Forced-choice questions require less cognitive effort, because respondents can rely on recognition. They merely have to identify the applicable response from a list of alternatives (Bridgeman & Morgan, 1996; Butler, 2012; Ku, 2009). Forced-choice questions are efficient for large-scale assessment of critical thinking, but this type of assessment only reveals cognition. The forced-choice format is not very suitable as a performance-based assessment for measuring respondents' disposition toward critical thinking (Ku, 2009).

Ku (2009) advocates the use of a multi-response format for assessment of critical thinking, and emphasizes that a critical thinking assessment should address two components: cognition (e.g., a set of skills, rules of formal logic), and disposition (e.g., the intention and motivation to engage in critical thinking). Together, these components determine actual critical thinking performance. Taube (1997) empirically showed that a two-factor model of critical thinking provided a better fit with the data than a one-factor model. One factor consisted of dispositional indicators of critical thinking, including assessments of

tolerance of ambiguity, need for cognition, and intellectual development. The other factor consisted of cognitive indicators of critical thinking, including Scholastic Aptitude Test (SAT)-scores, Grade Point Average (GPA)-scores, and scores on the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980). In short, there appear to be distinct dispositional and cognitive components of critical thinking, which also appear to be associated with different response formats, such that constructed responses reveal more about the former than the forced-choice format alone can do (Ku, 2009).

Halpern (2012) evaluated the factor structure of the HCTA in more detail, with a U.S. normative sample of 450 respondents. The best result for these data was a two-factor model (differentiating between the constructed-response and forced-choice formats), with each factor including the five categories of critical thinking skills. Hau et al. (2006, April) and Ku (2009) also evaluated the factor structure of the HCTA using a sample of respondents from the U.S. and Hong Kong, and they arrived at the same conclusion. These empirical studies thus support the use of both question types for assessing critical thinking (Halpern, 2012; Hau et al., 2006, April; Ku, 2009).

There is considerable evidence of the reliability and validity of the HCTA (Halpern, 2012). Internal consistency (Cronbach's alpha) scores and inter-rater reliability findings range from .85 to .97. Content validity is presumably high because the items are based on the constructs that are most frequently mentioned in descriptions of critical thinking. Three small-scale studies mentioned in the manual for the HCTA reported moderate correlations between scores on the constructed-response and forced-choice items of .39, .49, and .51. A fourth small study was conducted in Belgium by Verburgh, Francois, Elen, and Janssen (2013), where they translated the HCTA to Dutch and investigated the psychometric properties of that version. They found a medium to strong correlation between scores on the constructed-response and the forced-choice items. These studies suggest that there is a reasonable relationship between formats, and that the dispositional and the cognitive component of critical thinking are addressed by distinct measures (Halpern, 2012).

Positive correlations (r = .12–.59) with level of education and scores on academic ability tests such as the SAT and GPA point to construct validity (Butler, 2012; Halpern, 2012; Hau et al., 2006, April; Ku, 2009). In addition, scores on need for cognition and conscientiousness scales are moderately related to critical thinking skills (Clifford, Boufal, & Kurtz, 2004; Halpern, 2012; Ku, 2009; Ku & Ho, 2010; Spector, Schneider, Vance, & Hezlett, 2000).

The HCTA has been translated into a number of different languages, including Chinese, Portuguese, Dutch, Spanish, and Vietnamese (Butler et al., 2012; Halpern, 2012). With respect to total HCTA score, studies in Spain (M [355] = 106, SD = 16.1) and Belgium (M [173] = 113.15, SD = 11.5) found similar results compared to the U.S. normative sample (M [450] = 109.71, SD = 18.23). The relatively consistent scores illustrate the quality of the HCTA-scenarios. Translators needed few changes to make the assessment culturally accessible, possibly because cultural differences were already taken into account in the development of the HCTA-scenarios (Halpern, 2012).

Recently, Butler (2012) conducted an interesting study on the external validity of the HCTA. The study examined the relationships between real-world outcomes and level of critical thinking. Real-world outcomes were measured by Butler (2012) with an adapted version of the Decision Outcomes Inventory (DOI) developed by (Bruine de Bruin, Parker, & Fischhoff, 2007). This adapted version, the Real World Outcomes inventory (RWO), assesses frequencies of negative life events over the past six months in domains such as finances, education, and personal relationships. Butler (2012) reasoned that higher scores on the HCTA would be associated with fewer negative life events. This argument was based on the idea that more skilled critical thinkers use their skills in various domains of life in order to be successful and avoid failure caused by poor decision making. As Halpern (1998) states: "critical thinkers will have more desirable outcomes than 'noncritical' thinkers (where 'desirable' is defined by the individual, such as making good career choices or wise financial investments)" (p. 450). Butler's prediction was confirmed. The HCTA and RWO scores of 131 respondents showed a modest yet statistically significant relationship (r = –.38, p = <.001) in the predicted direction. Dwyer, Hogan, and Stewart (2012) replicated this study with an Irish sample of 70 respondents. Their study confirmed Butler's results, as a moderate but significant correlation between critical thinking scores and real-world outcomes was found, r = –.28, p = .019.

Besides the work by Butler (2012) and Dwyer et al. (2012), there are no studies that provide validity evidence linking critical thinking assessments with real world outcomes outside of the academic world and business. But there are other indications that critical thinking is negatively associated with negative life events. For example, scores on a random events knowledge test (which resembles the skills of using likelihood and uncertainty and decision making and problem solving) were negatively correlated with a measure that indicates a pathological gambling problem (Turner, Macdonald, & Somerset, 2008). So there is a sign that there is a relationship between critical thinking and real world outcomes, but more research is needed here.

In the present study, the psychometric qualities of the Dutch version of the HCTA were examined. Schuhfried GmbH has already translated the HCTA to Dutch. However, there is only one study mentioned in the HCTA manual that provided psychometric data for the Dutch version (Verburgh et al., 2013). Therefore, this study explored the internal structure and reliability of the Dutch HCTA. In addition, this study examined the external validity of the HCTA by relating it to a Dutch version of the Real-World Outcomes inventory (RWO-NL).

As in the work by Halpern (2012), Hau et al. (2006, April), and Ku (2009), the factor analysis is expected to show two related, but separable latent factors for the constructed and forced-choice responses on the HCTA. Additionally, the Cronbach's alpha score for the total item set is expected to be 0.85 or higher, and reliabilities of the two subscales should resemble those found with the U.S. normative sample (constructed-response format α = .84, and forced-choice format α = .79). Finally, as Butler (2012) found, it is predicted that there will be a modest, but significant negative correlation between HCTA scores and scores on a real world outcomes inventory.

2. Method

2.1. Participants

The respondents were first- and second-year communications and psychology students from the University of Twente in the Netherlands, Enschede. A total of 258 students signed up voluntarily and earned 1.5 study credit points by completing both the Dutch HCTA and RWO-NL. Data from 18 respondents were excluded from this set because their answers on one or both tests indicated that they did not take the assessment seriously (e.g., giving funny comments for constructed-response questions with no attempt to give the right answer, or only responding with a dot or slash for all questions) or they prematurely aborted the session. Of the remaining 240 respondents, 80% were female ($n = 191$) and 20% were male ($n = 49$). This female/male distribution represents the distribution in the student population at large for the included programs of study. The respondents were aged 18–32 ($M = 20.53$, $SD = 2.07$). A large majority of students came from the Netherlands (46%) or from Germany (34%). German students are required to pass the Dutch language test in order to study in the Netherlands. There were missing data for the remainder of the students on country of origin.

2.2. Materials

2.2.1. Halpern Critical Thinking Assessment (HCTA)

The Halpern Critical Thinking Assessment (HCTA) is available in two forms (S1 and S2) and two versions (A and B). Form S1 presents 25 everyday scenario's accompanied by questions in two response formats: first constructed responses (open-ended) and then forced choices (e.g., multiple-choice, rating of alternatives or ranking). Form S2 consists of only the forced-choice questions, and can be used as a short form. Version A and version B are parallel versions of the HCTA. The present study used form S1 and version A. Administration of the HCTA (Form S1, Version A) takes 45–80 min. The maximum score for the HCTA is 194.

The HCTA draws its everyday scenarios from disciplines such as medical research, social policy analysis and other disciplines respondents may encounter in daily life. The HCTA measures five categories of critical thinking skills, with each category making a different contribution to the total score: (a) verbal reasoning (12%); (b) argument analysis (21%); (c) hypothesis testing (24%); (d) likelihood and uncertainty (12%); and (e) decision making and problem solving (31%). Five everyday scenarios are presented for each construct.

The original HCTA includes the Vienna Test System (VTS), but this study administered the HCTA and RWO-NL online with Thesistools.[1] Scoring followed the HCTA manual: scores for forced-choice responses were automatically computed. Scoring of the constructed responses was guided by VTS. Guided grading uses computerized prompting of the grader. For example, for the constructed-response question regarding the scenario about of Svetlana (mentioned in the introduction of this article), the grading system displays the following question: "Did the respondent indicate that all combinations of numbers have an equal chance?" For each question, the grader determines whether the respondent's answer: (a) clearly indicates this; (b) less clearly indicates this; or (c) does not indicate this at all (Halpern, 2012). Standardized scoring, in combination with computerized prompting, reduces scoring bias. Halpern (2012) reported a high ($r = .83$) inter-rater reliability for constructed responses, which indicates that objective scoring can be assumed.

2.2.2. Real-World Outcomes inventory (RWO-NL)

The Dutch version of the Real-World Outcomes inventory (RWO-NL) was adapted from Bruine de Bruin et al. (2007) and Butler (2012). The original items were first translated to Dutch and then adaptations were made for language-dependent expressions, involving usage of culturally unfamiliar or uncommon terms or items (Butler et al., 2012). In the RWO-NL, respondents must indicate whether or not they have experienced a particular event in the past six months by selecting a check box for 'yes' or 'no'. The inventory contains items from a wide variety of domains, such as finances, education, and personal relationships. There are items with and without sub-questions. The former always start with a situation that makes negative events possible (e.g., "Gone shopping for food or groceries"). The negative events are assessed by the sub-questions (e.g., "Threw out food or groceries you had bought, because they went bad"). Therefore this inventory considers the possibility of actually experiencing a negative life event due to a previous decision. Nine items have no sub-questions (e.g., "Been in a public fight or screaming argument").

The total set of items was created in three steps. First, 33 items were taken from the original inventory of Bruine de Bruin et al. (2007). Second, 10 additional items were taken from the inventory of Butler (2012). Of these original items, 12 were adapted to make them suitable for the Dutch population. For example, an original item contained the words "Used checks". This was altered to "Used a debit card", because checks are rarely used in the Netherlands. And third, an additional set of seven items was included to represent events that play a great role nowadays in the life of Dutch students. For example: (a) "Had a mobile phone", with subsequent negative events: (b) "Lost a mobile phone" and (c) "Had to pay at least three times extra on your phone bills because you went over your call/text/data limit". In all, the final version of the RWO-NL consisted

[1] http://www.thesistools.com.

Table 1
Summary of reliability estimates (Cronbach's alpha and Guttman's lambda) for (sub)scale scores for the HCTA.

(Sub)scale	Cronbach's alpha			Guttman's lambda		
	Constructed response	Forced choice	Total	Constructed response	Forced choice	Total
Critical thinking	.61	.64	.75	.63	.67	.77
Verbal reasoning	.30	.19	.33	.33	.23	.37
Argument analysis	.24	.22	.38	.30	.26	.42
Hypothesis testing	.29	.38	.53	.32	.42	.55
Likelihood and uncertainty	.31	.18	.39	.33	.20	.42
Decision making and problem solving	.30	.43	.52	.34	.47	.54

of 50 items. See Appendix A for the complete list of the 50 presented item sets and their corresponding response frequencies. The administration of the RWO-NL takes 5–15 min.

The scoring of the RWO-NL was adapted from Bruine de Bruin et al. (2007). The negative outcomes are weighted to make the calculation of the total score on the RWO-NL more fair. That is, some outcomes are not as bad as others (e.g., taking the wrong train or bus versus going to jail). Generally, truly bad outcomes probably happen to only a few people. To weight for severity, the proportion of avoided outcomes per item is computed (1 – proportion of experienced outcomes). Each respondents' weighted RWO-NL score is the sum of the weighted scores divided by the total number of experienced opportunities for making bad decisions. This results in a score between 0 and 1, where 0 stands for good decision making, and 1 for poor decision making. Respondents can skip uncomfortable questions by checking a third option. If the nonresponse concerned an opportunity that could make a negative life event possible but the respondent did answer "yes" to a subsequent negative life event, then the response does not make sense. The opportunity, for instance driving a car, is necessary in order to experience the related negative life event, such as getting a speeding ticket. In these cases, it is assumed that the respondent forgot to answer the first question and therefore the nonresponse is changed to "yes". If the nonresponse concerned a negative life event, then the whole item is excluded from analyses by excluding the opportunity that could make the negative life event possible from the calculation of the proportion score.

2.3. Procedure

First, all respondents signed an informed consent statement describing the purpose of the study and stating confidentiality. Next, each respondent engaged in the same online assessment procedure; starting with the HCTA, followed by the RWO-NL. After completion of the two measures, respondents were asked whether they had experienced problems in answering the questions on the HCTA and the RWO-NL. After the entire study was completed, the respondents received a debriefing via email. In addition, references to papers about critical thinking were included for interested students.

3. Results

The mean score on the HCTA ($n = 240$) was 108.23 ($SD = 13.91$) out of the maximum score of 194. All measures met the criteria for univariate normality (skewness and kurtosis between −1 and 1). There were no differences in HCTA scores based on age or gender (all $ps > 05$). There were also no differences observed in HCTA scores between Dutch and German students, $t(189) = 1.52$, $p = .13$, and between Dutch students and the remaining students with other ethnicities, $t(115) = 0.58$, $p = .57$. It was therefore concluded that students with a native language other than Dutch were presumably able to understand the content of the items correctly.

Cronbach's alpha for all items was $\alpha = .75$, which seems to support the reliability of the HCTA (Kline, 2000). However, the large number of items and the use of multiple subscales within the HCTA could have artificially inflated the value of Cronbach's alpha (Cortina, 1993; Tavakol & Dennick, 2011). Therefore, Cronbach's alpha was calculated for every skills category under both formats (see Table 1).

As can be seen, total subscale scores had moderate to low values for Cronbach's alpha. However, the use of other reliability estimations such as Guttman's lambda is endorsed by Sijtsma (2009). This analysis results in a slightly better overall reliability value, $\lambda_2 = .77$, and subscale values.

A confirmatory factor analysis was carried out with three measurement models specified in Halpern (2012), to test the hypothesis that the factor analysis would show two related, but separable latent factors for the constructed and forced-choice formats, each including all five critical thinking categories. This analysis evaluated the goodness of fit of each model to the data. The first model (M1) represented two latent factors, the constructed-response ("critical thinking – free recall") and the forced-choice format ("critical thinking – recognition"), each including the five sub-scale scores hypothesized to load on the associated factor. The correlated unique errors of the associated sub-scale scores were allowed to be freely estimated, since items with both formats have the same origin. The latent factors were allowed to be correlated to test the hypothesis that there was a relationship between the two formats, and simultaneously give evidence regarding the separability of free recall and recognition in this context. The only change in the second (M2) and third (M3) models compared to M1 was that

Table 2
Model fit statistics for the two-dimensional measurement models using the HCTA sample.

Model	χ^2	*df*	*p*	CFI	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI
M1	32.050	29	.318	.990	.021	–	–	–	–
M2	42.917	30	.060	.957	.042	10.867	1	.001	0.033
M3	95.614	30	<.001	.782	.096	63.564	1	<.001	0.208

Note: CFI, comparative fit index (Bentler, 1990); RMSEA, root mean square error of approximation (Browne and Cudeck, 1993).

the standardized latent correlation of the two latent factors was set to 1 and 0, respectively, to test the hypotheses that the two factors have indistinguishable or completely separate characteristics.

The calculations of the factor structure of the HCTA were carried out with IBM® SPSS® Amos (TM) 22 (Arbuckle, 2013). Maximum likelihood was used to estimate the model parameters, since the data were normally distributed. The following cut-off values were used to evaluate the goodness of fit of the models: non-significant χ^2-test, CFI ≥ .95, and RMSEA < .05 (Jackson, Gillaspy, & Purc-Stephenson, 2009; Marsh, Hau, & Wen, 2004). To test whether model M2 and model M3 differed significantly from model M1, criteria of a significant $\Delta\chi^2$ statistic and ΔCFI ≥ 0.01 (Cheung & Rensvold, 2002) were used. The model fit statistics for the three models are summarized in Table 2.

The statistics in Table 2 show a significantly better fit of model M1, $\Delta\chi^2$ (1, N = 240) = 10.87, p < .01, which indicated that the latent factors of "Critical Thinking – free recall" and "Critical Thinking – recognition" had a strong relationship (r = .785). However, the factors were separable, because M2 (in which the standard latent correlation of the two factors was set to 1) yielded a poorer fit than M1.

Hancock and Mueller (2001) argue that coefficient *H* is a more appropriate way of reporting reliability as compared to Cronbach's alpha when performing a confirmatory factor analysis. This measure reflects "the extent to which the latent construct is reproducible from its own measured indicators" (Gagne & Hancock, 2006, p. 112). A minimum value of .70 for coefficient *H* is desirable in order to reach reasonable reliability. As can be seen in Fig. 1, the latent factors of the forced-choice and constructed-response format reached a value of .60 and .65 respectively, which represented insufficient reliability for this measure.

The standardized factor loadings of the constructed-response and multiple choice sub-scales on their associated latent factors were all significant (all *ps* < .001). The correlated unique errors between the associated sub-scales turned out to be rather small and only reached significance for "Hypothesis testing" and "Decision making and problem solving". The lack of significance indicates good fit of the model, because it is preferable that the correlations between the latent variables and the associated subscales explain all variance. The structural relations and standardized factor loadings are depicted in Fig. 1.

To provide further insights into the factor structure of the HCTA, we analyzed a bifactor model with both latent factors (the constructed-response and the forced-choice format) and a general factor of critical thinking loaded by the ten subscale

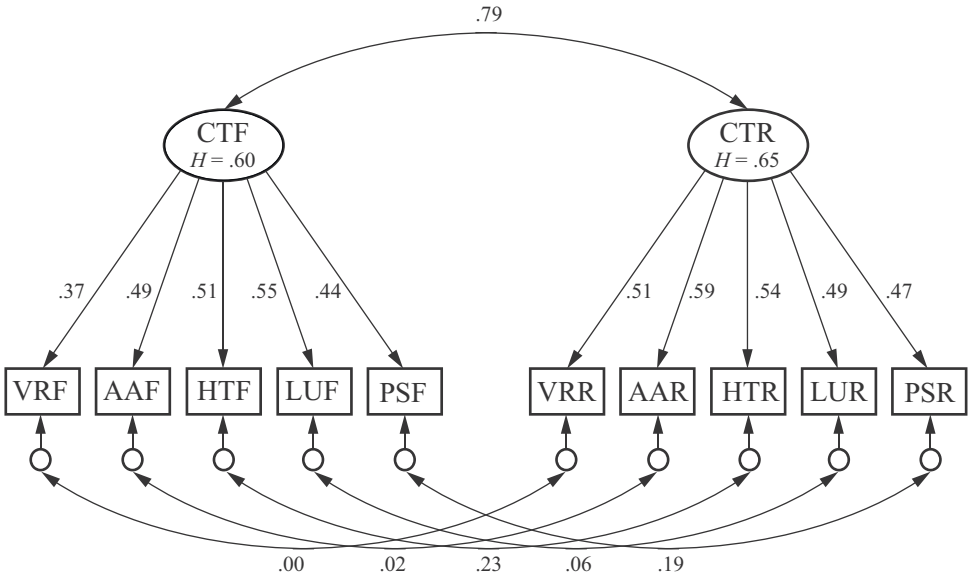

Fig. 1. Standardized factor loadings of measurement model 1. CTF, latent factor critical thinking – free recall; CTR, latent factor critical thinking – recognition; VRF, sub-scale score verbal reasoning – free recall; AAF, sub-scale score argument analysis – free recall; HTF, sub-scale score hypothesis testing – free recall; LUF, sub-scale score likelihood and uncertainty – free recall; PSF, sub-scale score decision making and problem solving – free recall, VRR, sub-scale score verbal reasoning – recognition; AAR, sub-scale score argument analysis – recognition; HTR, sub-scale score hypothesis testing – recognition; LUR, sub-scale score likelihood and uncertainty – recognition; PSR, sub-scale score decision making and problem solving – recognition. *H* = coefficient H (Hancock & Mueller, 2001), also referred to as maximal reliability.

Table 3
Standardized factor loadings of the bifactor model.

Subscale	Second-order factor	First-order factors	
	Critical thinking	Constructed-response	Forced-choice
Verbal reasoning (F)	.294[a]	.189	
Argument analysis (F)	.406[a]	.311[a]	
Hypothesis testing (F)	.613[a]	−.109	
Likelihood and uncertainty (F)	.459[a]	.458[a]	
Decision making and problem solving (F)	.425[a]	.114	
Verbal reasoning (R)	.374[a]		.603[a]
Argument analysis (R)	.481[a]		.268
Hypothesis testing (R)	.538[a]		.182
Likelihood and uncertainty (R)	.420[a]		.187
Decision making and problem solving (R)	.527[a]		−.014

Note: (F), subscale measured by the constructed-response format; (R), subscale measured by the forced-choice format.
[a] Standardized factor loading is significant at the 0.05 level (2-tailed).

scores. Model fit statistics were similar to that of M1, $\chi^2(25, N = 240) = 32.052$, $p = .157$; CFI = .977; RMSEA = .034, and no difference between M1 and the bifactor model was detected, $\Delta\chi^2(4, N = 240) = 0.002$, $p = .99$. This indicated that the bifactor model is also an adequate model to evaluate the factor structure of the HCTA. The bifactor model facilitates the interpretation by showing each factor's unique contribution to the subscales (Brown, 2006; Wolff & Preising, 2005). The factor loadings in Table 3 show that the impact of the second-order factor of critical thinking to the subscale scores is generally higher than the impact of the first-order factors. The impact of the forced-choice format only outweighs the impact of the general critical thinking factor on the subscale of verbal reasoning. In this case, 36% of its variance is explained by the first-order factor of the forced-choice format ($.603^2 = .364$), and only 14% of its variance is explained by the higher-order factor of critical thinking ($.374^2 = .140$). This indicates that a specific element of this subscale is only reflected in the forced-choice format and not captured in the critical thinking factor.

The mean score on the RWO-NL ($n = 238$) was 0.14 ($SD = 0.08$). Only two respondents did not complete the RWO-NL. Items to which participants were not comfortable with answering, were not included in these calculations. The RWO-NL scores met the criteria for univariate normality (skewness and kurtosis between −1 and 1). Correlational analysis was carried out to test the final hypothesis that HCTA and RWO-NL score would show a modest, but significant negative relationship. However, there was no significant relationship between total HCTA score and score on the RWO-NL, $r = -.11$, *ns*. When we investigated how both response formats on the HCTA are related to scores on the RWO-NL, the relationship between the forced-choice format of the HCTA and the RWO-NL turned out to be small but significant, $r(238) = -.14$, $p = .034$. When we investigated how the subscale scores on the HCTA are related to scores on the RWO-NL, the relationship between the subscale 'decision making and problem solving' and scores on the RWO-NL also turned out to be small but significant, $r(238) = -.14$, $p = .026$. An overview of all bivariate correlations between scores on the HCTA and RWO-NL is shown in Table 4.

Table 4
Summary of bivariate correlations of (sub)scale scores on the HCTA and the RWO-NL.

HCTA scale	RWO-NL
Critical thinking total score	−.11
Constructed response	−.06
Forced choice	−.14*
HCTA subscale	RWO-NL
Verbal reasoning total score	−.05
Constructed response	−.01
Forced choice	−.08
Argument analysis total score	−.08
Constructed response	.01
Forced choice	−.15*
Hypothesis testing total score	−.01
Constructed response	−.04
Forced choice	.02
Likelihood and uncertainty total score	−.08
Constructed response	−.06
Forced choice	−.08
Decision making and problem solving total score	−.14*
Constructed response	−.08
Forced choice	−.15*

* Correlation is significant at the 0.05 level (2-tailed).

4. Discussion and conclusions

The results of the present study partly replicated those reported by Halpern (2012). First, the confirmatory factor analysis revealed that the model that reflected two correlated latent factors (constructed-response and forced-choice formats) each containing the five subscales of the HCTA, best fitted the factorial structure of the Dutch HCTA. This agrees with the findings of Halpern (2012), Hau et al. (2006, April), and Ku (2009), in which the same model fitted the data from a U.S. and a Chinese sample. Most importantly, the analysis confirmed that the latent factors of the constructed-response and the forced-choice formats are closely related, yet separate in their properties. This separability reflects the difference between measuring more of the dispositional component with the constructed-response format, and measuring the cognitive component with the forced-choice format. This indicates that the use of both response formats in the HCTA is a valid method to obtain an accurate indication of critical thinking ability. Analysis of a bifactor model revealed a greater influence of the general critical thinking factor than both lower-order factors of the response formats. Caution is warranted when interpretating scores on these lower-order factors separately, because these factors do not have an unique influence on all related subscales. But generally, the calculation of a total score on the HCTA is justified. Second, the estimated value of Cronbach's alpha did not confirm the hypothesis of $\alpha \geq .85$; neither did the alphas of the subscale scores (constructed-response format $\alpha = .61$ in lieu of $\alpha = .84$, and forced-choice format $\alpha = .64$ in lieu of $\alpha = .79$). Also, coefficient H calculated from the factor analysis was not sufficient to achieve a reasonable reliability for the subscales. Still, the overall values of Cronbach's alpha ($\alpha = .75$) and Guttman's lambda ($\lambda_2 = .77$) indicate good reliability of the Dutch HCTA. Taken together, these results and the universality of the two-factor model confirm the quality of the Dutch translation for population level studies and justifies the use of the Dutch version of the HCTA for the Dutch population. But given the lower reliabilities of the subscale scores, caution is advised with regard to the individual level.

The third aim was to assess the relationship between HCTA and RWO-NL scores. We expected that the skill of critical thinking is negatively related to the frequency of negative life events. The analysis showed a non-significant relationship, while Butler (2012) and Dwyer et al. (2012) found a modest negative relationship, $r(131) = -.38$, $p < .001$ and $r(70) = -.28$, $p = .019$, respectively. When we investigated further how both formats and subscores for the HCTA were related to scores on the RWO-NL, it was revealed that the forced-choice format and the subscale 'decision making and problem solving' do have a significant relationship with scores on the RWO-NL. That the subscale of decision making and problem solving significantly correlates with the RWO-NL is plausible, because the RWO was originally adapted from the Decision Outcomes Inventory (DOI) of Bruine de Bruin et al. (2007). The DOI measures "decision-making success in terms of avoiding negative decision outcomes" (p. 943). However, the correlations found in this study are very small and should be interpreted with caution. In addition, the relation between scores on the HCTA and RWO-NL in this study could possibly be compromised because of the use of self-report. For example, students with higher critical thinking skills could give more socially desirable answers on the RWO-NL. The inclusion of only university students could also have influenced the RWO-NL scores, because it is possible that university students would be more likely to display themselves as good decision-makers compared to students from other educational levels. As in the study by Butler (2012), respondents were given the option not to answer any uncomfortable questions, to lessen this concern during administration of the RWO-NL. The absence of an adequate relationship could also be due to the limited range of level of educational attainment, but this assumption cannot be checked in this study, however. A closer look at item nonresponse revealed a significant negative relationship between HCTA scores and number of unanswered questions on the RWO-NL, $r(239) = -.20$, $p = .002$. This is the opposite of what would be expected according to the proposition that respondents with a higher critical thinking score gave more social desirable answers and therefore responded to fewer items on the RWO-NL. Perhaps respondents with a lower HCTA score were less motivated to complete the RWO-NL seriously, and therefore did not answer all questions. Or respondents with a higher HCTA score could have been more aware of the anonymity of the RWO-NL, and therefore could have answered more questions even when they felt embarrassed. Further development and validation of the RWO-NL can possibly yield a more reliable tool for measuring negative life events.

This study also has its limitations. First, the sample of respondents consisted only of university students in the same two fields of study who chose to participate in this study. There was no random selection, although the sample is rather large and is representative for students in these fields of study in terms of gender distribution. Second, in line with previous research, it is assumed that the constructed-response format measures more of the dispositional aspect of critical thinking. The argument that the constructed-response format provides the opportunity to come up with self-generated solutions, think critically in an unprompted context, and show the extent to which someone will think critically makes sense. The separability of both formats is also reflected in the internal structure of the HCTA. But external evidence supporting the assertion that the constructed-response format measures the dispositional aspect of critical thinking is scarce. Bridgeman and Morgan (1996) indicated that constructed-response and forced-choice questions measure separate cognitive abilities and Martinez (1999) added that the range of cognitions needed for answering constructed-response questions is larger than the range of cognitions for answering forced-choice questions. Further research should investigate this issue more precisely for critical thinking. Examining the relation of scores on the constructed-response format of the HCTA and RWO-NL scores could clarify this; respondents who score low on the RWO-NL willingly transfer and exercise their critical thinking skills in various domains in life, which should result in a high score on the constructed-response format. However, this relation was not seen in this sample (see Table 4). Other methods of measuring the disposition toward critical thinking in particular, such as the California Critical Thinking Disposition Inventory (CCTDI) (Facione, 2000), could provide further

evidence. Comparing scores on the CCTDI and scores on the constructed-response format of the HCTA could shed some light on whether the disposition to think critically is revealed in the constructed response. Third, the online administration of the test was unproctored. There are mixed opinions about this method of administration within the cognitive domain; that proctored and unproctored conditions may be equivalent (Lievens & Burke, 2011), or that the unproctored condition yields higher test scores than the proctored conditions because of the presence of a proctor (Carstairs & Myors, 2009). This effect did not occur in the present study, because no inflation of the unproctored test scores was observed compared with the mean of the U.S. normative sample ($M = 109.71$, $SD = 18.23$) (Halpern, 2012). Just the opposite, a lower mean of 108.23 ($SD = 13.91$) was found. However, the U.S. normative sample has a wider spread in terms of age ($M = 29$, $SD = 12.53$), which could mean that those respondents had more years of education. More years of education is related to level of critical thinking (Butler, 2012), which could explain the lower mean in comparison with the U.S. normative sample.

Respondents had the possibility of giving feedback on the HCTA after completion of the assessment. The option to give feedback was used by 55 respondents, 23% of the total sample. In their feedback, 37 respondents stated that complex and incorrect syntax and the use of scientific language made it difficult to comprehend the scenarios and questions on the HCTA. Based on this feedback, we infer that the issues raised could have affected the comprehensibility of some of the items. The lower mean in comparison to the U.S. norm sample could also be caused by the translation of the HCTA, where cultural differences could have influenced the test score. The Netherlands is a welfare state, where different opinions about politics, diets, drugs and alcohol could influence responses on items that are related to these issues. For example, respondents might react more leniently regarding the scenario about alcohol abuse, due to the lower age at which alcohol may be consumed in the Netherlands and greater social acceptance of alcohol. This milder attitude might lead Dutch respondents not to report the alcohol abuse to an authority figure as readily in that particular scenario, possibly lowering the score on this item. We suggest that these issues be addressed before further research is done.

This study established a proper foundation for future validation research on the HCTA in the Netherlands. The hypothesized factor structure was confirmed, which justifies the Dutch version of the HCTA. Because our sample of respondents showed a somewhat restricted range in age and educational level, it is recommended to include a larger and more diverse sample of respondents in future studies. In addition, the RWO-NL can be revised and improved, or another method for measuring negative life events can be introduced. For instance, an unobtrusive experiment could be developed to observe actual behavior that indicates good or bad decision making. Although this is a very laborious method of collecting data, it can circumvent the social desirability issue. Another interesting direction for further validation research for the Dutch population is investigating version B of the HCTA. A repeated measures design requires the use of a parallel version of the HCTA, allowing measurement of gains in critical thinking from pretest to posttest. With a sound Dutch critical thinking assessment instrument, it would be possible to determine the effectiveness of critical thinking instruction. Overall, a thoroughly researched HCTA offers a promising tool for assessing the 21st century skill of critical thinking among Dutch learners.

Acknowledgements

This research was conducted for the master thesis of Hannie de Bie, supervised by Pascal Wilhelm and Hans van der Meij. The authors would like to thank Diane F. Halpern and Heather A. Butler for their knowledge and helpful suggestions, and Schuhfried GmbH and Science Plus Group BV for their practical support.

Appendix A.

See Table A.1

Table A.1
Real world outcomes inventory (RWO-NL) with response frequencies.

Item		Percentage who made the decision	Percentage who experienced outcome
In the last six months, have you ever . . .			
1a	Borrowed a book at the library?	43.5	
b	Borrowed a book at the library without ever reading it?		45.2
c	Had to pay a fee because you returned the book too late?		24.0
2a	Bought new clothes or shoes?	96.7	
b	Bought new clothes or shoes you never wore?		31.2
3a	Gone shopping for food or groceries?	99.6	
b	Threw away food or groceries you had bought, because they went bad?		77.7
4a	Done your own laundry?	68.2	
b	Ruined your clothes because you didn't follow the laundry instructions on the label?		12.3
5a	Been enrolled in any kind of school?	100	
b	Missed a class because you forgot to set your alarm?		30.5
c	Pulled an 'all-nighter'?		33.5

Table A.1 (*Continued*)

Item		Percentage who made the decision	Percentage who experienced outcome
d	Forgotten to do a class assignment?		22.3
e	Arrived to class and realized that you'd forgotten about an exam that day?		3.9
f	Been suspended from school for at least one day for any reason?		24.9
6a	Been late for something, because you've overslept?		40.2
7a	Had any kind of job?	60.7	
b	Quit a job after a week?		2.8
c	Quit a job without giving at least two weeks notice?		6.9
8a	Ridden a bicycle?	94.6	
b	Fallen off your bicycle?		19.9
c	Been fined when you were riding a bicycle?		3.1
9a	Driven a moped or scooter?	11.7	
b	Had a souped-up moped or scooter?		32.1
c	Crashed your moped or scooter?		3.6
d	Been fined when you were driving a moped or scooter?		.0
10a	Driven a car?	86.2	
b	Been accused of causing a car accident while driving?		1.0
c	Gotten parking tickets?		8.7
d	Gotten speeding tickets?		14.6
e	Gotten lost or gone the wrong way for more than 10 minutes while driving?		39.8
f	Left your car keys in your car?		4.4
11a	Had a driver's license?	75.3	
b	Had your driver's license taken away from you by the police?		.0
12a	Bought a car?	6.3	
b	Had to spend at least €400,- to fix a car you had owned for less than half a year?		13.3
13a	Taken a trip by airplane?	33.1	
b	Missed a flight?		6.3
14a	Taken the train or the bus?	93.7	
b	Taken the wrong train or bus?		21.0
15a	Had any form of ID (driver's license, passport, birth certificate)?	97.9	
b	Had your ID replaced because you lost it?		3.4
16a	Lived in a rented apartment or other rental property?	72.0	
b	Been kicked out of an apartment or rental property before the lease ran out?		7.0
17a	Carried a key to your home?	96.2	
b	Had the key to your home replaced because you lost it?		5.7
c	Locked yourself out of your home?		16.1
d	Unintentionally left the door to your home unlocked?		16.5
18a	Been responsible for electricity, internet, phone, gas or water payments?	58.6	
b	Received a payment reminder for electricity, internet, phone, gas or water because you didn't pay on time?		12.1
c	Received a formal notice from a debt collection agency for your electricity, internet, phone, gas or water because you didn't pay on time?		2.1
d	Had your electricity, internet, phone, gas or water shut off because you didn't pay on time?		1.4
19a	Had a loan or purchased something on installment?	17.6	
b	Received a payment reminder for repayment of your loan or installment?		14.3
c	Received a formal notice from a debt collection agency for repayment of your loan or installment?		7.1
20a	Been responsible for rent or mortgage payments?	51.9	
b	Received a payment reminder for your rent or mortgage?		10.5
c	Received a formal notice from a debt collection agency for your rent or mortgage?		1.6
21a	Purchased lottery tickets?		14.2
22a	Used a debit card?	94.6	
b	Had a negative bank balance for more than two months?		7.1
c	Had a debit card rejected for insufficient funds while trying to make a purchase?		14.2
23a	Had a mobile/cellular phone?	99.6	
b	Lost a mobile/cellular phone?		6.3
c	Had to pay at least three times extra because you went over your call/text/data limit?		26.5
24a	Had a credit card?	21.8	
b	Had more than €4000,- in credit debt?		1.9
25a	Gambled for money (e.g., in a slot machine, casino or online games)?	15.9	
b	Lost more than €250,- from gambling?		10.5
26a	Been to a bar, restaurant, or hotel?	97.1	
b	Been kicked out of a bar, restaurant, or hotel by someone who works there?		2.6

Table A.1 (*Continued*)

Item		Percentage who made the decision	Percentage who experienced outcome
27a	Loaned more than €40,- to someone?	31.8	
b	Loaned more than €40,- to someone and never got it back?		13.2
28a	Borrowed more than €40,- from someone?	18.8	
b	Borrowed more than €40,- from someone and never paid it back?		2.2
29a	Been in a romantic relationship?	67.4	
b	Broke off a relationship?		21.1
30a	Had sex?	72.4	
b	Been diagnosed with an STD?		2.3
c	Had an unplanned pregnancy (or got someone pregnant, unplanned)?		.0
31a	Had a romantic relationship that lasted for at least one year?	54.0	
b	Cheated on your romantic partner?		4.7
32a	Had sex with a condom?	51.5	
b	Had a condom break, tear, or slip off?		15.5
33a	Had an alcoholic drink?	95.0	
b	Consumed so much alcohol you vomited?		33.5
c	Received a DUI for drunk driving?		.0
34a	Been out in the sun?	88.3	
b	Got blisters from sunburn?		30.3
35a	Purchased a product from the television (e.g., an infomercial)?	1.7	
b	Purchased a product from the television without reading information about that product's effectiveness?		50.0
36a	Read your horoscope?	59.4	
b	Found that your horoscope was accurate?		40.8
37a	Been in jail overnight for any reason?		.4
38a	Been in a public fight or screaming argument?		19.7
39a	Forgotten the birthday of someone close to you and did not realize until the next day or later?		30.5
40a	Had health problems because you were overweight (e.g., shortness of breath, type 2 diabetes, heart disease, high blood pressure)?		2.9
41a	Broken a bone because you fell, slipped, or mis-stepped?		7.9
42a	Owned an object with lucky properties (e.g., a rabbit's foot, fortune-doll, etc.)?		34.3
43a	Paid to speak to a psychic (i.e. in person or over the phone)?		2.1
44a	Started a diet that guaranteed fast weight loss?	15.5	
b	Gained weight after a diet, ending up at your original weight?		56.8
45a	Bought something from a second-hand online store?	24.3	
b	Paid in advance for something off a second hand online store, but did not receive anything?		5.2
46a	Purchased herbal remedies that enhance thinking or memory?		9.2
47a	Owned an object with healing properties (e.g., healing crystals, magnetic bracelets, mystical stones, etc.)?		4.6
48a	Received a chain letter or email?	48.5	
b	Forwarded a chain letter or email?		11.2
49a	Signed up for a subscription (e.g., a newspaper, magazine, phone, etc.)?	31.0	
b	Signed up for a subscription (e.g., a newspaper, magazine, phone, etc.) and after a while you wanted to get rid of it, but it turned out you were committed to it for a long time?		12.2
50a	Accepted an offer from a vendor on the street or at the door (except for charities)?	2.1	
b	Accepted an offer from a vendor on the street or at the door (except for charities) that you regretted later?		20.0

Note: Non-response to items is not included in the calculation of the percentage of decisions or the percentage of outcomes experienced.

References

Ananiadou, K., & Claro, M. (2009). 21st century skills and competences for new millennium learners in OECD countries. *OECD Education Working Papers*, *4*, 1. http://dx.doi.org/10.1787/218525261154

Arbuckle, J. L. (Ed.). (2013). *Amos 22 reference guide*. Crawfordville, FL: Amos Development Corporation.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.

Black, B. (2012). An overview of a programme of research to support the assessment of Critical Thinking. *Thinking Skills and Creativity*, *7*(2), 122–133. http://dx.doi.org/10.1016/j.tsc.2012.04.003

Bridgeman, B., & Morgan, R. (1996). Success in college for students with discrepancies between performance on multiple-choice and essay tests. *Journal of Educational Psychology*, *88*(2), 333–340. http://dx.doi.org/10.1037/0022-0663.88.2.333

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, *92*(5), 938–956. http://dx.doi.org/10.1037/0022-3514.92.5.938

Butler, H. A. (2012). Halpern Critical Thinking Assessment predicts real-world outcomes of critical thinking. *Applied Cognitive Psychology*, *26*(5), 721–729. http://dx.doi.org/10.1002/acp.2851

Butler, H. A., Dwyer, C. P., Hogan, M. J., Franco, A., Rivas, S. F., Saiz, C., et al. (2012). The Halpern Critical Thinking Assessment and real-world outcomes: Cross-national applications. *Thinking Skills and Creativity*, *7*(2), 112–121. http://dx.doi.org/10.1016/j.tsc.2012.04.001
Carstairs, J., & Myors, B. (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. *Computers in Human Behavior*, *25*(3), 738–742. http://dx.doi.org/10.1016/j.chb.2009.01.011
Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5
Clifford, J. S., Boufal, M. M., & Kurtz, J. E. (2004). Personality traits and critical thinking skills in college students: Empirical tests of a two-factor theory. *Assessment*, *11*(2), 169–176.
Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. http://dx.doi.org/10.1037/0021-9010.78.1.98
Dwyer, C. P., Hogan, M. J., & Stewart, I. (2012). An evaluation of argument mapping as a method of enhancing critical thinking performance in e-learning environments. *Metacognition and Learning*, *7*(3), 1–26. http://dx.doi.org/10.1007/s11409-012-9092-1
Ennis, R. H. (1996). *Critical thinking*. Upper Saddle River, NJ: Prentice Hall.
Ennis, R. H., & Weir, E. E. (1985). *The Ennis-Weir critical thinking essay test: An instrument for teaching and testing*. Boise, ID: Midwest Publications.
Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell critical thinking tests level X & level Z: Manual*. Boise, ID: Midwest Publications.
Facione, P. A. (1990). *The California critical thinking skills test: College level*. Millbrae, CA: California Academic Press.
Facione, P. A. (1998). *Critical thinking: What it is and why it counts*. Millbrae, CA: California Academic Press.
Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic*, *20*(1), 61–84.
Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65–83. http://dx.doi.org/10.1207/s15327906mbr4101_5
Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449–455. http://dx.doi.org/10.1037/0003-066X.53.4.449
Halpern, D. F. (2012). *Halpern critical thinking assessment: Test manual*. Mödling, Austria: Schuhfried GmbH.
Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future – A festschrift in honor of Karl Jöreskog*. Lincolnwood, IL: Scientific Software International.
Hau, K. T., Halpern, D. F., Marin-Burkhart, L., Ho, I. T., Ku, K. Y. L., & Chan, N. M. (2006). Chinese and United States students' critical thinking: Cross-cultural construct validation of a critical thinking assessment. In *American Educational Research Association Annual Meeting* San Fransisco, CA, April.
Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23. http://dx.doi.org/10.1037/a0014694
Kline, P. (2000). *The handbook of psychological testing*. London: Routledge.
Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, *4*(1), 70–76. http://dx.doi.org/10.1016/j.tsc.2009.02.001
Ku, K. Y. L., & Ho, I. T. (2010). Dispositional factors predicting Chinese students' critical thinking performance. *Personality and Individual Differences*, *48*(1), 54–58. http://dx.doi.org/10.1016/j.paid.2009.08.015
Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational & Organizational Psychology*, *84*(4), 817–824. http://dx.doi.org/10.1348/096317910x522672
Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2
Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218.
Moseley, D., Baumfield, V., Elliott, J., Higgins, S., Miller, J., Newton, D. P., et al. (2005). *Frameworks for thinking: A handbook for teaching and learning*. Cambridge, UK: Cambridge University Press.
Rotherham, A. J., & Willingham, D. T. (2010). 21st century skills: The challenges ahead. *Educational Leadership*, *67*(1), 16–21.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. http://dx.doi.org/10.1007/s11336-008-9101-0
Sonoma State University. (1996). *ICAT Critical Thinking Essay Test*. Rohnert Park, CA: Authors.
Spector, P. E., Schneider, J. R., Vance, C. A., & Hezlett, S. A. (2000). The relation of cognitive ability and personality traits to assessment center performance. *Journal of Applied Social Psychology*, *30*(7), 1474–1491. http://dx.doi.org/10.1111/j. 1559-1816.2000.tb02531.x
Sternberg, R. J. (1986). *Critical thinking: Its nature, measurement, and improvement*. Washington, DC: National Inst. of Education.
Taube, K. T. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. *The Journal of General Education*, *46*(2), 129–164.
Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. http://dx.doi.org/10.5116/ijme.4dfb.8dfd
Turner, N. E., Macdonald, J., & Somerset, M. (2008). Life skills, mathematical reasoning and critical thinking: A curriculum for the prevention of problem gambling. *Journal of Gambling Studies*, *24*(3), 367–380. http://dx.doi.org/10.1007/s10899-007-9085-1
Verburgh, A., Francois, S., Elen, J., & Janssen, R. (2013). The assessment of critical thinking critically assessed in higher education: A validation study of the CCTT and the HCTA. *Education Research International*, *2013*, 13. http://dx.doi.org/10.1155/2013/198920
Voogt, J., & Roblin, N. P. (2010). *21st century skills: Discussion paper*. University of Twente.
Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, *44*(3), 299–321. http://dx.doi.org/10.1080/00220272.2012.668938
Vrije Universiteit Amsterdam. (2010, September 13). (*Sports scientists VU: Balance wristbands Dutch football team do not work*) *Bewegingswetenschappers VU: Balansbandjes Oranje werken niet*. Vrije Universiteit Amsterdam., retrieved from http://www.vu.nl/nl/nieuws-agenda/nieuws/2010/jul-sep/bewegingswetenschappers-vu-balansbandjes-oranje.asp
Watson, G. B., & Glaser, E. M. (1980). *Manual for the Watson Glaser critical thinking appraisal*. Cleveland, OH: Psychological Corporation.
Wolff, H., & Preising, K. (2005). Exploring item and higher order factor structure with the Schmid-Leiman solution: Syntax codes for SPSS and SAS. *Behavior Research Methods*, *37*(1), 48–58. http://dx.doi.org/10.3758/BF03206397