Psychological Test and Assessment Modeling, Volume 56, 2014 (2), 195-216

Examining the measurement equivalence of the Conditional Reasoning Test for Aggression across U.S. and Croatian samples

Zvonimir Galić[1], *Kelly T. Scherer*[2] *& James M. LeBreton*[3]

Abstract

The Conditional Reasoning Test for Aggression (CRT-A; James et al., 2005) is based on the ideas that aggressive individuals use motive-based cognitive biases to see their behavior as reasonable and that those biases can be measured with specially designed inductive reasoning tasks. The test has shown promising psychometric characteristics for U. S. samples but has not been validated in other cultural contexts. In our study, we examined whether the items from the CRT-A were invariant across culture by testing whether these items displayed differential item functioning (DIF) across Croatian (N=530) and U.S. (N=1479) samples. The Lord's Chi Square (Lord, 1980), the Raju UA index (Raju, 1988), the Mantel-Haenszel procedure (Mantel & Haenszel, 1959), and the logistic regression procedures (Swaminathan & Rogers, 1990) revealed that DIF was pervasive. Although an implicit measure of personality, the CRT-A seems susceptible to differential item functioning in another culture.

Keywords: Conditional Reasoning Test for Aggression, differential item functioning, implicit personality

[1] *Correspondence concerning this article should be addressed to:* Zvonimir Galić, PhD, Department of psychology, University of Zagreb, Luciceva 3, 10 000 Zagreb, Croatia; email: zvonimir.galic@ffzg.hr

[2] Purdue University

[3] Penn State University

On July 22nd 2011 Anders Breivik killed 77 people in Oslo and Utoya, Norway. The court declared him sane and sentenced him to 21 years in jail. During the process Breivik advocated his violence as a justifiable, political act (Lewis & Lyall, 2012). Breivik ascribed the main responsibility to the ruling Labor party that had permitted a wave of Muslim emigrants and, thus, taken ethnic rights from their citizens. Breivik claimed the killing was a completely reasonable act because it was against a group of people who oppressed him and Norwegian people, and deserved to be punished (Spiegel Online, 2012).

Breivik's defense gives an insight into how extremely aggressive individuals reason. According to James and associates (James et al., 2005; James & LeBreton, 2010, 2012) there is a pattern in reasoning that can be used to identify individuals prone to aggression. They use specific cognitive biases which help them to claim reasonability of their behavior (e.g., hostile attribution bias; derogation of target bias; potency bias).

James and LeBreton (2010, 2012) believe that these cognitive biases can be measured with the Conditional Reasoning Test for Aggression (CRT-A). Current research seems to suggest that the CRT-A has good psychometric characteristics and demonstrates modest to high levels of predictive validity. The CRT-A score predicted various undesirable behaviors both in the laboratory (e.g., lack of truthfulness about extra credit, James et al., 2005) and in real life situations such as aggressive behavior in basketball games (Frost, Ko, & James, 2007), traffic violations, or counterproductive work behavior (Bing et al., 2007).

So far, the CRT-A has not been tested in other cultural contexts. In our paper we report the results of a study that examined measurement equivalence of the CRT-A items across U. S. and Croatian samples. We find Croatia to be an interesting context for the test of this approach because of its differences from the U. S. in main cultural dimensions (e.g., power distance, Hofstede, 2001) and specific experiences related to aggression (e.g., war in 1990s and process of transition). Information about how this assessment system works in different circumstances might be important for its future development.

In this paper we will first briefly describe the main tenets of the conditional reasoning approach to aggression measurement. Additionally, we will review the results of the other studies that examined differential item functioning (DIF) of personality measures cross-culturally. In the remaining parts of the paper, we will describe the results of the study testing the measurement equivalence of the CRT-A items between Croatian and the U. S. samples.

Conditional reasoning approach to measurement of aggression

The conditional reasoning approach rests on the assumption that individuals are motivated to believe their behaviors are reasonable, appropriate, and rational as opposed to unreasonable, inappropriate, and irrational (James & LeBreton, 2012). James and LeBreton (2010; 2012) argued that individuals with a strong motive to aggress are able to reconcile the desire to harm others with societal expectations for socially appropriate deportment.

This reconciliation occurs by invoking implicit cognitive biases designed to enhance the logical appeal of aggressive behavior. These biases were denoted justification mechanisms by James (1998) to emphasize the critical role they play in rationalizing motive-driven behaviors. According to James (1998) the biases are relatively stable, and operate largely out of conscious awareness. They influence the reasoning process through selective perception, confirmatory biases during the information search, and the process of reaching a causal inference, thus, making reasoning conditional on one's personality.

James et al. (2005) identified six justification mechanisms that aggressive individuals use to enhance the appeal of their harmful actions. First, aggressive people are prone to see hostility and threat in others' behavior, and, thus, perceive their own aggressive behavior largely as an act of self-defense. This *hostile attribution bias* can be in action even when others' behavior is benign or even friendly. Second, aggressive individuals see interactions with other people as "contests to establish dominance versus submissiveness" (James & LeBreton, 2010, p. 31). The *potency bias* reflects a positive evaluation of one's aggressive behavior as a means of expressing dominance in a social situation. In contrast, the lack of aggression is framed as a sign of weakness and impotence. Moreover, aggressive individuals are inclined to frame retaliation as a more logically appropriate response to conflict than reconciliation (*retribution bias*), which makes them prone to aggressive behavior when they find themselves in a conflict. Further, aggressive individuals often see themselves as victims of powerful others who exploit or oppress them (e.g., supervisors, teachers, government agencies, large corporations). The *victimization bias* creates feelings of anger and injustice which result in believing that acts of aggression are a seemingly reasonable strategy of coping with exploitation and victimization. Fifth, aggressive individuals derogate targets of their behavior and ascribe to them negative characteristics such as evilness, stupidity, unethicality, or immorality. Such ascriptions enable aggressive individuals to aggress because the targets of aggression are seen in some way as deserving of harm (*derogation of target bias*). Finally, aggressive individuals often find social norms to be repressive and restrictive and their aggressive behavior is thus justified as a means of liberation (*social discounting bias*). According to James and his associates (James et al., 2005; James & LeBreton, 2010; 2012; LeBreton, Barksdale, Robin, & James, 2007) these six biases are relatively stable across time and situations and reflect the implicit aspects of the motive to aggress.

James et al. (2005) devised a test intended to measure an individual's inclination towards the justification of aggression – the Conditional Reasoning Test of Aggression (CRT-A). This test rests on the idea that it is possible to determine an individual's aggressiveness by observing the above described biases in his/her reasoning. The CRT-A consists of 25 inductive reasoning problems in which a story in the problem's stem is followed by four possible answers. Respondents are instructed to solve each inductive reasoning problem by selecting the answer which seems most reasonable to them. Three of the CRT-A problems are regular inductive reasoning problems with only one correct answer, and three illogical solutions. They are included on the test to improve the face validity of the CRT-A (i.e., to put respondents into a problem-solving mindset looking for correct vs. incorrect solutions to inductive reasoning problems). These items are not used to draw inferences about aggression.

Instead, the remaining 22 CRT-A items are used to infer aggression. Each of these items consists of an item stem along with four possible inferences. Two of the inferences are designed to be inductively illogical solutions to the problem, and two are designed to be inductively plausible solutions. For each item, one of the plausible alternatives was developed using one or more justification mechanisms for aggression, and thus is designed to be logically compelling to individuals with a motive to aggress. The other plausible alternative was developed using socially adaptive reasoning and is designed to be attractive to individuals with a weak motive to aggress. More detailed descriptions of these items are available in James and LeBreton (2010; 2012) and LeBreton et al., (2007).

A sample item from the CRT-A is presented in Table 1. Respondents are asked to find the inference representing the biggest problem with the information presented in the item's stem. The answers (a) and (c) are obviously incorrect, and irrelevant. Both alternatives (b) and (d) are logically plausible. Answer (d) is expected to be endorsed by individuals who habitually use *hostile attribution bias* in their reasoning, whereas the non-aggressive alternative (b) should appear the most logical to non-aggressive individuals because their reasoning is "shaped by the social adaptive values and the ideologies they have internalized" (James & LeBreton, 2012, p. 36).

Unlike most of the other psychological tests intended to measure implicit aspects of personality, the CRT-A shows good psychometric characteristics (James et al. 2005, James & LeBreton, 2010, 2012). Studies reported by James and associates revealed acceptable internal consistency (.76 based on calculation using a derivative of the KR-20 formula) and test-retest reliability (.82, based on two alternative forms of the test measured within a two month period). The scores on the CRT-A are found in most cases to be uncorrelated with cognitive ability or self-report personality, and reveal a factor structure consistent with the theory underlying the implicit motive to aggress (James & LeBreton, 2012; Galić, Scherer & LeBreton, 2014). Moreover, the CRT-A seems to be insensitive to situational contexts, and therefore unsusceptible to deliberate response distortion (LeBreton et al., 2007). Although there are still disputes about exact values of validity coefficients (James & LeBreton, 2010; 2012; Berry, Sackett, & Tobares, 2010), the CRT-A

Table 1:
An Illustrative Conditional Reasoning Problem

Store employees are told to watch out for people who look like shoplifters. If a customer looks like a shoplifter, then employees are supposed to watch the customer closely. Which of the following is the biggest problem with this practice? a. Most retail stores don't open until 10:00 in the morning. b. Many customers who look like shoplifters are honest and do not steal. c. Parking is getting harder to find in shopping malls. d. Abuse by store employees who use it as an excuse to bother people they don't like.

scores predicted aggressive behaviors/CWBs in both laboratory and field settings. Higher scores have been linked with outcomes such as verbal and physical aggression but also theft and lying (James & LeBreton, 2010). Additionally, higher CRT-A scores were related to more passive forms of aggression including higher absenteeism among nuclear facility operators, production deviance (e.g., not reporting to workplace assignments) among temporary workers, and higher turnover rates among restaurant employees (Bergman, McIntyre, & James, 2007). The average uncorrected criterion-related validity drawn from studies which relied on predictive designs and objective measures of behavior was .41 (James & LeBreton, 2012).

Differential item functioning of personality measures in cross cultural comparisons

Before a psychological instrument is used for group comparisons or original normative data are transferred to another cultural context, psychologists should establish that two (or more) versions of the instrument show measurement equivalence across cultures. Measurement equivalence between different versions of the same inventory exists if individuals who have the same trait levels but come from different groups have equal scores on each of the items (Drasgow, 1984). The items that depart from the measurement equivalence principle show differential item functioning (DIF), which makes cross-cultural comparisons that are based on them questionable.

Previous studies that used either the classical test theory (CTT) or the item response theory (IRT) approach to test for measurement equivalence of personality instruments across different cultures showed that the DIF among personality items is pervasive. These conclusions were largely the same irrespective of the particular personality inventory examined or the specific procedure used to test for DIF. Using confirmatory factor analysis (CFA), Nye, Roberts, Saucier, and Zhou (2008) revealed that more than half of the items of the 40-item adjective measure of the Big Five personality traits functioned differently across Chinese, Greek and American samples. Similar results were reported by Church, Alvarez, French, Katigbak and Ortiz (2011) who used the CFA approach and found that DIF was prevalent in the NEO-PI-R items. About 40-50% of the items exhibited some form of DIF across the U.S., Mexico and Philippines samples.

Similar patterns of findings have been obtained by researchers using IRT approaches to test for DIF. Huang, Church, and Katigbak (1997) compared the NEO-PI-R items between large samples of Filipino and American college students using parameter equating and model comparison IRT methods. Additionally, they examined DIF using the Mantel – Haenszel (M-H) procedure. The three approaches yielded fairly consistent findings with roughly 40% of the items displaying DIF. Johnson, Spinath, Krueger, Angleitner, and Riemann (2008) used German and Minnesotans twins' samples to test for DIFs on the Multidimensional Personality Questionnaire (MPQ). The IRT analyses revealed a large number of DIF items within the MPQ scales. The DIF items were shown to significantly influence conclusion about the differences on personality traits between the samples coming from the two cultures. Finally, Kulas, Thompson, and Anderson (2011)

tested the DIF of the items from the Dominance scale of the California Personality Inventory across four samples (American normative, U.K. managers, U.S. managers, and Indian managers). Again, their analyses revealed pervasive DIF, especially when Indian managers were compared to the other three groups.

The DIF analyses of personality instruments are almost completely limited to the self-report personality questionnaires. To the best of our knowledge, Hofer, Chasiotis, Freidlmeier, Busch, and Campos (2005) reported the only study that used DIF procedures to explore differential functioning of an implicit personality measure. In their study, Hofer et al. (2005) compared TAT-type picture-story items to find a culture independent set of stimuli intended for measurement of the affiliation and power motives across cultures. The comparison of responses to the eight cards across German, Cameroonian, and Costa Rican participants using the M-H procedure revealed that four items displayed DIF, making them questionable for use in cross-cultural comparisons.

Therefore, all described studies suggested that cross-cultural comparisons using above mentioned personality measures are questionable because they might be contaminated by DIFs. New types of personality assessment such as the CRT-A could be useful in international research if their cross-cultural measurement equivalence is supported.

DIF analysis of the Conditional Reasoning Test for Aggression (CRT-A)

There are two main reasons why psychometric properties of the conditional reasoning approach to personality measurement should be cross-culturally tested. First, much of the extant cross-cultural comparisons of personality has relied on self- and peer-report personality questionnaires, in the most of cases measuring traits within the Five Factor framework (e.g., McCrae, Costa, Pilar, Rolland, & Parker, 1998; McCrae et al., 2005). However, pervasive DIF on personality questionnaire items makes cross-cultural comparisons questionable. The probable cause of DIF and main problem with self- and peer report measures is that they are prone to the Reference Group Effect (RGE, Heine, Buchtel, & Norenzayan, 2008; Heine, Lehman, Peng, & Greenholtz, 2002). When responding on a Likert type item, individuals from different cultures compare themselves with a reference group, and these reference groups differ between cultures. This problem could cause the prevalence of the DIF on personality items but it also leads to the question of validity of the data collected with self and peer reports in a context other than that in which the instrument was originally developed. Conditional reasoning approach represents a reasonable alternative to "ordinary" personality measures. Considering that the CRT-A is a personality test that indirectly measures motive-based cognitive biases, it should not be prone to the RGE or similar effects, such as impression management, that could have caused the DIF on a self-report personality measure.

Second, the conditional reasoning approach to aggression measurement assumes that the motive based cognitive biases used to justify aggression are universal (James & LeBreton, 2010; 2012) and should hold in various cultural circumstances. However, the first step in making such claims is to establish the cross-cultural equivalence of measures assessing the biases.

In our study we tested measurement equivalence of the CRT-A in Croatia. We believe that Croatia represents an interesting context for examining the CRT-A's assumptions. First, it represents a significantly different cultural context than the U.S. (Hofstede, 2001). Second, the experience of war (1991-1995) and hard process of transition from socialistic to a free market country (Tanner, 1997) gave Croatian citizen extensive experience with different forms of aggressive behavior and different justifications that followed them. The evidence of similar relationship between aggressive responses on conditional reasoning problems and the scale scores in Croatian and the U.S. samples would further support the validity of the conditional reasoning approach to personality measurement. The information about CRT-A's cross-cultural measurement equivalence in this context would have practical implications for its use in various scientific (e.g., cross-cultural comparison of personality profiles) and practical purposes (e.g., using the U.S. norms for personnel selection or individual counseling).

Method

Participants and procedure

Data were collected from a sample of 2,074 undergraduate students enrolled in large, introductory courses in psychology and management in the U.S. and Croatia. Participants earned course credit for completing the CRT-A. As per the recommendations provided in the CRT-A test administration guidelines (James & McIntyre, 2000), data of participants who endorsed five or more illogical distractor responses were removed from the analysis as were the data of participants whose responses were missing. This resulted in the reduction of the U.S. sample by 53 participants, and of the Croatia sample by 12 participants, resulting in a final U.S. sample of 1,479, a final Croatian sample of 530, and thus a final combined sample of 2,009. Average age in the Croatian sample was 21.54 (sd = 2.47) with 53.4 % female participants. Although, the exact data on U.S. sample characteristics were not available, we believe that the samples were comparable considering that they came from the same type of studies (psychology and management) that are similar in age and gender structure in Croatia and the United States.

Measure

The CRT-A consists of 25 inductive reasoning items designed to measure the justification mechanisms (JMs) associated with the implicit motive to aggress (James & McIntyre, 2000; James & LeBreton, 2010; 2012). Each item has four response options which include: a) an inductively logical aggressive response; b) an inductively logical response based on non-aggressive or socially adaptive ideology and reasoning; and c) two illogical responses. As stated earlier, the three test items (1, 2, and 6) are classic inductive reasoning items designed to put respondents into a problem-solving mindset. Consequently, 22 out of the 25 test items are used to assess the JMs of aggression (James & McIntyre, 2000). As James and McIntyre (2000) recommended, we scored the remaining 22 items

such that each aggressive response earned a "1" and non-aggressive responses earned a "0." In our data analysis, described below, we analyzed these dichotomous scores for DIF across the U.S. and Croatia samples.

Translation and adaptation process

Three Croatian researchers proficient in the English language independently translated the CRT-A into Croatian and held several meetings to discuss the discrepancies in translations. Considering that all translators observed that some information included in the CRT-A items was culture-specific, certain changes were made to the Croatian version. Specifically, minor changes were made to four items (3, 6, 9 and 11) which included changing the names of individuals and places from American names/places to Croatian names/places. The only item that underwent major changes was the item exploring reasons of recent improvements of the American car industry. In order to keep the story's appeal to the respondents similar, American car industry was replaced with Croatian industry of refrigerators. The response options for this item in the Croatian version were formulated so that the same aggression justification mechanisms would be operative as were in the original instrument. Both the original and the English translation of the adapted version of the problem are listed in the Appendix.

The adapted version of the CRT-A was, together with the original version of the instrument, then sent to the two psychologists experienced in personality assessment, familiar with the conditional reasoning approach, and bilingual in Croatian and English. Minor objections raised by these psychologists were adapted in the final version of the Croatian version of the CRT-A. Our approach to test the translation was consistent with the recommendation for a cross-cultural test adaptation (e.g., Geisinger, 1994, Hui & Triandis, 1985), and the International Testing Commission Guidelines for Translating and Adapting Test (the International Testing Commission, 2010).

Analyses

In our analyses, we first calculated item descriptive statistics and item-total correlations separately for Croatian and the U.S. samples. As reported in James and LeBreton (2012), results of a factor analysis of the CRT-A items on a large sample of participants (n=4,772) revealed that items tended to cluster into three sub-factors labeled External Justifications, Internal Justifications, and Powerlessness. Item-total correlations were calculated as the biserial correlation between a response and a subscale result (Lord & Novick, 1968). Before DIF analyses, we tested for the unidimensionality of the items included in the three subscales using principal axis factoring. Because the response format was dichotomous we estimated tetrachoric correlations among the tests items using the "polycor" package (Fox, 2007) from the R-program (R-development Core Team, 2010). In all subsequent DIF analyses, the Croatian sample was considered as the focal group while the American sample was treated as the reference group.

DIF was tested using both IRT and non-IRT indices. Within the IRT model, we used Lord's Chi-Square (Lord, 1980) and Raju's unsigned area (UA) DIF indices (Raju, 1990). While Lord's Chi-Square shows whether the item parameters are equal in the two populations given the sample-based item parameter estimates, Raju's UA indicator reflects the size of the area between the item response functions (IRF) obtained on the two samples. The unsigned area means that the UA indicators accumulate differences between IRFs irrespective of their sign, and therefore reveal both differences in item-difficulties and item-discriminations.

In addition we used the Mantel – Haenszel (1959) procedure, and logistic regression (Swaminathan & Rogers, 1990) as non-IRT based techniques of DIF testing. The M-H technique compares item performance of focal and reference groups across different score levels. If an item does not exhibit DIF, the focal and reference group at the same score level are expected to show similar response patterns. Within the logistic regression approach, a model is fitted where the probability of correct response on an item is predicted based on the total test score, the group membership and the interaction between group membership and the total test score. If the group membership and the interaction explain the probability of correct answer above the test score level, item is said to exhibit DIF. Useful characteristics of the M-H and logistic regression procedures are effect size statistics which are helpful for evaluating the magnitude of an item's DIF. The "difR" package was used to perform all the DIF analyses (Magis, Beland, & De Boeck, 2010)

Considering that previous research demonstrates that α errors in DIF analyses are relatively high when the number of items is low and the samples are large (Teresi, Ramirez, Lai, & Silver, 2008), only DIF below $p < .01$ were considered to be statistically significant in all our analyses. A similar procedure of the DIF analyses was also reported by Huang et al. (1997) and Budgell, Raju, and Quartetti (1995).

Results

Descriptive statistics

The item difficulties (p-values) and item-total correlations for Croatian and the U.S. samples are shown in Table 2. Generally, p-values were low for both samples (average p-value was 0.24 for Croatian, and 0.18 for the U.S. samples) which is in accordance with previous findings that, on average, only a small proportion of respondents selected the aggressive answers (James & LeBreton, 2012). The item-total correlations were reasonably high indicating that the items within a subscale have a common measurement object.

For the first factor p-values (i.e., item difficulties) ranged between .10 and .38 for the Croatian sample, and .04 and .28 for the U.S. sample. Item-total correlations for the first factor ranged between .31 and .63 for the Croatian sample, and between .31 and .56 for the U.S. sample.

Table 2:
The Item Difficulties (p-values) and Item-Factor Biserial Correlations for Conditional Reasoning Problems of the Conditional Reasoning Test for Aggression on Croatian and the US samples.

Item Number and Theme		**Croatian sample (n=530)**		**U.S. sample (n=1479)**	
Factor 1: External justification items		*p-value*	*Item-total biseria1*[1,2]	*p-value*	*Item-total biserial*[1,2]
(11)	a homeless man	.10	.63	.19	.46
(15)	permits to carry guns	.10	.31	.14	.45
(16)	American cars/Croatian fridges	.29	.56	.11	.50
(17)	store employees vs. shoplifters	.16	.49	.19	.48
(18)	bonuses for employees	.24	.54	.04	.42
(19)	search on employees	.23	.54	.28	.54
(20)	gangs	.33	.51	.23	.51
(22)	hold up victims	.16	.43	.06	.31
(23)	divorces	.28	.44	.28	.53
(24)	employee's revenge	.38	.63	.13	.56
(25)	agreement between countries	.26	.56	.05	.46
Factor 2: Internal justification items					
(4)	aggressively going after customers	.22	.66	.22	.64
(5)	generals	.12	.52	.05	.50
(7)	an eye for an eye	.07	.55	.05	.50
(8)	bosses and employees	.04	.55	.08	.54
(13)	duels with swords	.15	.68	.22	.66
(21)	wild animals	.16	.66	.22	.68
Factor 3: Powerlessness items					
(3)	late for meetings	.21	.62	.33	.64
(9)	new technology and workplace	.24	.59	.32	.66
(10)	Girl Scouts and Boy Scouts	.31	.60	.06	.57
(12)	good product at a low price	.63	.62	.39	.69
(14)	a new girl at the high school	.53	.62	.27	.64

Note.[1]Item-total correlations were calculated as the biserial correlation between a response and a subscale result; [2] all correlations are significant at $p < .01$ level.

On the second factor, the range of p-values was between .04 and .22 for Croatian and .05 and .22 for the U. S. sample. Item-total correlations ranged between .52 and .68 for Croatian respondents and between .50 and .68 for the U.S. respondents.

Finally on the third factor, the range was between .21 and .63 and .06 and .39 for Croatian and the U.S. samples, respectively. The item-factor correlations were between .59 and .62 for the Croatian sample, and .57 and .59 for the U.S. sample, respectively.

Internal consistency coefficients based on item-total biserial correlations and calculated using a derivative of the KR-20 formula (Guliksen, 1950; Equation 21, p. 389) were 0.72 and 0.66 for the External Justification factor, 0.65 and 0.62 for Internal Justification factors, and 0.58 and 0.64 for Powerlessness for Croatian and the U.S. samples, respectively. Correlations between the three factors were similar for both samples. Correlations between External and Internal Justification were .27 and .27, between External Justification and Powerlessness .06 and .10, and between Internal Justification and Powerlessness .20 and .14 for Croatian and the U.S. samples, respectively. Except for the correlation between the first and the third factor in the Croatian sample, all other correlations were significant at $p <. 01$ level. Internal consistency of the total test was .73 for the Croatian sample and .71 for the U.S. sample, which was consistent with prior research (see James & LeBreton, 2012 for a review).

Based solely on p-values, item-factor correlations, and reliabilities, the two forms of the CRT-A seem reasonably similar. The differences in item difficulties (i.e., p-values) did not appear excessive and item-total correlations were roughly equivalent. However, p-values are dependent on a trait distribution and the size of item-total correlations on item difficulties obtained on a specific sample (Raju & Ellis, 2002). Thus, we proceeded with stronger tests of DIF.

Test of unidimensionality

To test for the unidimensionality of External Justification, Internal Justification and Powerlessness subscales, we conducted factor analyses of the three subsets of items using principal axis factoring separately on each sample. In each of the six cases, the scree plot revealed one dominant factor underlying intercorrelations among items. The requirement that the first factor accounts for at least 20% of the total variance among items (Reckase, 1979) was met for all three factors on the Croatian sample, and for Internal Justification and Powerlessness subscales on the U.S. sample. For the External Justification factor, the percentage of explained variance was just below the threshold (18.29%, eigenvalue=2.01). Considering that the scree-plot indicated one dominant factor and that previous studies showed that the IRT assumptions are relatively robust to violations (Drasgow & Hulin, 1990), we believe that the subscale factor structure did not influence the DIF analyses.

DIF analyses

The items were calibrated with a two-parameter logistic (2PL) model separately for the Croatian and U.S. samples. In this model, the probability of an aggressive response is modeled by $P(\theta)= \frac{e^{1.7a_i(\theta-b_i)}}{1+e^{1.7a_i(\theta-b_i)}}$, where θ is the respondent's level of aggressiveness on a factor, *a* is an item discrimination, and *b* is an item difficulty coefficient. The parameters on both samples were estimated using maximum likelihood functions from the "ltm" package of the R-program (Rizopoulos, 2006). The fit of the 2PL model to the CRT-A items was tested with the item fit statistics from the same program package.

The item fit statistics showed that the large majority of the items fit the 2PL model well. The only exceptions were the "hold up victims" (item number 22) item which did not show good model fit on both samples, and the "generals" (item number 5) item which did not fit the model in the Croatian sample. However, the overall fit of the items to the 2-PL was considered adequate across both samples, and therefore all items were included in DIF analyses.

In order to test whether the CRT-A items demonstrated significant DIF between the two versions of the instrument, we calculated Lord's Chi-Square (Lord, 1980) and Raju's Unsigned Area indices (Raju, 1988) using the 2PL IRT model. Items in the focal group were rescaled to those of the reference group using equal means anchoring (Cook & Eignor, 1991).

Additionally, we conducted the non-IRT based M-H and logistic regression procedures. Irrespective of the method used, DIF was observed on most of the CRT-A items – Lord's Chi Square found DIF on 16 items, Raju's UA indicators found DIF on 13 items, the M-H procedure found DIF on 18 items, and the logistic regression on 19 items. The three methods showed a moderate agreement in DIF detection. The agreement was better within the type of analysis (i.e., an IRT DIF method agreed more with the other IRT method than with a nonIRT procedures and vice versa). Considering that it is common for different DIF criteria to lead to somewhat different conclusions (Borsboom, 2006), we decided to define as "true" DIFs those items for which the results of the four procedures converged. Those items and related outcomes of the DIF analyses are shown in Table 3.

For the eight items that revealed DIF using all four procedures we wanted to see how large the observed DIFs were, and whether they were uniform (i.e., constant across the trait levels) or non-uniform (i.e., different between the trait levels). In Table 4 we reported the effect size of the DIFs obtained through the M-H procedure ($\hat{D}_i$) and the effect sizes obtained within logistic regression (*Δ*Nagelkerke R^2). The effect sizes from the logistic regressions were reported separately for uniform (the effect of group membership) and nonuniform DIF (the effect of the interaction between the group membership and the trait level). The results shown in Table 4 reveal that DIFs on the CRT-A items between Croatian and the U.S. respondents were mostly uniform and large in size. According to Dorans and Holland (1993), an item has large DIF when the absolute value of

Table 3:
Results of the Differential Item Functioning Analyses: Lord's Chi Square DIF, Raju's Unsigned Area, Mantel-Haenszel and Logistic Regression Procedure Indices

Item Number and Theme		**IRT DIF Methods**		**Non-IRT Methods**	
Factor 1: External justification items		*Lord's Chi Square DIF*	*Raju's DIF Unsigned Area indices*	*Mantel-Haenszel α*	*Logistic regression (uniform and non-uniform, ΔR^2)*
(15)	permits to carry guns	2846.65***	-2.97**	0.38***	0.22***
(16)	American cars/croatian fridges	43.26**	-3.78***	2.13***	0.07***
(17)	store employees vs. shoplifters	18.33**	-2.75**	0.41****	0.14***
(18)	bonuses for employees	24.43**	-3.73***	4.66***	0.22***
(22)	hold up victims	2162.90***	25.44***	2.02***	0.09***
(24)	employee's revenge	132.35***	-7.75***	3.01***	0.10***
(25)	agreement between countries	50.28***	-5.37***	4.38***	0.19***
Factor 3: Powerlessness items					
(9)	new technology and workplace	593.32***	8.49***	0.28***	0.13***

Note. α= common odds ratio, **p < .01; ***p < .001.

$\hat{D}_i$ statistic is higher than 1.50, and is significantly greater than 1.00. The effect size given within the M-H procedure revealed that all eight items had large DIF. Seven of them had an effect size ($\hat{D}_i$) significantly larger than 1. The exception was the "hold up victims" item whose lower margin of the 95% confidence interval was just below threshold of 1 (-0.87). This conclusion was further supported with the effect sizes obtained within the logistic regression. If classified according to the Jodoin and Gierl (2001) scale, uniform DIF on seven items can be categorized as a large, and the one (on the "American cars/Croatian fridges" item) as a moderate. At the same time, a significant non-uniform DIF was shown only on one item ("permit to carry guns") and, by its effect size, that DIF can be categorized as negligible (Jodoin & Gierl, 2001).

Seven of the eight items that showed DIF were found on the External Justification factor, and one item on the Powerlessness factor. The fact that DIFs were uniform indicated that they are a function of the differences in the item difficulties rather than in the item discriminations. Of the seven items from the External Justification factor that showed DIF, five have lower item difficulties (see Table 2) in the Croatian sample suggesting Croatian

Table 4:
The Effect sizes for the Eight DIF Items obtained in the Mantel-Haenszel and the Logistic Regression Procedures.

Item Number and Theme	**Mantel-Haenszel procedure**	**Logistic regression**	
	$\hat{D}_i$ (95% CI)	Uniform (Δ Nagelkerke R^2)	Nonuniform (Δ Nagelkerke R^2)
Factor 1: External justification items			
(15) permits to carry guns	2.30 (1.48, 3.12)	0.19***	0.04**
(16) American cars/Croatian fridges	-1.78 (-2.42,-1.12)	0.06***	0.00
(17) store employees vs. shoplifters	2.11 (1.39, 2.63)	0.13***	0.01
(18) bonuses for employees	-3.62 (-4.42, -2.82)	0.22***	0.00
(22) hold up victims	-1.65 (-2.44, -0.87)	0.09***	0.00
(24) employee's revenge	-2.59 (-3.32; 1.95)	0.10***	0.00
(25) agreement between countries	-3.47 (-4.25; 2.70)	0.19***	0.00
Factor 3: Powerlessness items			
(9) new technology and workplace	2.97 (2.32; 3.62)	0.13***	0.00

Note. $\hat{D}_i$ = DIF effect size from the M-H DIF procedure; **p < .01; ***p < .001.

participants were more likely to select items based on the aggressive justification mechanisms compared to the U.S. participants when the total score on that factor is held constant. The U.S. respondents endorsed three DIF items more often than the Croatians (the "permit to carry guns" and the "store employees vs. shoplifters" items from the External justification factor and the "new technology workplace" item from the Powerlessness factor).

Discussion

In our study we tested for differential functioning of the items from the CRT-A across Croatian and U.S. samples using both IRT (Lord's Chi Square and Raju's Unsigned Area) and nonIRT DIF methods (M-H procedure and Logistic Regression). The four DIF methods showed complete convergence for eight items which were considered as "true" DIF. Those items make 36.36% of the total test leading us to the conclusion that DIF was pervasive on the CRT-A items across the two samples.

According to Ellis, Becker, and Kimmel (1993) there are three reasons why translated items could show DIF. First, DIF could be caused by ineffectively translated items. Second, the meaning of an item might be different between the cultures, and, third, knowledge/experience related to an item might be culturally specific, and, thus, result in different response patterns. Considering that we tried to avoid the first two causes of DIF with a carefully-designed translation and adaptation procedure of the instrument, we believe that the third reason was likely the dominant cause of DIF observed in our study.

The means through which culturally-specific experience could influence responses of our participants could be comprehended if we carefully analyze the eight items which showed DIF across all three procedures. The fact that the DIFs were uniform rather than non-uniform indicates that DIFs appeared to be driven by differences in item difficulties than in item discriminations. Seven of these eight items are related to the first factor underlying the CRT-A (External Justification, James & LeBreton, 2012), and five of these seven had higher p-values (i.e., lower item difficulties) for Croatian participants. This means that, when equated with the U.S. participants in the level of the latent factor, the participants from Croatian sample more readily endorsed responses that are based on hostile attribution and victimization by powerful others biases, the two justification mechanisms that form the External Justification factor. The cause for this could be found in general differences between the two cultures and some culturally-specific experience of Croatian participants related to aggression.

First, Croatian participants might be more ready to endorse the alternatives James and McIntyre (2000) consider as aggressive because some aspects of the culture not related to aggression but connected with described justification mechanisms are different between Croatia and the United States. For example, the power distance is significantly larger in Croatia in comparison with the United States (Hofstede, 2001). This may lead Croatian participants to really experience victimization by powerful others more often than the U.S. based respondents and, consequently, select the answer that should be related to this justification mechanisms. Additionally, many of the Croatian participants and their families experienced the war for independence in the 1990s as well as a turbulent postwar period (Tanner, 1997). Moreover, they experienced childhood during the period of almost permanent economic crisis that followed the transition from a socialistic, planned economy to a democratic free-market society. This period was marked with a significant amount of corruption, strong feelings of injustice and growing social inequalities (Galić & Plećaš, 2012; Nestić, 2002; Vojnić, 1994). These social circumstances probably reflected in situations in which participants or people around them could really have the

experience of being victims of powerful others and observe hostile intentions in others' behavior. These situations could enhance appeal of responses that are based on hostile attribution and victimization by powerful others biases, and make them more habitual in a respondent's thinking.[4] Stated alternatively, the hostile attribution and victimization "biases" were not operating similarly in the U.S. and Croatian samples. Specifically, due to some characteristics of the culture and the "reality" of life in Croatia during the last 20 years Croatian participants might have been victims of powerful others and might have encountered others who truly had hostile intentions. The base rates for such interactions were likely higher than those found in the United States during this same time period. Thus, what is defined as an aggressive cognitive bias among respondents in the U.S. samples may simply reflect relatively veridical perceptions of social interactions in Croatia.

A similar explanation may be offered for the three DIF items which revealed lower difficulties in the U.S. sample ("permit to carry guns", "store employees vs. shoplifters" and "new technology and workplace"). The permit to carry guns and the related hostile intention of people who ask for one seem to be more salient concerns and so they are a commonly-discussed issue in the United States but not as common in Croatian society (see for example the discussions regularly posted on the website of the National Rifle Association; www.nra.org). Finally, we suspect that the "new technology and workplace" item from the Powerlessness factor and "store employees vs. shoplifters" were more commonly endorsed by the U.S. respondents because the changes in workplace technology are more readily observed in the United States than in Croatian companies whereas the shoplifting control is more strictly enforced in U.S. in comparison to Croatian stores.

In sum, we hypothesize that observed DIF on specific items was likely driven by the fact that a response considered to be consistent with justifying aggression, might also be the result of the respondent's developmental experiences (e.g., war for independence; active gun lobby and discussion of firearms civil liberties). However, our explanations are speculative and warrant further research.

What are the practical implications of our study? Do our results implicate that the Croatian version of the CRT-A should not be used for personality assessment? The best answer to these two questions is: further research is needed. Relatively large number of DIFs between the Croatian and U.S. samples leads to two major implications. First, we should avoid any cross-cultural comparison of these samples using the CRT-A because we could detect spurious cultural differences that are only the product of the measurement procedure, and fail to reveal true cultural differences that have been masked by measurement artifacts (Chen, 2008). Second, we should develop the local Croatian norms for interpretation of individual results on the CRT-A. However, it is important to note that the DIF analyses do not implicate conclusions about intragroup comparisons. Reasonably high internal consistency indices, item-total correlations and similar item diffi-

[4] It is important to note that the reason for DIF on the “American car"/ “Croatian fridges" item could be confounded with the fact that this item was adapted. Previous research has shown that culture-specific item substitutions generally result in DIF (Church et al., 2011).

culties for the Croatian adaptation of the CRT-A makes us optimistic about its usefulness for scientific and practical purposes (e.g., personnel selection). However, future construct- and criterion-related validity evidence is needed before we can draw definite conclusions about the usefulness of the Croatian adaptation of the CRT-A.

Limitations and future research

There are several limitations and related suggestions for future research. First, the focal sample in our study was from a specific country (i.e., Croatia) and DIF might not be observed if respondents came from another country, more similar to the United States. Although this might be true, the main point of our study is that the CRT-A problems are prone to DIF between cultures, and their measurement equivalence should be tested before cross-cultural comparisons are made, or results interpreted based on the original norms. Second, respondents in both the focal and referent samples were students, and perhaps different conclusions would be reached if we used more heterogeneous samples. Considering that previous research has shown that the CRT-A score is not related to gender, race, or participants' intellectual skills (James & LeBreton, 2012), nor is it easily faked in high-stakes testing situations (LeBreton et al., 2007), we believe that our results would likely replicate even if different samples (e.g., employees, general population) were compared.

Third, we explained the observed DIFs between Croatian and the U.S. samples with the effect of the cultural related knowledge. One might argue that the differences between the samples may be more easily explained with the differences in the gender structure across the samples. Although we did not have exact data about the gender structure in the U.S. sample and could not directly test the issue, we believe that that should not be the case because extensive resarch in the U. S. showed that implicit aggression, as measured with the CRT-A, is in no way related to gender. James and LeBreton (2012) report that in eight out of 10 published studies the correlation between the CRT-A score and gender was insignificant. The accumulated data lead them to conclude that "in all, a generally low and non-significant correlation between gender and scores on CR test for aggression is indicated" (p. 139). Additionally, James and LeBreton (2012) reported the results of the study that tested the DIFs due to gender on a large U.S. sample (n=2119) on the CRT-A. The results of that analysis revealed that only on one out of the 22 CRT-A a mild DIF due to gender was encountered. So, the U.S. based studies indicated that the construct measured with the CRT-A is in not related to gender. Although the interaction between gender and culture is a possibility, we opt for a more parsimonious explanation related to the differences between the cultures.

Finally, our analyses rest on the assumption that the factors structure of the CRT-A are in both our samples similar to those reported by James and LeBreton (2012). It could be argued that our conclusion of severity of DIFs could have been caused by the different factor structures between the two samples. We believe that this was not the case. First, if the dimensionality of the CRT-A was significantly different between the samples it should reflect in at least some non-uniform DIFs (i.e., constant across the trait levels).

However, in our study almost all DIFs were uniform (i.e., different between the trait levels). The non-uniform DIF was observed only on one item, and even in that case its effect was negligible. However, future research would benefit if the factor structure of the CRT-A is replicated on international samples before additional DIF studies on this promising system of personality assessment were conducted.

References

Bergman, S. M., McIntyre, M. D., & James, L. R. (2007). Identifying the aggressive personality. *Journal of Emotional Abuse*, *4*(3-4), 81-93. doi: 10.1300/J135v04n03_06

Berry, C. M., Sackett, P. R., & Tobares, V. (2010). A meta-analysis of Conditional reasoning test of aggression. *Personnel Psychology*, *63*(2), 361-384. doi:10.1111/j.1744-6570.2010.01173.x

Bing, M. N., Stewart, S. M., Davison, H. K., Green, P. D., McIntyre, M. D., & James, L. R. (2007). An integrative typology of personality assessment for aggression: Implications for predicting counterproductive workplace behavior. *Journal of Applied Psychology*, *92*(3), 722-744. doi:10.1037/0021-9010.92.3.722

Borsboom, D. (2006). When Does Measurement Invariance Matter? *Medical Care*, *44*(11), S176–S181. doi:10.2307/41219517

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of Differential Item Functioning in Translated Assessment Instruments. *Applied Psychological Measurement*, *19*(4), 309-321. doi:10.1177/014662169501900401

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005-1018. doi:10.1037/a0013193

Church T. A., Alvarez, J. M., Q, T., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology*, *101*(5), 1068-1089. doi:10.1037/a0025290

Cook, L. L., & Eignor, D. R. (1991). An NCME Instructional Module: IRT Equating Methods. *Educational Measurement: Issues and Practice*, *10*(3), 37-45.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*(1), 134-135. doi:10.1037/0033-2909.95.1.134

Drasgow, F*., &* Hulin, C. L. *(*1990*).* Item response theory. *In* M. D. Dunnette *&* L. M. Hough *(Eds.), Handbook of industrial & organizational psychology: Vol. 1 (*2nd ed*., pp.* 577-636*).* Palo Alto, CA*:* Consulting Psychologists Press*.*

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology*, *24*(2), 133-148. doi:10.1177/0022022193242001

Fox, J. (2007). Polycor: polychoric and polyserial correlations. *R package version 0.7-5, URL http://CRAN. R-project. org/package= polycor*.

Frost, B. C., Ko, C. H. E., & James, L. R. (2007). Implicit and explicit personality: A test of a channeling hypothesis for aggressive behavior. *Journal of Applied Psychology*, *92*(5), 1299-1319. doi:10.1037/0021-9010.92.5.1299

Galić, Z., & Plećaš, M. (2012). Quality of working life during the recession: The case of Croatia. *Croatian Economic Survey*, *14*(1), 5-41.

Galić, Z., Scherer, K. T., & LeBreton, J. M. (2014). Validity evidence for a Croatian Version of the Conditonal Reasoning Test for Aggression. Manuscript submitted for publication

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304-312. doi:10.1037/1040-3590.6.4.304

Gulliksen, H. (1950). *Theory of Mental Tests.* New York: Wiley

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science, 19*, 309-313. doi:10.1111/j.1467-9280.2008.02085.x

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology, 82*(6), 903 -918. doi: 10.1037/0022-3514.82.6.903.

Hofer, J., Chasiotis, A., Friedlmeier, W., Busch, H., & Campos, D. (2005). The Measurement of implicit motives in three cultures: power and affiliation in Cameroon, Costa Rica, and Germany. *Journal of Cross-Cultural Psychology*, *36*(6), 689-716. doi:10.1177/0022022105280510

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*(2), 192-218. doi:10.1177/0022022197282004

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, *16*(2), 131-152. doi:10.1177/0022002185016002001

International Testing Commission (2010) International Test Commission Guidelines for Translating and Adapting Test. Retrieved June 1st, 2014 from: http://www.intestcom.org/upload/sitefiles/40.pdf .

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, *1*(2), 131-163. doi: 10.1177/109442819812001

James, L. R., & LeBreton, J. M. (2010). Assessing aggression using conditional reasoning. *Current Directions in Psychological Science*, *19*(1), 30-35. doi:10.1177/0963721409359279

James, L. R., & LeBreton, J. M. (2012). *Assessing the Implicit Personality Through Conditional Reasoning* (1st ed.). Washington, D.C.: American Psychological Association.

James, L. R., & Mclntyre, M. D. (2000). *Conditional Reasoning Test of Aggression test manual.* Knoxville, TN: Innovative Assessment Technology.

James, L. R., McIntyre, M. D., Glisson, C. A., Green, P. D., Patton, T. W., LeBreton, J. M., Frost, B. C., et al. (2005). A conditional reasoning measure for aggression. *Organizational Research Methods*, *8*(1), 69-99. doi:10.1177/1094428104272182

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, *14*(4), 329-349. doi:10.1207/S15324818AME1404_2

Johnson, W., Spinath, F., Krueger, R. F., Angleitner, A., & Riemann, R. (2008). Personality in Germany and Minnesota: An IRT-Based Comparison of MPQ Self-Reports. *Journal of Personality*, *76*(3), 665-706. doi:10.1111/j.1467-6494.2008.00500.x

Kulas, J. T., Thompson, R. C., & Anderson, M. G. (2011). California Psychological Inventory dominance scale measurement equivalence: General population Normative and Indian, U.K., and U.S. managerial samples. *Educational and Psychological Measurement*, *71*(1), 245-257.

LeBreton, J. M., Barksdale, C. D., Robin, J., & James, L. R. (2007). Measurement issues associated with conditional reasoning tests: indirect measurement and test faking. *Journal of Applied Psychology*, *92*(1), 1-16. doi:10.1037/0021-9010.92.1.1

Lewis, M., & Lyall, S. (2012, August 24). Breivik Gets 21-Year Sentence in Norway for 77 Killings. *The New York Times*. Retrieved from: http://www.nytimes.com/2012/08/25/world/europe/anders-behring-breivik-murder-trial.html

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

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Inc.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847-862. doi:10.3758/BRM.42.3.847

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

McCrae, R. R., Costa, P. T., Pilar, G. H. D., Rolland, J.-P., & Parker, W. D. (1998). Cross-Cultural Assessment of the Five-Factor Model. The Revised NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *29*(1), 171-188. doi:10.1177/0022022198291009

McCrae, R. R., Terracciano, A. & 79 Members of the Personality Profiles of Cultures project (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology*, *89*(3), 407-425. doi:10.1037/0022-3514.89.3.407

National Rifle Association (n. d.) Retrieved from: http://home.nra.org

Nestić, D. (2002). Ekonomske nejednakosti u Hrvatskoj 1973-1998. (Economic Inequalities in Croatia 1978-1998), *Financijska teorija i praksa*, *26*(3), 595-613

Nye, C. D., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality*, *42*(6), 1524-1536. doi:10.1016/j.jrp.2008.07.004

Raju, N. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495-502. doi:10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207. doi: 10.1177/014662169001400208

Raju, N. S. & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 156-188). San Francisco, CA: Jossey-bass, Inc.

R Development Core Team (2010). R: *A language and environment for statistical computing*. R Retrieved from http://www.R-project.org.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4*(3), 207-230. doi:10.2307/1164671

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25.

Spiegel Online. (2012, June 22). Breivik and Defense Claim Sanity. Retrieved from http://www.spiegel.de/international/europe/anders-breivik-seeks-judgement-of-sanity-in-closing-statement-a-840452.html

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.

Tanner, M. (1997). *Croatia: A Nation Forged in War; Third Edition* (3rd ed.). Yale University Press.

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly, 50*, 538-612.

Vojnić, D. (1994) European integration processes and the countries in transition – with special reference to Croatia and former Yugoslavia, *Ekonomski pregled*, *9-10*, 203-239.

Appendix

Table A:
The conditional reasoning problem that underwent major change during the test adaptation process.

Original item in the CRT-A.

American cars have gotten better in the past 15 years. American car makers started to build better cars when they began to lose business to the Japanese. Many American buyers thought that foreign cars were better made.

Which of the following is the most logical conclusion based on the above?

a. America was the world's largest producer of airplanes 15 years ago. (IL)

b. Swedish car makers lost business in America 15 years ago. (IL)

c. The Japanese knew more than Americans about building good cars 15 years ago. (PA)

d. American car makers built cars to wear out 15 years ago, so they could make a lot of money selling parts. (AA)

Adapted item in the Croatian adaptation of the CRT-A.

Croatian fridges have gotten better in the past 5 years. Croatian fridge makers started to build better fridges when they began to lose business to the Slovenians. Many Croatian buyers thought that foreign fridges were better made.

Which of the following is the most logical conclusion based on the above?

a. Croatia was the world's largest producer of stoves 5 years ago. (IL)

b. Swedish fridge makers lost business in Croatia 5 years ago. (IL)

c. The Slovenians knew more than Croatians about building good fridges 5 years ago. (PA)

d. Croatian fridge makers built fridges to wear out 5 years ago, so they could make a lot of money selling parts. (AA)

Note: IL= illogical answer; PA = prosocial answer; AA =aggressive answer.