Psychological Topics 20 (2011), 3, 425-447

Original scientific paper – UDC – 159.942.072-057.875(497:48)

South and North: DIF Analyses of University-Student Responses to the Emotional Skills and Competence Questionnaire

Bo Molander, Stefan Holmström

Department of Psychology, Umeå University, Sweden

Vladimir Takšić

Department of Psychology, Faculty of Humanities and Social Sciences, University of Rijeka, Croatia

Abstract

In a study of the Emotional Skills and Competence Questionnaire instrument (ESCQ; Takšić, 1998) three samples of university students from Balkan countries (Croatia, Serbia, and Slovenia) were contrasted with two samples of university students from Nordic countries (Finland and Sweden). In total, 1978 students participated. Effects of country and gender were obtained from the ESCQ total scores, as well as from the subscale scores. The subsequent analyses of item bias, that is, differential item functioning (DIF), revealed a number of DIF items in pair wise comparisons of the samples, thus creating doubts about the fairness in comparing mean scores. Further analyses of the DIF items showed, however, that most of the item curve functions were uniform, and that effect sizes were low. It was also shown that the number of DIF items depended on which countries were compared. Spearman correlations between measures of number of DIF items and cultural values as measured by World Value Survey data were very high. Implications of these findings for future cross-cultural studies of the ESCQ instrument are discussed.

Keywords: Emotional Skills and Competence Questionnaire (ESCQ), cross-cultural differences, DIF analyses, Balkan and Nordic countries, World Value Survey

✉ Bo Molander, Department of Psychology, Umeå University, 901 87 Umeå, Sweden. E-mail: *bo.molander@psy.umu.se*.

Finnish, Serbian, and Slovenian data were provided by the courtesy of Hannu Rätu, University of Joensuu, Finland, Petar Čolović, University of Novi Sad, Serbia, and Andreja Avsec, University of Ljubljana, Slovenia. We are indebted to Andreas Stenling for assistance in the DIF-analyses work and to an anonymous reviewer who provided valuable comments on an earlier version of the paper.

A minor part of the study was presented at the *The 12th European Congress of Psychology*, Istanbul, Turkey, July 4-8, 2011.

Introduction

During the last 15 years, there has been an increasing interest in assessment procedures and attempts to create a modernization of test user standards. Examples of this movement are the Standards for Educational and Psychological Testing published by the American Educational Research Association (AERA, 1999) in USA and the European Federation of Psychology Associations' (EFPA) Standing Committee on Tests and Testing, this committee being established in 1999. Progress has been done in areas such as test use and qualification of test users, computer-based and internet-delivered testing, and criteria for test reviews (see Bartram, 2011, for a review of the EFPA work).

The reason for mentioning the recent development in the use of tests in the present paper is that the intended upgrading has consequences for testing in cross-cultural research. Guidelines for test translation and adaptation (Hambleton, 1994, 2005; van de Vijver & Hambleton, 1996; van de Vijver & Leung, 1997) were important publications influencing cross-cultural psychological research but also influencing the modernization of test standards in general.

The development and assessment of the Emotional Skills and Competence Questionnaire (ESCQ; Takšić, 1998) has followed the standards for tests and testing quite well, and in some respects even more rigorously (e.g., Faria et al., 2006; Takšić, 2011; Takšić, Mohorić, & Duran, 2009; Takšić, Tkalčić, & Brajković, 2001) than several other instruments used for cross-cultural evaluations. The results, so far, show good validity and reliability with respect to factor structure, as measured in different countries and in various samples. Recently, and as a timely response to the upgrading of test standardization, the ESCQ instrument has also been evaluated on item level, that is, item bias (Holmström, Molander, & Takšić, 2008; Molander, Holmström, & Takšić, 2009).

Although item bias is a long recognized problem (e.g., Mantel & Haenszel, 1959), surprisingly few cross-cultural studies in the past have focused on that aspect of assessment. In the 1990s, however, one type of item bias called *differential item functioning* (DIF) started to come into focus and methods for assessment were suggested (e.g., Camilli & Shepard, 1994; Smith, 2002; Swaminathan & Rogers, 1990; Zumbo, 1999). DIF was also discussed in connection with cross-cultural methods by, for example, van de Vijever and Leung (1997). From a very low number of papers assessing DIF in the 1990s there has been a large increase of such papers in the last five years, although the number of papers making assessments of DIF in the cross-cultural context is still modest and almost non-existing in the area of emotional intelligence. In addition to our own work, cited above, studies by Ekermans, Saklofske, Austin, and Stough (2011) and Gignac & Ekermans (2010) are rare exceptions. However, there are now suggestions to include item-bias analyses among the procedures necessary for

passing American and European standards for test review criteria (e.g., Geisinger, 2011; Nielsen, 2011).

In the Molander et al. (2009) paper we examined DIF in ESCQ data collected from Croatian, Slovenian, and Swedish students. The main results showed that a large proportion of the Total-Scores' items were flagged as potential DIF items; the highest proportion was found for the Croatian-Swedish comparison, the smallest for the Croatian-Slovenian comparison, the proportion flagged items in the Slovenian-Swedish comparison falling in between. These results suggest that there is a possible cultural effect in interpreting the meaning of the items among these three countries, even if errors in test translation from the Croatian original or methodological errors also could contribute. Another result to note was that analyses of how men and women in each country differed in responding to the items showed very few DIF-items. This finding validates the impression from the Faria et al. (2006) and Takšić et al. (2009) studies that ESCQ works well within each cultural/linguistic setting.

Although the Molander et al. (2009) study yielded important information for the future cross-cultural use of ESCQ, this study also raised several questions, for example: Is the utilized DIF method too sensitive? Were the samples comparable with respect to participant's qualifications? Could some other combination of countries make it easier to separate linguistic and cultural effects? Are such effects separable anyway? Does the pattern of DIF analyses of the three ESCQ scales differ from the DIF pattern of total scores with respect to differences between countries? How are differences in DIF-patterns between countries related to differences in mean scores? In the present study we tried to get more affirmative answers to some of these questions by adding Serbia and Finland to the former three countries. Thus, we will get a better opportunity for understanding the differences obtained previously between Croatia, Slovenia and Sweden.

As pointed out in the Molander et al. (2009) paper there is need of a theory in the area of emotional intelligence that could help predict the variation obtained in cross-cultural studies of the type that has been performed with the ESCQ instrument and that also could help explain the pattern of DIF differences between countries. Awaiting such a psychological theory we have made use of sociological/psychological data, as expressed in the World Value Survey (WVS; Inglehart & Welzel, 2005, 2010, 2011).

WVS covers all five countries included here, and gives a rough estimate of how cultural values differ between these countries. Inglehart and Welzel (2011) state that a large number of basic values is closely correlated and that two dimensions cover about 70% of the variance. These two dimensions are named "Traditional/Secular-rational" and "Survival/Self-expression". The first dimension makes a contrast between societies where religion and religious values are important and societies where these values are less important. Societies close to the traditional part of the dimension follow traditional family values and nationalistic

values, favouring authority and rejecting abortion and divorce. Societies on the secular-rational part of the dimension show the opposite opinions. The second dimension separates societies which emphasize economic and physical security from societies which emphasize subjective well-being, self-expression, and quality of life. These latter societies also show larger tolerance of out-groups, interpersonal trust, and emphasis on individual freedom.

No doubt there are important similarities in scope and meaning between these sociologically and psychologically based dimensions and the psychological dimensions of the ESCQ instrument. An important bridge between ESCQ and WVS is to be found in the emphasis in ESCQ on social relationships and how they are valued. Other such bridges are exemplified below in the discussion section. We believe that the WVS-measure might work acceptably for ESCQ as a measure of cultural proximity.

The WVS project provides the World Value Survey Cultural Maps, the last version based on data from 2005-2008 (Ingelhart & Welzel, 2010, 2011). From that map and the scale values based on nation-level mean scores (see Ingelhart & Welzel, 2005) we note the following order of scale values for the Traditional/Secular-rational dimension: Croatia (0.08), Serbia (0.35), Slovenia (0.73), Finland (0.82), and Sweden (1.86). For the Survival/Self-expression dimension the order is: Serbia (-.62), Croatia (.31), Slovenia (.36), Finland (1.12), and Sweden (2.35). Dimension scales vary from -2.0 to +2.5. Values of Croatia were obtained from wave 4 (2000), all others from wave 5 (2006).

From these data it seems reasonable to expect Serbia and Croatia together in one cluster, Finland and Sweden in one cluster, and Slovenia somewhere in between. Language-wise there are four groups: Serbia and Croatia again in one group, in the other three groups Slovenia, Finland, and Sweden are alone. Interestingly, Slovenian language differs little from Serbian and Croatian languages, whereas Finnish differs quite a lot from the Swedish language, being part of another language family. Thus, Swedish (belonging to Indo-European languages) could be said to be closer to the Slavic languages of Serbia, Croatia, and Slovenia than to Finnish (belonging to Finnic-Uralic languages).

Differences between Finland and Sweden thus are differences in the way items are expressed linguistically, we believe, more than cultural differences even if Finland in the World Value Survey cultural map actually is situated at some distance away from Sweden and in the direction of Slovenia. Finally, Finnish and Swedish samples should both deviate from the three other samples, possibly with the Slovenian sample somewhat closer than the Serbian and Croatian samples. These expectations are based on the assumption that DIF items reveal translation variation and/or cultural variation. Methodological differences during data collection could also cause DIF items, but we have reason to believe that for these five countries the method has been quite similarly applied. It is important to point out that in this paper we are not concerned with individual variation in emotional

intelligence, neither the variation in emotional intelligence between countries. We are concerned with the assessment of the ESCQ instrument for use in cross-cultural psychological studies. To summarize, this study is (1) examining if DIF patterns found in the Molander et al. (2009) study are replicated in the present slightly changed and extended group of samples of university students; (2) illuminating the relative importance for DIF of linguistic and cultural variation in the samples; and (3) examining the relationships among Word Value Survey basic cultural dimensions, ESCQ scores, and DIF measures.

Method

Participants

Participants were 1978 university students from University of Rijeka, Croatia; University of Joensuu, Finland; University of Novi Sad, Serbia; University of Ljubljana, Slovenia; and University of Umeå, Sweden. Data in the Serbian sample were collected (2009) in students from the Philosophical Faculty. The Finnish data were collected (2004) in students from Humanities, Education, Psychology, and Natural Sciences and the Swedish data were collected (2005) in students from Psychology and Education. The collection of the data in the Croatian and Slovenian samples started in 2001, and 2005, respectively, with new subjects being added over time. Students are mostly from Humanities and Social Sciences in these two latter samples. The mean ages of the Swedish and Finnish samples are somewhat higher (approximately 25 years of age) as compared to the other three samples (approximately 22 years of age). The Croatian and Slovenian samples are larger in number of participants as compared to the Molander et al. (2009) study, while the Swedish sample is smaller. All samples were checked to ensure that the data files comprised university students and no other category of participants. It was discovered that the former Swedish sample comprised a number of students with the expected age of a university student but with unclear university student status. These students were discarded in the present sample. See Table 1 for sample characteristics.

Table 1. *Number and Proportions of Participating Men and Women, and Estimated Mean Age in the Croatian, Finnish, Serbian, Slovenian, and Swedish Samples*

	Men *N* (proportion)	Women *N* (proportion)	Total sample	Mean Age
Croatia	236 (.31)	522 (.69)	758	≈22
Finland	55 (.24)	173 (.76)	228	≈25
Serbia	64 (.28)	162 (.72)	226	≈22
Slovenia	169 (.31)	385 (.69)	554	≈22
Sweden	79 (.37)	133 (.63)	212	≈25

Instrument

The *Emotional Skills and Competence Questionnaire* (ESCQ) was developed by Takšić (1998) as a self-report version of the Mayer and Salovey ability-based instrument for the measurement of emotional intelligence (1996). The questionnaire consists of a total of 45 items including three subscales: *Perceive and Understand* (15 items); *Express and Label* (14 items); and *Manage and Regulate* (16 items). Responses to items are given in accordance with a five-point scale: Never (1), Seldom (2), Occasionally (3), Usually (4), Always (5). Total scores are formed by adding the values rated for all items. Subscale scores are formed in a similar way. Translation of the instrument is routinely performed from an English version of the original Croatian instrument according to the back translation procedure, and by communication with the author of the instrument. Sometimes, as in the case of the Slovenian version, the instrument is translated directly from the Croatian original and items were checked also by people having Croatian as their native language, or were highly skilled in the language. For further details about the development of the instrument and the present cross-cultural distribution, see Takšić et al. (2009).

Procedure

The ESCQ was administered to students in classes or individually during regular academic hours. Participation was voluntary and no financial reward was given. Before the start of the testing, the participants were introduced in general terms to the purpose of the study, and the informed consent was obtained. Instructions were given about how to use the scale of the instrument. The questionnaire took about 20 minutes to finish.

Statistical analysis

Internal reliabilities were determined by Cronbach's alpha for total scores, as well as for subscale scores. Acceptable level for alpha was set to .70 (Nunnally & Bernstein, 1994). Differences in alfa were tested by the statistic $(1-\alpha_1)/(1-\alpha_2)$ with N_1-1 and N_2-1 degrees of freedom. This statistic follows the F-distribution (see e.g., Fischer & van de Vijver, 2010). Differences in mean scores were analyzed by analyses of variance (ANOVAs) with Bonferroni adjustment applied in post-hoc tests. Level of significance was set to .01 due to the large sample sizes. Effect sizes are shown by partial eta squared (η^2). Rank correlations were calculated by Spearman's rho.

DIF analyses in this paper follow the Zumbo (1999) method and recommendations. This method is based on the ordinal logistic regression equation

$$y = b_0 + b_1\text{TOTAL} + b_2\text{GRP} + b_3\text{TOTAL x GRP}_i + \varepsilon_i$$

where TOTAL stands for total scores and GRP stands for group. In the equation ε_i is distributed with mean zero and variance $\pi^2/3$. Scripts for calculating DIF in SPSS are provided by Zumbo (1999). Effect size criteria for DIF items were set according to Jodoin and Gierl (2001) recommendations. DIF analyses are performed by pair wise comparisons of all samples, but also with the Croatian sample as a reference group for some of the analyses. All statistical calculations except the calculations of alpha differences were performed on PASW Statistics v.18.

Results

Internal consistencies (Cronbach's alpha) for total scores, as well as for each of the three subscales are presented in Table 2. Differences in reliability between the Croatian sample and each of the other samples were tested by means of the statistic $(1-\alpha_1)/(1-\alpha_2)$. Significant differences in Table 2 are marked with *.

Table 2. *Cronbach's alpha for Total Scores and Subscale Scores*

ESCQ	Croatia	Finland	Serbia	Slovenia	Sweden
Total	.88	.86	.87	.91*	.85
PU	.84	.88*	.84	.90*	.84
EL	.81	.75*	.84	.85*	.80
MR	.69	.56*	.72	.74	.59*

Note. ESCQ = Emotional Skills and Competence Questionnaire; PU = Perceive and Understand (15 items); EL = Express and Label (14 items); MR = Manage and Regulate (16 items).

$^*p<.01$.

Table 2 shows high alpha values over all samples, especially for total scores and PU-scale scores. The values for the EL scale are somewhat lower, although quite acceptable. The MR-scale values are lower than the Nunnally & Bernstein (1994) criterion for three of the samples. Low values for the MR scale have been found in most of the other ESCQ samples and is probably due to the more complex ability this scale covers than the other two scales (cf. Mayer & Salovey, 1996). It seems as an established pattern that the PU scale has the highest internal reliability and the MR scale the lowest, with the EL scale in between (see Takšić et al., 2009; for more examples of how the three scales vary according to this pattern). What is interesting in the present context is that the reduction of the size of the alpha coefficient from PU to MR is especially strong in the Finnish and Swedish samples, suggesting perhaps that the MR scale, in addition to multidimensional content, contains items that could be difficult to translate from the Croatian original.

From a cross-cultural perspective it is, of course, important to study the size of the variation among the samples with respect to mean scores, including differences

in mean scores between men and women. In particular, the difference in effect size between effects of country and gender is of interest here. Means are presented in Table 3.

Table 3. *Means and Standard Deviations of ESCQ Scores*

		Croatia		Finland		Serbia		Slovenia		Sweden	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ESCQ	All	160.4	16.29	160.4	12.78	168.2	16.24	162.8	17.69	161.0	12.60
	Men	158.0	16.59	161.9	11.94	163.0	17.23	159.2	17.38	159.1	13.25
	Women	161.4	15.94	160.1	13.08	170.2	15.42	164.1	17.54	162.1	12.15
PU	All	53.4	6.96	52.7	6.34	56.2	7.06	56.5	8.15	52.2	5.81
	Men	51.9	7.11	52.8	6.61	54.3	7.52	54.5	8.28	51.1	5.75
	Women	54.0	6.81	52.7	6.30	56.9	6.75	57.2	7.56	52.8	5.79
EL	All	48.6	7.22	50.9	5.60	51.2	7.96	47.8	7.69	51.0	6.28
	Men	47.9	6.54	51.0	5.50	49.6	8.13	45.6	7.79	50.5	6.26
	Women	48.8	7.45	50.9	5.64	51.9	7.82	48.6	7.41	51.2	6.33
MR	All	58.5	5.99	56.8	4.52	60.7	6.68	58.5	6.35	57.8	4.71
	Men	58.1	6.20	58.1	4.52	59.1	7.67	59.0	6.47	57.5	5.09
	Women	58.6	5.87	56.5	4.47	61.3	6.16	58.2	6.30	58.0	4.50

Note. ESCQ = Emotional Skills and Competence Questionnaire; PU = Perceive and Understand (15 items); EL = Express and Label (14 items); MR = Manage and Regulate (16 items).

Analyses of variance were performed on total scores and the scores of each subscale. A 5(Country) x 2(Sex) ANOVA on total scores showed significant effects of Country, $F(4,1951)=7.02$, $p<.001$, $\eta^2=.014$, and of Sex, $F(1,1951)=13.63$, $p<.001$, $\eta^2=.007$. There was no significant Country x Sex interaction, $\eta^2=.004$. As shown by the Bonferroni post hoc tests the effect of Country was due to significant differences between the Serbian sample and all other samples, $p<.001$. The effect of Sex replicated the common finding that women generally score higher than men on the ESCQ (e.g., Faria et al., 2006; Takšić et al., 2009).

For the PU subscale the analysis showed significant effects of Country, $F(4,1951)=19.17$, $p<.001$, $\eta^2=.038$, and of Sex, $F(1,1951)=20.07$, $p<.001$, $\eta^2=.010$. Effect size for the non-significant Country x Sex interaction was $\eta^2=.003$. The effect of Country was somewhat more complex than in the case of total scores, as Croatian, Finnish and Swedish samples differed from Slovenian and Serbian samples, all these differences with $p<.001$. The first three samples did not differ, neither did the latter two. Analysis of the EL scale showed effects of Country, $F(4,1951)=18.47$, $p<.001$, $\eta^2=.040$, and of Sex, $F(1,1951)=11.22$, $p<.001$, $\eta^2=.006$. The Country x Sex interaction was not significant, $\eta^2=.005$. The effect of Country was due to significant differences between the Croatian sample, and the Serbian, Finnish and Swedish samples. The Serbian, Finnish, and Swedish samples differed also from the Slovenian sample ($p<.001$). Finally, the analysis of the MR scale showed only a significant effect of Country, $F(4,1951)=6.54$, $p<.001$, $\eta^2=.013$, and effect sizes of Sex and Country x Sex were $\eta^2<.001$, and $\eta^2=.007$, respectively. The

effect of Country was due to differences between the Serbian sample and all other samples ($p<.001$). In addition, the Croatian sample differed from the Finnish sample ($p<.01$), this latter sample also differing from the Slovenian sample ($p<.01$).

From a cross-cultural point of view the relatively simple outcome of the analysis of total scores quickly becomes quite complicated when the results of subscale analyses are considered. At present there is no cross-cultural theory in the emotional intelligence field which could explain the obtained variation. However, the subscale scores seem to be more informative than the total scores, as the latter are a sum of the separate scales. It is quite possible that proficiency in the EI dimensions may vary within a specific cultural context. Total scores may hide such a variation. Emphasis in previous studies of the cross-cultural validation and psychometric description of the ESCQ instrument has been on total scores more than the scores of the subscales. Maybe it is time to look more systematically at the cross-cultural variation of the subscales. Such an undertaking may very well contribute to an increased understanding of what a cross-cultural theory of emotional intelligence might look like.

The above analyses of means make it necessary to examine also the items in the scales from a cross-cultural perspective. Means as revealing differences or similarities between groups and cultures presuppose that the items which make up the means are understood similarly in the groups that are compared (e.g., Fischer & van de Vijver, 2010; van de Vijver & Leung, 1997).

The logical regression method used in this paper to analyze if the items have been understood similarly in different samples is suggested by Zumbo (1999) and is based on previous work by e.g., Swaninathan & Rogers (1990). There are by now quite many methods available for calculating DIF (see, e.g., Roussos & Stout, 2004). The choice of the Zumbo method was due to the facts that 1) the method handles binary data as well as ordinal data; 2) the method is not dependent on overly big sample size; 3) the method has been evaluated and compared with other methods (cf. Kristjansson, Aylesworth, McDowell, & Zumbo, 2005); and 4) the method involves a measure of effect size. According to Zumbo *differential item functioning* (DIF):

"*…occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure…*" (Zumbo, 1999, p. 12).

This definition should not be mixed with the definition of *item bias*, which "*…occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not sufficient, for item bias…*" (Zumbo, 1999, p. 12).

Thus, if the analysis shows no DIF in an item, there is no item bias, but if DIF is shown it is not a sufficient evidence of item bias. In such a case further

investigations have to be performed to decide what type of problem this item may have.

As mentioned in the method section, scripts for SPSS (PASW Statistics) are provided by Zumbo (1999), where details of the testing procedure also can be found. These tests involve level of significance for chi square and a measure of effect size. To be able to compare our results with the Molander et al. (2009) study we use .01 as level of significance and the Jodoin & Gierl (2001) criteria for effect size. Thus, DIF is negligible for effect-size values below .035, moderate between .035 and .070, and large for levels above .070. However, it should be noted that Zumbo (1999) recommends the use of a measure developed by Zumbo and Thomas (1997) with the criterion of at least 0.13 for an item considered to be DIF. This criterion is a parallel to effect-size criteria for other statistics (Cohen, 1992) and is also used in assessing personality items (e.g., Pope, 1997). As the consensus about effect-size criterion for DIF still does not seem to be too strong (e.g., French & Maller, 2007; Kristjansson et al., 2005), we will in this study report also the Zumbo-Thomas criterion.

Tables 4-7 depict the results of DIF analyses performed on total scores of the ESCQ and of each subscale. All samples are compared with each other in a pair wise fashion, as we consider all relationships to be of interest for better understanding of the cultural and/or linguistic variation. However, in some of the calculations performed below, we have used the procedure of reference group and focal group, with Croatia as a reference group.

The tables illustrate that the frequency of potential DIF items varies with countries compared. For total scores there is such a pattern that the least number of DIF items occurred for the comparison of the Croatian and Serbian samples and for the comparison of the Swedish and Finnish samples. Somewhat higher frequency occurred for comparisons with the Slovenian sample, less for comparisons with the samples of Croatia and Serbia than for comparisons with the samples of Finland and Sweden. The highest frequency was obtained for Croatian and Serbian comparisons with Finnish and Swedish samples. The low number of DIF items for the Croatian-Serbian-Slovenian comparisons and for the Finnish-Swedish comparisons are quite a good outcome, considering the very liberal Jodoin and Gierl (2001) criteria for effect size that were applied. The more conservative Zumbo-Thomas measure indicated very few DIF items.

Similar sample patterns were found for the three subscales and the PU scale showed fewer DIF items than the EL scale, the latter scale showing fewer DIF items than the MR scale. At the most the proportion flagged DIF items was .33 for total scores (Croatia-Finland), .27 for the PU scale (Croatia-Sweden), .29 for the EL scale (Serbia-Sweden), and .44 for the MR scale (Finland-Serbia).

Table 4. *Number of DIF Items for Total Scores in Each Pair of Samples*

	Croatia	Finland	Serbia	Slovenia	Sweden
Croatia	-	*2*	*0*	*0*	*1*
Finland	**15**	-	*3*	*1*	*1*
Serbia	**2**	**13**	-	*0*	*1*
Slovenia	**5**	**10**	**1**	-	*0*
Sweden	**14**	**4**	**13**	**9**	-

Note. Numbers marked in bold: Jodoin & Gierl (2001) effect-size criterion (>.035); Numbers marked in italics: Zumbo & Thomas (1997) effect-size criterion (>.13). Number of total items: 45.

Table 5. *Number of DIF Items for the Perceive and Understand (PU) Scale*

	Croatia	Finland	Serbia	Slovenia	Sweden
Croatia	-	*0*	*0*	*0*	*0*
Finland	**3**	-	*0*	*0*	*0*
Serbia	**1**	**1**	-	*0*	*1*
Slovenia	**1**	**2**	**2**	-	*0*
Sweden	**4**	**1**	**3**	**3**	-

Note. Numbers marked in bold: Jodoin & Gierl (2001) effect-size criterion (>.035); Numbers marked in italics: Zumbo & Thomas (1997) effect-size criterion (>.13). Number of total items: 15.

Table 6. *Number of DIF Items for the Express and Label (EL) Scale*

	Croatia	Finland	Serbia	Slovenia	Sweden
Croatia	-	*0*	*0*	*0*	*0*
Finland	**2**	-	*1*	*0*	*0*
Serbia	**1**	**3**	-	*0*	*1*
Slovenia	**2**	**3**	**1**	-	*0*
Sweden	**2**	**1**	**4**	**2**	-

Note. Numbers marked in bold: Jodoin & Gierl (2001) effect-size criterion (>.035); Numbers marked in italics: Zumbo & Thomas (1997) effect-size criterion (>.13). Number of total items: 14.

Table 7. *Number of DIF Items for the Manage and Regulate (MR) Scale*

	Croatia	Finland	Serbia	Slovenia	Sweden
Croatia	-	*1*	*0*	*0*	*1*
Finland	**5**	-	*1*	*1*	*1*
Serbia	**1**	**7**	-	*0*	*1*
Slovenia	**2**	**4**	**0**	-	*0*
Sweden	**5**	**2**	**5**	**4**	-

Note. Numbers marked in bold: Jodoin & Gierl (2001) effect-size criterion (>.035); Numbers marked in italics: Zumbo & Thomas (1997) effect-size criterion (>.13). Number of total items: 16.

Spearman's rho was applied to correlations between DIF measures and the ESCQ means for total scores and subscale scores, as well as the WVS measures of the two basic value dimensions presented above in the introductory section. Two DIF measures were formed: 1) DIFallTS, which is the sum of DIF items in total scores for each sample's comparison with all other samples. From Table 4 this gives Croatia (36), Finland (42), Serbia (29), Slovenia (25), and Sweden (40); 2) DIFcroTS, which, with Table 4 and Croatia set to (0) as a reference sample, gives Finland (15), Serbia (2), Slovenia (5), and Sweden (14). Similar DIF measures were formed for subscale scores. In Table 8 measures of the two WVS map dimensions are correlated with mean scores of the ESCQ instrument and DIF measures with Croatia as a reference sample.

Table 8. *Spearman Correlations: Measures of WVS Dimensions, ESCQ Mean Scores, and Measures of DIF for Total Scores and Subscale Scores*

	Mean TS	Mean PU	Mean EL	Mean MR	DIFcroTS	DIFcroPU	DIFcroEL	DIFcroMR
WVS1	-.05	-.60	.20	-.72	.90*	.98**	.89	.98**
WVS2	-.41	-.70	-.10	-.87	.80	.82	.78	.87

Note. WVS1 = Traditional/Secular; WVS2 = Survival/Self-expression.
*$p<.05$, **$p<.01$.

Table 8 shows that DIF measures correlate very highly with the value dimensions, somewhat higher with the Traditional/Secular dimension of the WVS map than with the Survival/Self-expression dimension. We also used the DIFall version measures in these correlations and found that, in general, the coefficients were lower with those measures. The DIFallTS measure correlated .50 and .60 with WVS1 and WVS2, respectively. Similar and lower levels were found for the subscale correlations. It is obvious from Table 8 that mean total scores and mean subscale scores yield lower coefficients than the DIFcro measures and that the mean PU and MR subscale measures are more sensitive to WVS values than the mean total scores measure. However, some caution is recommended in reading Table 8, as the number of samples used for the rank order correlations is small and quite high correlations are required for stable results.

DIF analyses were also performed on total scores items on gender differences between and within country samples. It should be pointed out, however, that these analyses may be less reliable than the analyses of total score items and subscale items, as the samples are smaller than recommended for DIF analyses. This is especially the case for the Finnish, Serbian, and Swedish samples. The results are presented in Tables 9-10.

Table 9. *Number of DIF Items for Men/Women Based on Total Scores (Pair Wise Samples)*

	Croatia	Finland	Serbia	Slovenia	Sweden
Croatia	-	*4/2*	*1/0*	*3/0*	*3/1*
Finland	**16/14**	-	*3/3*	*1/1*	*1/1*
Serbia	**7/1**	**18/15**	-	*0/0*	*3/2*
Slovenia	**9/6**	**11/10**	**2/2**	-	*1/1*
Sweden	**21/11**	**12/7**	**17/16**	**16/10**	-

Note. Numbers marked in bold: Jodoin & Gierl (2001) effect-size criterion (>.035); Numbers marked in italics: Zumbo & Thomas (1997) effect-size criterion (>.13). Number of total items: 45.

Table 10. *Number of DIF Items for Men vs. Women Based on Total Scores (Within Samples)*

Croatia	Finland	Serbia	Slovenia	Sweden
6	**2**	**4**	**0**	**6**
0	*0*	*0*	*0*	*0*

Note. Numbers marked in bold: Jodoin & Gierl (2001) effect-size criterion (>.035); Numbers marked in italics: Zumbo & Thomas (1997) effect-size criterion (>.13). Number of total items: 45.

Only the values based on the Jodoin and Gierl (2001) effect-size criterion show interpretable trends here. From the two tables it seems that the number of DIF items is larger than what is obtained when men and women are not separated. The trend is also that men have more DIF items than women and that the pattern over countries is similar to what was found for the overall DIF analysis of the total scores, and that gender effects within samples are quite small.

Item comparisons with Croatia as a reference sample are presented in Table 11 together with item formulations. This table comprises the items with the five highest effect values for each pair wise comparison. In Table 11 DIF items are presented by item number.

We note from the content of Table 11 that one item (11) is present in all comparisons with Croatia, and that one item (14) is present in three out of four comparisons. Further, it can be seen that the comparisons Croatia-Serbia and Croatia-Slovenia are quite similar (11, 14, 15, 16), and more similar than the comparisons of Croatia-Finland and Croatia-Sweden, where two items (7, 11) occur in both comparisons. There is no obvious pattern for these "five highest" in how DIFs are distributed among the subscales, more than that items from all three scales are flagged as DIF.

Table 11. *Five Highest DIF Effect Values for Total Scores with Croatian Sample as a Reference*

Croatia-Serbia	Croatia-Slovenia	Croatia-Finland	Croatia-Sweden
14* When I see how someone feels, I usually know what has happened to him (PU)	11* I study and learn best, when I am in a good mood and happy (MR)	10* When I am with a person who thinks highly of me, I am careful about how I behave (MR)	7* When I don't like a person, I find ways to let him/her know (MR)
11* I study and learn best, when I am in a good mood and happy (MR)	16* I can easily think of a way to approach a person I like (EL)	7* When I don't like a person, I find ways to let him/her know (MR)	41* I have found it easy to display fondness for a person of the opposite sex (EL)
15 I am able to tell the difference if my friend is sad or disappointed (PU)	15* I am able to tell the difference if my friend is sad or disappointed (PU)	11* I study and learn best, when I am in a good mood and happy (MR)	25* If I observe a person in the presence of others, I can determine precisely his/her emotions (PU)
16 I can easily think of a way to approach a person I like (EL)	39* I notice when somebody feels down (PU)	22* I can recognize most of my feelings (EL)	13* When I meet an acquaintance, I immediately notice his/her mood (PU)
24 I can say that I know a lot about my emotional state (EL)	14* When I see how someone feels, I usually know what has happened to him (PU)	14* When I see how someone feels, I usually know what has happened to him (PU)	11* I study and learn best, when I am in a good mood and happy (MR)

Note. PU = Perceive and Understand (15 items); EL = Express and Label (14 items); MR = Manage and Regulate (16 items); *effect size >.035.

Table 12. *DIF-Items for Total Scores in all Sample Comparisons*

	Croatia	Finland	Serbia	Slovenia	Sweden
Croatia	-	*7, 10*	--	--	*7*
Finland	**1, 2, 6, 7, 10, 11, 13, 14, 17, 20, 22, 23, 37, 40, 43**	-	*7, 10, 24*	*10*	*10*
Serbia	**11, 14**	**1, 2, 5, 6, 7, 8, 10, 13, 16, 17, 22, 24, 43**	-	--	*7*
Slovenia	**11, 14, 15, 16, 39**	**1, 5, 6, 7, 10, 15, 16, 23, 40, 43**	**24**	-	--
Sweden	**1, 2, 4, 7, 10, 11, 13, 14, 17, 19, 22, 25, 35, 41**	**10, 18, 25, 41**	**1, 2, 3, 4, 7, 10, 13, 15, 16, 17, 24, 25, 41**	**1, 4, 7, 15, 16, 19, 25, 35, 41**	-

Note. Items marked in bold: effect size >.035. Items marked in italics: effect size >.13. Number of items: 45.

Inspection of Table 12 reveals that in the ten comparisons there were 23 items from the PU scale, 31 items from the EL scale, and 32 items from the MR scale. In total, 28 different items (63%) were flagged one or more times as DIF, as assessed by the Jodoin & Gierl (2001) effect-size criterion. Of these 28 items 1, 7, and 10 occurred in six comparisons, item 16 five times, and items 2, 11, 13, 14, 15, 25 and, 41 four times. All others occurred three times or less. Thus, most of the DIF items seem to be connected to a specific sample or are rather unique for a specific sample comparison. As a contrast only 3 items (7%) were flagged as DIF according to the Zumbo-Thomas (1997) criterion.

It is possible in the Zumbo (1999) logistic regression method to get information about the form of a DIF item, that is, if the item is uniform or non-uniform. If DIF is uniform the curves describing item characteristics (ICC) for two groups are parallel. If DIF is non-uniform the two curves will cross. If the two curves are identical or close to identical there is no DIF (see Zumbo, 1999, pp. 15-21). Here we have applied a somewhat simpler method (by ANOVA) described by van de Vijver and Leung (1997) to produce curves. In the graphs below (Figures 1-4) curves are shown for item 11 for each of the comparisons with Croatia, as presented in Table 11.

Figure 1. *Item 11 Mean Score for Croatian and Finnish Samples as a Function of Total Score Level (p<.001; Effect Size:* η^2*=.072)*

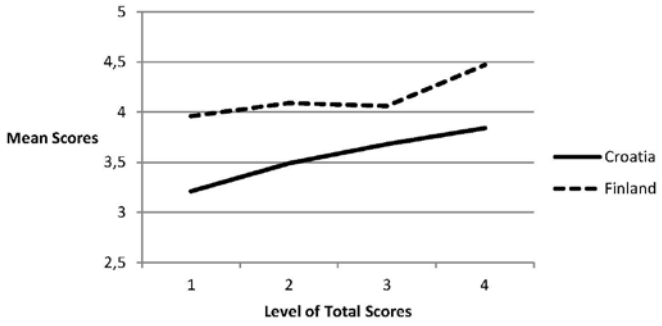


Figure 2. *Item 11 Mean Score for Croatian and Serbian Samples as a Function of Total Score Level (p<.001; Effect Size:* η^2*=.034)*

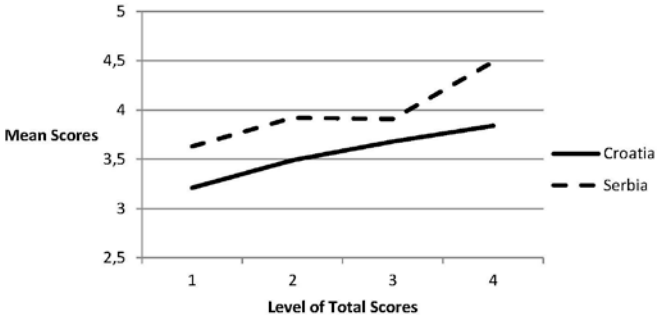


Figure 3. *Item 11 Mean Score for Croatian and Slovenian Samples as a Function of Total Score Level (p<.001; Effect Size:* η^2*=.086)*

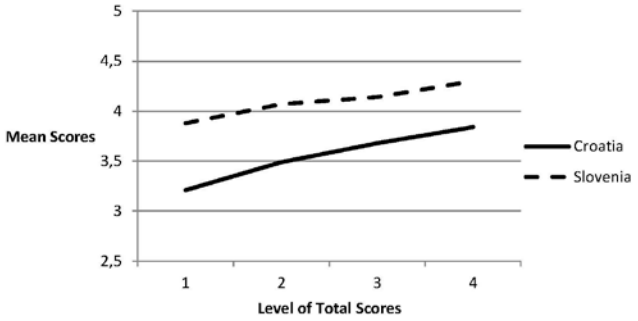
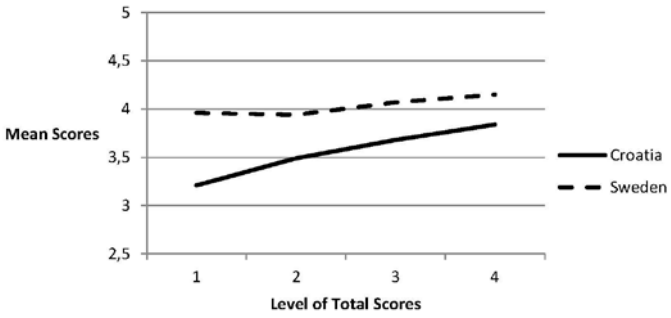


Figure 4. *Item 11 Mean Score for Croatian and Swedish Samples as a Function of Total Score Level (p<.001; Effect Size:* η^2*=.045)*



All four figures indicate significant uniform functions. There are no significant interactions between sample and score level (p>.05; all η^2<.007), which would indicate non-uniform relationships. It should be noted that comparisons, where the three samples with the smaller number of participants are involved, that is, Serbia, Finland, and Sweden, show more deviations from parallel curves than when larger samples, as in the Croatian and Slovenian comparison, are involved. We examined all items listed in the second column of Table 12 with respect to uniform and non-uniform functions, using Croatia as a reference sample. All the 36 items showed uniform functions, the majority with moderate effect sizes (between >.035 and <.070). Only four of the items (two in the Croatian-Finnish comparison and two in the Croatian-Swedish comparison) had effect sizes considered large. There were also six items that showed significant non-uniform functions or close to significance in addition to the uniform functions. However, all of those had very low and negligible effect sizes.

Discussion

The results based on mean total scores and mean scores of the subscales in this study are overall quite similar to the results reported by Molander et al. (2009), where Croatia, Slovenia, and Sweden was compared. There are a few notable discrepancies, though. The addition of Serbia and Finland gave rise to a significant effect of Country on the total score measure, showing the Serbian sample to be significantly higher than all other samples, among which there was no significant difference. In the former study there was no effect of Country. In both studies there were effects of Sex, women having higher scores than men, but no Country x Sex interactions. The results for the PU and EL subscales were identical, that is, effects of Country and Sex, but no Country x Sex interactions. However, the MR scale yielded somewhat unexpected results, as this scale did not give any significant effects in the former study. Here, there was an effect of Country, the Serbian sample scoring higher than the other samples, and the Croatian sample scoring higher than the Finnish sample. The conclusions from these comparisons are, firstly, that although the Croatian, Slovenian, and Swedish samples were changed in size and proportions of male and female participants from the 2009 study to the present study the effects were almost identical, and the relationships between Croatian, Slovenian and Swedish samples for total scores and subscale scores were the same. Thus, the ESCQ score measures seem to be robust with respect to Country and Sex effects. The second conclusion is that the Serbian sample scores higher on the ESCQ instrument than the other samples in the study, which brings up the question if this proficiency is culturally based or just part of sampling variation. The cultural alternative will be examined further below. It can be noted anyway, that the women in the Serbian sample are scoring particularly high, although not high enough for creating a Country x Sex effect. This study conformed to the common finding from other cross-cultural studies (e.g., Schwartz & Rubel, 2005) that the effects of culture are larger than the effects of gender, as measured here by the partial eta squared measure.

The analyses of the Cronbach alphas, as shown in Table 2, demonstrate that the internal consistency in the ESCQ instrument is very high and in this study particularly high in the Slovenian sample. What Table 2 also shows is that it is possible to pass the criterion of .70 in the MR scale, as illustrated by both the Slovenian and Serbian samples. Why most samples don't reach this criterion in the MR scale is still not perfectly clear, but both the Serbian and Slovenian samples, as well as the Finnish and Swedish samples, which yielded quite low coefficients, could help throw light on this issue after further analyses of the correlation patterns. A cultural effect rather than a linguistic effect seems likely here. As we suggested above, some items in the scale are complex and may lead to some ambiguity. The degree of ambiguity might vary between samples. An answer may not come easily, however, and we agree with Loewenthal (1996), who said concerning the definition of Cronbach's alpha that "I have always found this coefficient delightfully

metaphysical" (p. 10). Note also, that the alpha coefficient is defined in such a way that it can be larger than 1.0, a fact that is seldom pointed out in the literature.

The question of cultural impact on the scores becomes more critical when we consider the results of the DIF analyses. As in the Molander et al. (2009) study there were many items flagged as DIF when the Jodoin & Gierl (2001) criteria for effect size were applied. In the former study the proportion of DIF items for total scores was .62 for the Croatian-Swedish comparison, and .40 and .27 for the Slovenian-Swedish and Croatian-Slovenian comparisons, respectively. The corresponding values in the present study are .31, .20, and .11. Thus, there are much lower DIF proportions in this study and that goes also for subscales and gender. The discrepancy between the two studies demonstrates the sensitivity of the present DIF procedure with respect to changes in the samples, such as the number of participants and the category of participants. It is generally recommended that DIF analyses should be performed on large samples (e.g., French & Maller, 2007; Kristjansson et al., 2005). The logistic regression procedure we used here is relatively robust in smaller samples though, but sample sizes of at least 200 are still recommended (Zumbo, 1999). The DIF analyses on gender shown in Table 8 are likely not to be reliable then, except for the comparisons between Croatia and Slovenia. Another important issue is the choice of effect-size criteria for DIF. The two alternatives used in the present study represent somewhat extreme points on a continuum. The literature (e.g., French & Maller, 2007) seems at the moment to prefer the liberal criterion of Jodoin & Gierl (2001) rather than the conservative Zumbo-Thomas criterion (Zumbo & Thomas, 1997). At the same time authors argue for collecting a large base of DIF items to increase the possibility of selecting the important ones (e.g., culturally/linguistically based) from those that are due to irrelevant factors (e.g., inappropriate item due to offensive language). After examining such a large database it may be possible to suggest better criteria for DIF, it is believed.

After the identification of potential DIF items several steps may be taken. One procedure is to examine item curves as shown in Figures 1-4. Items showing uniform curves are easier to keep in the questionnaire than the items showing non-uniform curves, because such items are still possible to be included in correlation analyses. In both cases it is important to evaluate the effect size. If the effect size is small the item might be kept. In the present DIF analyses it is the case that most DIF items are uniform and that most effect sizes are low. Another procedure is to run one more DIF analysis but without those items that were flagged included in the first run. Perhaps such a second analysis will give acceptable results. Zumbo (1999), for example, argues for "purifying" the instrument, but warns against excluding items too fast, as the domain that is of interest to measure may become too limited. However, purifying by repeated DIF analyses is not recommended by all experts (e.g., French & Maller, 2007). In this paper we have not applied purifying, mostly because we would like to keep as many items as possible, in

agreement with the Zumbo domain argument above, and also because we want to learn more about the characteristics of the population of flagged items. Some of these items may simply be ambiguous and easy victims for cultural influences in interpretation (e.g., item 7 in Table 10) and others may just be difficult to answer correctly (e.g., item 13), but still possibly influenced by cultural customs. Also, there is one item (41), which should be formulated in a different way in accordance with recent test regulations, as it seems to violate the rights of lesbian and gay relationships.

The most interesting result in the present study is the strong relationship received between DIF structures and cultural values, as measured by the WVS study (Inglehart & Welzel, 2010, 2011). One conclusion from these rank correlations is that DIF results are more related to the Traditional/Secular dimension than the Survival/Self-expression dimension of the WVS map, and that DIF results seem to be a somewhat better indicator of cultural differences in ESCQ than mean scores. Furthermore, it looks as if culture affects the interpretation of an item more strongly than the rated strength of it, as means of total scores do not differ much between the samples. The results also suggest that differences in cultural values are more important than differences in languages as contributing to differences in the number of DIF items among the samples. It should be noted, though, that items have occasionally been discovered to have caused problems for the participants to understand. Those items are few, however, and the changes of meaning due to cultural values seem to be quite subtle. At the present state it is thus difficult to predict which items will be flagged as DIF and which will not. Admittedly, relatively few countries took part in the present study, and DIF analyses and calculations have to be performed again with more countries involved before the present results can be relied on to a satisfactory extent. Also, the present results raise questions. For example, one question has to do with how many of the flagged DIF items are influenced by cultural factors and how many by other factors. Another question is what DIF patterns would look like if people who are more representative for a specific culture participated instead of university students.

How can we conceptually understand the high correlations between DIF results and survey results on values, should our findings be reliable? The ESCQ self-report instrument is derived from the MSCEIT ability battery (Mayer & Salovey, 1996). In the construction and comments of MSCEIT the authors did not say much about the possible influence of basic values (e.g. Mayer, Salover, & Caruso, 2008). However, it is obvious that concepts such as emotion and intelligence in itself are related to cultural values and considered to be important aspects of human life. In the case of MSCEIT such values in all likelihood have an American touch. This goes also for several of the tasks and items that are part of that battery. So far we have not seen any publication with DIF analyses performed on MSCEIT, although Mayer and his colleagues (Mayer, Roberts, & Barsade, 2008; Mayer et al., 2008) actually suggested that greater attention should be directed to cultural effects. It

would really be of general interest that such studies are performed. DIF studies would also make it easier to compare ability testing with self-report testing in the area of emotional intelligence.

Presumably the ESCQ questionnaire is somewhat more vulnerable to the influence of cultural values than the MSCEIT instrument, as several items contain a larger degree of social complexity than what is arranged in the MSCEIT. It is likely that some specific Croatian values have slipped into such items when the questionnaire was constructed. Nevertheless, as shown in the introductory section, the factor structures of ESCQ and MSCEIT are quite similar. The cultural values we are talking about and which make up the two WVS dimensions are distilled from a very large set of survey questions dealing with areas such as *Important in life, Environment, Work, Politics and society, Religion and morale, National identity, Socio-demographics*. Under *Important in life* are included issues such as *Family, Friends, Leisure time, Politics, Work, Religion, Feeling of happiness, State of health*, and *Important child qualities.* Psychological important issues, as suggested by Schwartz (see e.g., Schwartz, 2006, 2009, 2011) are: *To think up new ideas and be creative, To be rich, Living in secure surroundings, To have a good time, To help people nearby, Being very successful, Adventure and taking risks, To always behave properly, Looking after the environment*, and *Tradition*. WVS and other surveys (e.g., European Social Survey) have convincingly demonstrated that there is a considerable variation among countries in opinions about the issues just listed. Any questionnaire or test instrument assessing psychological capacities and abilities are related, more or less, to those basic values. It is not surprising, then, that the ESCQ instrument comprises some items where cultural differences are brought into light by means of DIF analyses. What is surprising is that the number of DIF items seems to be such a strong indicator of the cultural distance between countries. If this finding is supported in future studies, where ESCQ data from a larger number of countries are examined, we will have a better base for understanding the cross-cultural variation of emotional intelligence, and more precise hypotheses about this variation might be formulated.

References

AERA (1999). American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. In *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bartram, D. (2011). Contributions to the EFPA Standing committee on tests and testing to standards and good practice. *European Psychologist, 16*, 149-159.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Ekermans, G., Saklofske, D.H., Austin, E., & Stough, C. (2011). Measurement invariance and differential item functioning of the Bar-On EQ-i: S measure over Canadian, Scottish, South Aftican and Australian samples. *Personality and Individual Differences, 50*, 286-290.

Faria, L., Lima Santos, N., Takšić, V., Räty, H., Molander, B., Holmström, S., ... Toyota, H. (2006). Cross-cultural validation of the Emotional Skills and Competence Questionnaire (ESCQ). *Psicologia, 20*, 95-127.

Fischer, R., & van de Vijver, F.J.R. (2010, July). *Conceptual and practical issues in cross-cultural data analysis*. Presentation at the XX Congress of the International Association for Cross-Cultural Psychology (IACCP). University of Melbourne, Australia.

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

Geisinger, K. (2011, July). Test review at the Buros Centre for Testing. In D. Bartram (Chair), *Test and testing track*. Symposium conducted at 12th European Congress of Psychology, Istanbul, Turkey.

Gignac, G.E., & Ekermans, G. (2010). Group differences in EI within a sample of black and white South Africans. *Personality and Individual Differences, 49*, 639-644.

Hambleton, R. (1994). Guidelines for adapting educational and psychological tests. A progress report. *European Journal of Psychological Assessment, 17*, 187-200.

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

Holmström, S., Molander, B., & Takšić, V. (2008). Analysis of item bias in the Emotional Skills and Competence Questionnaire: A cross-cultural comparison [Abstract]. *International Journal of Psychology, 43*, 354 (Abstract).

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change and democracy*. New York: Cambridge University Press.

Inglehart, R., & Welzel, C. (2010). Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics, 8*, 554.

Inglehart, R., & Welzel, C. (2011). The WVS cultural map of the world. Retrived from http://www.worldvaluessurvey.org/wvs/articles/folder_published/article_base_54.

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*, 329-349.

Kristjansson, E., Aylesworth, R., McDowell, I. & Zumbo, B.D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*, 935-953.

Loewenthal, K.M. (1996). *An introduction to psychological tests and scales*. London, England: UCL Press.

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

Mayer, J.D., Roberts, R.D., & Barsade, S.G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology, 59*, 507-536.

Mayer, J.D., & Salovey, P. (1996). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds.), *Emotional development and emotional intelligence: Implications for educators* (pp. 3-31). New York: Basic Books.

Mayer, J.D., Salovey, P., & Caruso, D.R. (2008). Emotional intelligence. New ability or eclectic traits? *American Psychologist, 63*, 503-517.

Molander, B., Holmström, S., & Takšić, V. (2009). Cross-cultural and sex differences in the Emotional Skills and Competence Questionnaire scales: Challenges of differential item functioning analyses. *Horizons of Psychology, 18*, 35-53.

Nielsen, S. (2011, July). The need for an updated international test standard. In D. Bartram (Chair), *Test and testing track.* Symposium conducted at 12th European Congress of Psychology, Istanbul, Turkey.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.

Pope, G.A. (1997). *Nonparametric item response modeling and gender differential item functioning of the Eysenck Personality Questionnaire* (Unpublished master's thesis). University of Northern British Columbia, Prince George, BC, Canada.

Roussos, L.A., & Stout, W. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-116). Thousand Oaks, CA: Sage.

Schwartz, S.H. (2006). A theory of cultural value orientations: Explication and applications. *Comparative Sociology, 5*, 137-182.

Schwartz, S.H. (2009). Basic values: How they motivate and inhibit prosocial behavior. In M. Mikulincer & P. Shaver (Eds.), *Prosocial motives, emotions, and behavior: The better angels of our nature* (pp. 221-241). Washington, DC: American Psychological Association Press.

Schwartz, S.H. (2011). Studying values: Personal adventure, future directions. *Journal of Cross-Cultural Psychology, 42*, 307-319.

Schwartz, S.H., & Rubel, T. (2005). Sex differences in value priorities: Cross-cultural and multimethod studies. *Journal of Personality and Social Psychology, 89*, 1010-1028.

Smith, L.L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin, 28*, 754-763.

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

Takšić, V. (1998). *Validacija konstrukta emocionalne inteligencije* [Validation of the emotional intelligence construct] (Unpublished doctoral dissertation). University of Zagreb, Zagreb.

Takšić, V. (2011, July). *Emotional Skills and Competence Questionnaire* (ESCQ): *Cross-cultural validation.* Symposium conducted at 12th European Congress of Psychology, Istanbul, Turkey.

Takšić, V., Mohorić, T., & Duran, M. (2009). Emotional Skills and Competence Questionnaire (ESCQ) as a self-report measure of emotional intelligence. *Horizons of Psychology, 18*, 7-21.

Takšić, V., Tkalčić, M., & Brajković, S. (2001, July). *Emotional intelligence: An empirical validation of the construct.* Poster presented at the VIIth European Congress of Psychology, London.

van de Vijver, F., & Hambleton, R. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

van de Vijver, F.J.R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. Berry, Y. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (Vol. 1, pp. 257-300). Needham Heights, MA: Allyn & Bacon.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense, Canada.

Zumbo, B., & Thomas, D.R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Working Paper of the Edge worth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, BC, Canada.

Received: December 2, 2011

PRAZNA STRANICA