Published in: Studia Psychologica - Volume 52 / No. 3 / 2010, pp. 219-228

In search of "the correct answer" in an ability-based Emotional Intelligence (EI) test*

[1]Tamara Mohoric

[1]Vladimir Taksic

[2]Mirjana Duran

[1]Department of Psychology, Faculty of Science and Arts, University of Rijeka, Omladinska 14, 51 000 Rijeka – Croatia

[2]Faculty of Education, University of Osijek, Lorenza Jaegera 9, 31000 Osijek - Croatia

Corresponding author: Vladimir Taksic, PhD; Department of Psychology, Faculty of Science and Arts, University of Rijeka; Omladinska 14; 51 000 Rijeka – Croatia

E-mail: vtaksic@ffri.hr

Tel number: + 385 51 345 051

Fax number: + 385 51 345 207

* The work and the results in this article are a part of the Project "Operationalization and cross-cultural validation of emotional intelligence construct" supported by the Croatian Ministry of Science, Education and Sports

Abstract

One of the main questions in ability-based emotional intelligence tests is the problem of the correct answer. A 197 high school students answered on Vocabulary Emotions Test (VET, Taksic, Harambasic and Velemir, 2004), along with ten experts in the field. We compared different methods for determination the correct answer, with particular regard to expert and consensus scoring methods. The aim of the study was to explore which of the different methods of calculating the correct answer (proportion, mode, lenient mode, distance and adjusted distance) can be used as best estimation of correct answer taken from Croatian Dictionary.

The results showed that experts scoring methods were closely connected to the correct answer, than consensus scoring methods according to the rest of the group.

Key words: emotional intelligence tests, ability EI, consensus scoring, expert scoring, target scoring

Summary

Emotional intelligence tests are based on the idea that EI involves problem solving with and about emotions. A major issue for the ability-based EI tests is the "problem of the correct answer". Because emotion-based systems have no absolute algorithm to determine the correct answer, in everyday contexts the appropriateness of emotional responses is usually determined by agreement with: a) the rest of the group (*consensus scoring*), or b) specialists in a field of emotions (*expert scoring*). In every scoring method individual score is usually generated according to: a) *mode* (category chosen by the largest proportion of the sample is scored as correct), and b) *proportion* of the sample that choose the same alternative.

In order to enhance and deepen the understanding of scoring methods, these methods were compared using results on the Vocabulary Emotion Test (VET, Taksic, Harambasic and Velemir, 2004).

Scoring methods according to experts' consensus were closely connected with the correct answer, higher than scoring methods according to consensus with the rest of the group. Also, it is better to use method-mode, compared to method proportion. This method showed higher correlation coefficients with the correct answer, than did proportion method in both samples (experts and consensus).

1. Introduction

There exist several different definitions of emotional intelligence (EI). Some authors (Mayer and Salovey, 1997) define emotional intelligence by means of achievement tests, while others have the opinion that EI is a composite of different facets (Bar-On, 2000). Although it operates in a mostly unitary fashion, EI is still sub divisible into four different branches according to Mayer, Salovey and Caruso (2000). The first of these branches, emotional perception and identification, involves recognizing and inputting information from the emotional system. The second and third branches, emotional facilitation of thought and emotional understanding, involve the further processing of emotional information and emphasize problem solving. In general, the emotional facilitation of thought branch (2nd branch) involves using emotion to improve cognitive processes, whereas the emotional understanding branch (3rd branch) involves cognitive processing of emotion. The fourth branch of this model, emotion management, concerns emotional self-management and the management of emotions in other people.

Current research suggests that ability models of EI can be described as a standard intelligence and empirically meet the criteria for a standard intelligence (Mayer, Salovey, Caruso and Sitarenios, 2001). Also, only maximum-performance (ability), but not self-report (trait) measures of EI can be seen as tapping the cognitive ability domain (Barchard and Hakstian, 2004). Differentiation of correct and incorrect answers presents a major problem to researchers using ability based EI tests. Usual procedure (when there isn't clear distinction between right and wrong answers) involves conducting a research on a large group of subjects (expert or layman) who’s average answers then serve to calculate subject's individual score. Over the past decade, different scenario-based scales have been developed to measure knowledge and expertise in performance domain such as leadership, as well as to assess emotional, social and general intelligence (Mayer, Salovey, Caruso and Sitarenios, 2003).

While most applications have utilized expert groups to develop scoring standards, others have constructed scoring keys based on data collected from large groups of respondents who were knowledgeable concerning the subject domain but could not be qualified as experts. The scoring keys from these groups of non-experts were believed to have closely approximated the scoring standards that would have been obtained by experts. This form of measurement was named *consensus-based measurement*. It provides a maximal performance based method to assess knowledge-related constructs and it is relevant to conceptualizations of emotional intelligence that propose a related set of knowledge, skills, and abilities (Legree, Psotka, Tremble and Bourne, 2005).

A major issue for performance-based or ability EI tests is the "problem of the correct answer". Because emotion-based systems have no absolute algorithm to determine the correct answer, in everyday contexts people usually determine the appropriateness (or correctness) of emotional response by agreement with the rest of the group interacting in the same emotional system. This form of determination of the correct answer is called *consensus scoring* (correct answer is what a group agrees upon). There are also *expert scoring* (an expert in e.g. EI says what is the correct answer), and *target scoring* (the creator of the item or the test says what is the correct answer). MacCann, Roberts, Matthews and Zeinder (2004), in their comparison of different scoring methods, emphasize several theoretical problems with the usage of these methods in the field of EI. First, there is no criterion for who is an"expert" in emotional intelligence. Also, some evidence suggests that scores can be higher for test takers who are similar to the experts. Target scoring also has some problems, as the target themselves may not be able to express the emotion they are feeling accurately, or that they may report only pleasant or pro-social emotions.

There are five different consensus scoring methods, called proportion, mode, lenient mode, distance and adjusted distance (for details see MacCann et. al., 2004). In the *proportion*

method a score is allocated to each response according to the proportion of people endorsing that response. In the *mode* method the rating or category chosen by the largest proportion of the sample (the mode) is scored as correct, and all other responses are scored zero. The *lenient mode* is similar and marks are awarded to the modal rating, and to the ratings one scale value either side of it. In the *distance* method the score awarded for an item is the difference between a participant's response and the optimum scale value. In the case of consensus distance scoring, the optimum Likert-scale value is calculated as the mean rating (over all participants) for that item. *Adjusted distance* is very similar to distance method; the only difference is that an individual's ratings for each item are converted to z-scores.

In spite of existing theoretical and other statistical problems (see MacCann et al., 2004), ability-based emotional intelligence tests are most commonly scored by consensus method. Consensus based measurement is especially relevant to measuring EI because emotional intelligence is an example of a domain that is still lacking in the availability of experts and objective knowledge. Also, one can expect differences in the determination of the correct answer due to cultural or group specificity or bias.

In order to enhance and deepen the understanding of different scoring methods for EI ability tests these methods were compared using results obtained on Vocabulary Emotions Test (VET; Taksic, Harambasic and Velemir, 2004). VET is emotional intelligence test designed on the bases of one of the branches from the Mayer-Salovey model of EI – understanding of emotions, in Croatian setting. Since this test has the correct answer (according to the Croatian Dictionary, Anic, 1994), it could be used to assess which of the scoring methods is better, in the sense of being more in accordance with the actual correct answer. VET has showed correlation with standard vocabulary test, but in the same time demonstrates high proportion of unique variance (Taksic, Harambasic and Velemir, 2004).

The aim of the present study was to explore which of the different methods of calculating the correct answer (proportion, mode, lenient mode, distance and adjusted distance) can be used as best estimation of "real" correct answer on VET (given by the Croatian Dictionary), which could be regarded as highly expertise solution (or target criterion).

2. Method

2.1. Participants

Subjects were 197 high school students (150 girls and 47 boys), with an age range from 15 to 19 years (M =16.69, SD =1.08). All subjects were recruited from different high schools in the city of Rijeka, Croatia.

Experts: Ten experts participated in this study. The experts were all university teachers, from Departments of Psychology at University of Rijeka and Zagreb. All of the experts were familiar with the problem of emotion, either through their research interest or practical work, e.g. as clinical psychologist. Nine experts were female and one was male.

2.2. Instruments

2.2.1. *Vocabulary Emotions Test* (VET; Taksic, Harambasic and Velemir, 2004)

This test was developed within Mayer and Salovey′s conceptualization of emotional intelligence, and represents a measure of ability from branch c: *understanding emotions*. The test consists of 102 adjectives describing various emotional states and moods. The first adjective is the target word, followed by six adjectives with similar meaning. A subject has to choose one adjective (out of six) which is closest in meaning to the target word. It is important to emphasize that this test has a correct answer, based on a solution from a Croatian dictionary (Anic, 1994). Test has been used in various researches and had shown satisfying

psychometric properties, with reliability coefficient α=.91 (Taksic, Harambasic and Velemir, 2004), and has 44% of unique variance over and above various tests of standard intelligence.

In order to compare different scoring methods, for every participant five different types of "correct answers" were obtained according to: 1) consensus proportion criterion, 2) consensus mode criterion, 3) expert proportion criterion, 4) expert mode criterion, and 5) Croatian Dictionary (Anic, 1994).

In the *consensus method*, each of a respondent's answers was scored against the *proportion* of the sample that endorsed the same answer. For example, if the same alternative was chosen by 50% of the sample, the individual's score would be augmented by the proportion of 0.50. The respondent's total raw score was the sum of those proportions across all of the 102 items of the test. The second consensus method used here was *mode* - the category chosen by the largest proportion of the sample was scored as correct (and got 1 point), and all other responses were scored as zero. The total raw score was the sum of all the correct answers.

Beside consensus method, an *expert scoring method* was also used. In this method the correct answer was determined also according to a consensus, not of the whole sample but according to a consensus among experts. So, if experts choose A as the correct answer for the test item by 70% and C by 30%, then answer A got proportion of 0.70 and C got proportion of 0.30 (method *proportion*). Each of the respondent's scores was then evaluated against the criterion formed by proportional responding of an expert group (in this case, the 10 university teachers). So, if the participant chose A on that particular item, he/she got a proportion of 0.70, if he/she choused C then got proportion of 0.30, and answers B and D did not brought any points. Because majority of experts always chose the solution that is the same as the one from the Dictionary, method *mode* came to the same solution as did correct answer, and as a consequence it was not mentioned in the further analysis.

In general, in consensus scoring method we used method proportion and mode, and in expert scoring method only proportion was displayed.

Reliability coefficients for different scoring methods of VET were also calculated. Results are shown in Table 1, along with some descriptive statistics. Reliability of VET is satisfying for 3 different scoring methods (correct answer, expert proportion and consensus mode) and is rather low for consensus proportion method.

Table 1 here

2.2.2. *Additional information*

We also collected some additional information about:

- academic achievement (GPA - grade point average);
- number of negative grades after the first semester;
- absences from school (frequency of unexplained or unjustified absence); and
- teacher's assessment of student's school behavior (from 1- unacceptable till 5 - remarkable).

2.3. Procedure

Testing was conducted in classrooms during school hours. Each participant received tests to answer and detailed instructions. Researcher was present during the testing to answer any potential questions. Participants were debriefed upon the completion of testing.

Each expert answered on the VET test individually.

3. Results

3.1. Analysis of expert answers

In order to examine the level of agreement between experts, the inter-item correlation of the experts' answers was calculated. Inter-item correlation was r=.82 (consensus among experts

about chosen answer) which shows that in most cases experts agreed about the correct answer on a test-item.

As it was said before, the majority of the experts always chose the correct answer (as in Dictionary solution), so the correlation between correct answer and expert-mode was the highest possible (r=1.00).

Analysis of experts' answers showed that for 55 (out of 102 adjectives) experts agreed 100% about the correct answer (all ten experts selected the same alternative, which was the same as the solution from the dictionary). For 47 other adjectives results were not so clear. For some adjectives a large number of experts chose the same answer (e.g. 9 vs. 1), but for some adjectives ratings were more evenly distributed (e.g. 6 vs. 4; or 5 vs. 4 vs. 1). These adjectives are the problematic ones and their characteristics should be examined in further researches.

3.2. Convergence of different methods

Every participant had five test scores on VET, based on: 1. correct answer (dictionary solution), 2. consensus scoring method – proportion, 3. consensus scoring method – mode, 4. expert scoring method – proportion, and 5. expert scoring method – mode.

The correlation coefficients among total scores based on different criteria are calculated, and shown in Table 2.

Table 2 here

In order to test the convergence of the different scoring methods, we calculated participants' test results using two general (consensus) criteria (proportion and mode), on the one hand, and the expert criterion, on the other. The correlation between the score sets was .34 (consensus-proportion and expert scoring) and .68 (consensus-mode and expert scoring). Correlation of consensus-mode and expert-mode scoring method was significantly larger than correlation of

consensus-proportion and expert method (t=8.17, p<0.01). Also, consensus-mode method had a higher correlation with expert scoring than the consensus proportion method.

As it is shown in Table 2, both models of scoring (the expert and the consensus-mode scoring method) have significant and large correlations with total score calculated according to the correct answer from the Croatian dictionary. The consensus proportion method also has a significant but relatively lower correlation with correct answer, than expert proportion method.

The correlation between consensus-mode and correct answer was .88 and the correlation between consensus-proportion method and correct answer was also significant but relatively low (r=.53)**.** The difference between these two correlation coefficients is statistically significant (t=13.96, p<0.01) indicating that consensus-mode is more appropriate for calculating the correct answer in this kind of EI test than consensus-proportion. It is also important to notice that two different consensus scoring methods have significant but relatively low correlation (r=.69).

Mayer and Salovey's (1997) model of EI hypothesizes that emotional knowledge is embedded within a general, social context of communication and interaction (Mayer et al., 2001). Emotion experts would be more likely than others to possess an accurate shared social representation of emotional knowledge, and as a consequence, they could provide an important criterion for the tests correct answers. If this hypothesis was true, then experts would agree more on the correct answer than the general group. Our results confirm this hypothesis - mean correlation between subjects ratings is .32, what is significantly lower than mean correlation between experts (r= .82).

After regarding all of the above mentioned results, it could be concluded that when using a consensus method for the determination of correct answer in this EI test, it is better to use method-mode (category chosen by the largest proportion of the sample is scored as correct,

and all other responses are scored zero), than method-proportion. This method had higher correlation coefficients with both correct answer and expert criteria than consensus-proportion method.

The four different ways of obtaining correct answers were compared with some criterion variables. Correlations were calculated for every scoring method and the criterion variables (Table 3).

Table 3 here

As expected, the consensus-proportion method had the lowest correlations with all of the criterion variables (and, as it was mentioned before, it also had a low correlation with the correct answer).

Comparison of different scoring methods shows that the highest correlation with various criterion variables was obtained with the expert method of scoring, even higher than the correct answer. The difference between obtained coefficients of correlation is statistically significant for academic achievement and number of negative grades. As stated before, for 55 adjectives all experts agreed upon the correct answer, but the other 47 adjectives they had two or more different alternatives chosen, that means alternatives different than ones in the Dictionary. So, for those adjectives participants got some points even for the alternatives which are not correct according to a dictionary solution. The problem with the correct answer (dictionary solution) may be that a solution from a dictionary is too academic, while some experts may have chosen alternatives which are closer to everyday life and language usage and so, closer to those of participants.

4. Discussion

Emotional intelligence tests are based on the idea that EI involves problem solving with and about emotions. Such ability tests are believed to measure something different than self-report

scales of EI, with which correlations are relatively low (Mayer, Salovey and Caruso, 2002). EI test used in this research (Vocabulary Emotion Test) measures understanding of emotions, ability that according to Mayer and Salovey's (1997) model represent third branch of EI model. The Mayer-Salovey model supposes that the third branch is the most cognitive and should have a relationship with abstract reasoning.

Since the most serious problem with EI tests is the problem of the correct answer, different scoring methods of emotional intelligence test were compared. The aim was to find out which of the scoring methods has the highest correlation with some criterion variables, and also to examine if experts and general sample (consensus) differ substantially in the selection of the correct answer to a test item.

Because humans live in environment where interaction and relations to other people ideas, rules, etc. are so important, it is justified to believe that human beings have come to a general consensus about many emotional meanings. That is also the main reason and justification for using consensus scoring in EI tests. But it is also important to notice that this does not mean that there is only one accurate way to feel or interpret feelings. According to our results general- and expert-consensus scoring relatively effectively come together, although there were some significant differences. The evidence from this study also reflects that experts are more reliable judges than general sample. If further studies confirm these results, the expert criteria may prove preferable to those of general consensus judgments.

In their early work, Mayer, Caruso and Salovey (1999) used an expert criterion based on only two experts and a general consensus method (to score the earlier version of MEIS) and found those methods only partially converged. But in the later research (Mayer et. al., 2003) they again used an expert criterion and a general consensus method (to score the MSCEIT measure), now using 21 experts. They found high level of convergence between the expert and consensus scoring methods. Some authors (Roberts, Zeidner and Matthews, 2001)

replicated the finding and raised the lack of convergence as a serious problem. Others have argued that, as more experts are used, and their answers aggregated, their performance would resemble that of the consensus of a large general group (Mayer et al., 2001).

The ten experts in this study did exhibit superior agreement levels relative to the general sample. At the same time, the expert and general consensus criteria often agreed about correct answer. Unfortunately, there were some test items for which experts did not agree on the correct solution and those items need further attention. Although for some items there was a general consensus among experts on correct answer, for other items the ratings were distributed evenly among two (or more) different alternatives.

Legree et al. (2005) emphasize several important implications that consensual scoring has for studying individual differences. The approach allows the construction and scoring of scales for knowledge domains for which experts do not exist, or cannot be easily defined. Also, it allows the assessment of knowledge domains that have not been traditionally addressed in psychological research, and broadens the domain of psychological assessment and intelligence research into horizontal aspects of intelligence, e.g. emotional or social intelligence. Also, consensus based scoring has the potential to allow for the same protocol to be scored against multiple standards. Much social knowledge represents the convergence between many perspectives and the truth is commonly believed to exist at the intersection of these perspectives. Useful knowledge can be distributed over individuals, so consensus based measures are needed to analyze this type of knowledge and its evidentiary sources for emerging fields like emotional intelligence.

Matthews, Zeidner and Roberts (2002) emphasize some additional problems with expert scoring. They say it is important to know the structure of the experts (e.g. are they white, middle-class, highly educated). Also, they feel that the view of experts may reflect cultural consensus rather than special expertise, especially since there is a problem of expertise in

many special sub-fields of emotion. In our sample experts were university professors (experts in a field of emotions), and mainly female, which could modify the results.

The consensus scoring method leaves little scope for the item metric analyses that are a central element of emotional intelligence test development. Normally, items of graduated difficulty are required to ensure comparable reliability of measurement across the full range of abilities. Consensus scoring, by its nature, excludes identification of difficult items on which, say, only the 10% most capable individuals reach the correct answer, and the consensus answer is then - incorrect.

In spite of these problems, expert scoring is often used in many researches because of its many advantages. Our research also showed that experts agreed more on correct answer than general sample, although there are some test items for which experts did not agreed. Nevertheless, consensus scoring method (whether expert or general sample) will be important for the development of future EI tests, and in order to work on psychometric issues of EI tests, researches first need to answer on the basic problem of the correct answer.

Some shortcoming of this study should be corrected in the future. For example, it would be very interesting and useful to examine VET against other ability based measures of EI, such as MSCEIT. Good starting point is that VET showed high proportion of specific variance in comparison with standard vocabulary test (Taksic, at al, 2004), but it still would be necessary to include other indicators of verbal comprehension in order to show that VET is not a measure of vocabulary but a measure of an aspect of emotional intelligence. VET is now validated only in Croatian settings, and it needs to be evaluating cross-culturally. The first translations were in English and Swedish language, and the results are encouraging.

Also, in the present samples of subjects and experts there was more females than males, what could influence the results know the fact that females showed higher EI abilities in many

studies (Zeidner and Kaluda, 2008). Balanced sample and using more measures of EI could be recommended for future researches.

References

Anic, V. (1994). *Croatian Dictionary* [Rjecnik hrvatskoga knjizevnog jezika, 3]. Zagreb: Novi liber.

BarOn, R. (2000). Emotional and social intelligence: insights from Emotional Quotient Inventory. In: R. Bar-On & D.A. Parker (Eds.). *The Handbook of Emotional Intelligence* (pp. 363-388). San Francisco: Jossey-Bass.

Barchard, K.A and Hakstian, A.R. (2004). The nature and measurement of Emotional intelligence abilities: Basic dimensions and their relationships with other cognitive ability and personality variables. *Educational and Psychological Measurement, 64*, 437-462.

Legree, P.J., Psotka, J., Tremble, T. and Bourne, D.R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze and R.D. Roberts, (Eds.). (2005). *Emotional Intelligence: An International Handbook,* (pp. 155-179). Cambridge: Hogrefe & Huber Publishers.

MacCann, C., Roberts, R.D., Matthews, G. and Zeinder, M. (2004). Consensus scoring and empirical option weighting of performance-based Emotional intelligence (EI) tests. *Personality and individual differences, 36*, 645-662.

Matthews, G., Zeidner M. and Roberts, R.D. (2002). *Emotional intelligence: Science & Myth*. Cambridge: The MIT Press.

Mayer, J.D. and Salovey, P. (1997). What is emotional intelligence?. In: P.Salovey and D. Sluyter (Eds.). *Emotional development and emotional intelligence: Implications for educators* (pp. 3-31). New York: Basic Books.

Mayer, J.D., Caruso, R.D., and Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence, 27*, 267-298.

Mayer, J.D., Salovey, P. and Caruso, D. (2000). Emotional intelligence as *zeitgeist*, as personality, and as mental ability. In: R. Bar-On & D.A. Parker (Eds.). *The Handbook of Emotional Intelligence*. San Francisco: Jossey-Bass.

Mayer, J. D., Salovey, P. and Caruso, D. R. (2002). *Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) user's manual.* Toronto, Ontario, Canada: MHS Publishers.

Mayer, J.D., Salovey, P., Caruso, D. and Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion, 1*, 232-242.

Mayer, J.D., Salovey, P., Caruso, D. and Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT v2.0. *Emotion, 3,* 97-105.

Roberts R.D., Zeidner M. and Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? Some new data and conclusion. *Emotion, 1*, 196-231.

Taksic, V., Harambasic, D. and Velemir, B. (2004). Emotional Vocabulary Test as an attempt to measure the emotional intelligence ability - understanding emotion aspect. *International Journal of Psychology, 39,* 5-6 (Suppl S).

Zeidner, M. and Kaluda, I. (2008). Romantic love: What's emotional intelligence (EI) got to do with it? *Personality and Individual Differences*, 44, 1684–1695

Tables

Table 1. Mean, standard deviation and reliability coefficients for different scoring methods on VET

	N	M	SD	Cronbach Alpha
VET_Correct answer	197	62.82	14.15	0.91
VET_Experts' proportion	197	42.83	8.67	0.81
VET_Consensus mode	197	58.66	12.59	0.87
VET_Consensus proportion	197	22.57	2.73	0.53

Table 2. Correlation coefficients for different total scores on VET

	Experts mod	**Experts proportion**	**Consensus mode**	**Consensus proportion**
Correct answer	1.00**	.88**	.88**	.52**
Experts' proportion			.68**	.34**
Consensus mode				.69**

** p< 0.01

Table 3. Correlation coefficients for different total scores and criterion variables

	GPA - achievement	**No of neg. grades**	***F* of unexplained absence**	**Assessment of school behavior**
Correct answer	.30**	-.15*	-.13	.14
Experts' proportion	.35**	-.23**	-.15*	.22**
Consensus mode	.22**	-.09	-.15*	.06
Consensus proportion	.14*	-.08	-.09	.01

* p< 0.05; ** p< 0.01