Psicothema 2006. Vol. 18, supl., pp. 49-54
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Bias in consensus scoring, with examples from ability emotional intelligence tests

Kimberly A. Barchard and * James A. Russell
University of Nevada, Las Vegas, and * Boston College

Consensus scoring occurs when the scoring key for a test is based upon the responses of the norm group. Consensus scoring is an attractive alternative to traditional methods of creating a scoring key for ability tests, especially useful when experts disagree about the correct answers to test items, as they do in the area of emotions and emotion perception. Of the many variations of consensus scoring, mode consensus scoring (the most frequent response in a norm group is given a score of 1, and all other responses a score of 0) and proportion consensus scoring (each respondent's score on an item is equal to the proportion of the norm group who match the respondent's answer) are the most widely used and the most psychometrically promising. This paper demonstrates that mode consensus scoring is biased against smaller sub-groups within the norm group: when sub-groups differ in their modal responses, the size of the sub-groups will influence the average group score. No known scoring option eliminates this bias. In contrast, proportion consensus scoring is not necessarily biased against smaller groups, although bias does occur in some extreme situations. Proportion consensus scoring is therefore the preferred consensus scoring option at this time.

Sesgos en la puntuación de consenso: ejemplos en pruebas de habilidad de inteligencia emocional. La puntuación de consenso se produce cuando las claves para puntuar un test están basadas en las respuestas de un grupo normativo. La puntuación de consenso es una atractiva alternativa a los métodos tradicionales que crean normas de puntuación para los test de habilidad. Es especialmente útil cuando los expertos no están de acuerdo en cuáles son las respuestas correctas a los ítems del test, como ocurre en el campo de las emociones y de la percepción de éstas. De las diferentes variantes de la puntuación de consenso, la puntuación de consenso basada en la moda (la respuesta más frecuente en un grupo normativo puntúa 1, todas las respuestas restantes puntúan 0) y la puntuación de consenso basada en la proporción (la puntuación de cada individuo en un ítem se corresponde con la proporción del grupo normativo que ha dado la misma respuesta a ese ítem) son los métodos más ampliamente utilizados y que muestran las propiedades psicométricas más prometedoras. Este artículo demuestra que la puntuación de consenso basada en la moda sesga los resultados de los pequeños subgrupos dentro del grupo normativo: cuando los subgrupos difieren en su respuesta modal, el tamaño de los subgrupos influirá en la puntuación media del grupo. Ningún sistema de puntuación conocido elimina este sesgo. En cambio, la puntuación de consenso basada en la proporción no sesga necesariamente los grupos pequeños, aunque puede mostrar sesgos en algunas situaciones extremas. Por lo tanto, la puntuación de consenso basada en la proporción es la opción más adecuada.

Tests of Emotional Intelligence and Emotion Perception are often scored using consensus scoring (Brackett & Mayer, 2006; Geher, Warner, & Brown, 2001; Mayer & Geher, 1996; Mayer, Salovey, & Caruso, 1999; Mayer, Salovey, Caruso, & Sitarenios, 2003). In consensus scoring, a respondent's score on an item is based upon the responses of the norm group. To illustrate two common types, consider an item that shows a picture of a face with a particular expression. The respondent is asked to indicate the emotion conveyed by that facial expression by choosing one response option from a list: (A) happy, (B) sad, (C) scared, and (D) angry. Imagine that in the normative group for the test (which might simply be all those who responded to it), 5% of the respondents chose A, 60% B, 20% C, and 15% D. In *proportion consensus scoring*, someone who selected A would obtain a score of .05; someone who selected B would obtain .60, and so on. In *mode consensus scoring*, someone who selected B would obtain a score of 1, while all other responses would receive a score of 0. In both types of scoring, higher scores would typically be interpreted as stronger ability to recognize emotions from faces.

Research from a variety of content areas has provided evidence for the validity of consensus scoring.[1] However, in this article, we show that mode consensus scoring is biased against smaller subgroups of respondents. For example, if there are more women than

Correspondence: Kimberly A. Barchard
Department of Psychology
University of Nevada, Las Vegas
PO Box 455030, Las Vegas NV, 89154 (Usa)
E-mail: barchard@unlv.nevada.edu

men in the norm groups used, across many experiments, men will receive lower average scores than if there were equal numbers of men and women in the norm groups. Available methods of implementing mode consensus scoring fail to overcome this bias. Proportion consensus scoring, in contrast, is not biased in this way and is therefore the method of choice.

The move to consensus scoring

A test builder who has developed a set of items, such as an ability test of emotional intelligence or emotion perception, must have a way of scoring responses to those items. In some cases, none of the more traditional techniques are fully adequate. Return to the facial-expression item mentioned above. One traditional method is to ask experts to indicate which option is correct. This route is problematic when experts disagree, as happens in the study of facial expressions specifically and emotion in general (e.g., Ekman, 1972; Fridlund, 1994). Another traditional way to develop a scoring key is factor analysis. When applied to items from an ability test, this route is more likely to group items by method than by content (for a discussion of method factors, see e.g., Bank, Dishion, Skinner, & Patternson, 1990; Levin, 1973) and, in any case, requires the test developer to have already specified how each item is scored before the factor analysis can be conducted. A third traditional way to develop a scoring key requires a criterion group: in the case of the facial-expression item, this would be a group of people who are known to be skilled in detecting emotions from faces. However, there is no such group commonly agreed upon. A fourth way is target-scoring (Mayer, Caruso, & Salovey, 2000). In the example of the facial-expression item, the test builder could ask the person whose face is shown in the test what emotion he or she was feeling. Often such information is simply unavailable: the person who posed might not have been asked. Furthermore, some items have no target: consider items for a test of emotion vocabulary or of comprehension of emotion metaphors. Even when a target is available, one could question the accuracy of that individual's introspective answer. More generally, traditional methods are likely to produce inadequate or suboptimal scoring keys whenever one is trying to assess an ability for which (a) there are group differences in test scores which might not reflect true differences in the underlying ability, (b) answers are context-specific, (c) experts believe there are correct answers but disagree about what these are, or (d) no bona fide experts exist. These last two situations were discussed by Legree et al. (2004). In such cases, consensus scoring presents an attractive alternative.

Several different consensus scoring methods have been invented. In addition to proportion and mode consensus scoring, introduced above, lenient mode, distance, and adjusted distance methods also exist. Of these methods, however, only proportion and mode consensus scoring result in unidimensional scores and demonstrate convergent validity (MacCann, Roberts, Matthews, & Zeidner, 2004). Therefore these two methods are the focus of our article.

Bias in mode consensus scoring

At first blush, all forms of consensus scoring appear vulnerable to an accusation of bias. By their sheer numbers, the largest sub-group of the normative group seems to determine the "correct" answer and therefore smaller sub-groups appear likely to get lower scores. Sub-groups could be based on socioeconomic status, ethnicity, age, sex, or on less visible characteristics such as education. Researchers and applied tests users will almost certainly be concerned about and will want to try to eliminate any such bias.

This intuition is correct for mode consensus scoring. First we will show that for an individual item, mode consensus scoring introduces bias most of the time when the smaller group differs from the larger group in its modal response to that item (for example, the smaller sub-group selects option #1 most frequently and the larger group selects option #2 most frequently). Next we will show that mode consensus scoring creates bias at the level of the total test score when one or more items are biased in this way. This bias can challenge the validity of the test (when larger and smaller groups simply have different opinions), lead to absurdity (when the smaller group consists of experts), or even be illegal (when the differences fail to correspond to differences in performance). In this section, we show precisely when mode consensus scoring creates this bias.

First let us consider how the average score for a group is calculated for a particular item. We begin with the case of two groups, Groups A and B. Each person in the two groups completes a single multiple-choice item that has J response options. The proportion of people giving the j^{th} response is given as a_j and b_j, in Groups A and B respectively, where j ranges from 1 to J. For example, the proportion of people giving the third response in Group B is given as b_3.

The average score for a particular group is equal to the sum of the products of the proportion of people giving a particular response and the score obtained for that response:

$$y = \sum_{j=1}^{J} p_j s_j$$

where y is the average score, p_j is the proportion of people giving response j, and s_j is the score given for response j. Furthermore, $\sum_{j=1}^{J} p_j = 1$ for any set of proportions, and in both mode and proportion consensus scoring $\sum_{j=1}^{J} p_j = 1$. For ease of discussion, we make one further stipulation. In Group A, one of the J responses will have a higher frequency than the others. Let us denote this response as k, where k is between 1 and J, so that the highest proportion is a_k.

We next show that the item mean for Group A varies depending upon whether the scoring key is based on the group being scored (within-group scoring) or a different group (between-group scoring). First let us consider the case when Groups A and B have the same mode, response k. Regardless of which group is used to develop the scoring key, response k will be scored as correct, and all other options will be scored 0. Thus, within-group and between-group scoring will result in identical scoring keys and therefore identical scores for Group A.

If the modes differ, however, within- and between-group scoring will result in different scoring keys and different average scores for Group A. If the scoring key is developed from Group A (within-group scoring), the response with the highest frequency in Group A (response k) will be given a score of 1, and all other responses will be given a score of 0. Therefore, the average score of Group A, using the scoring key developed from Group A, is equal to a_k. This is the highest average score it is possible to get in

Group A: if one of the non-modal responses were scored as correct, the average score could not be any larger than a_k, because all other responses have frequencies that are less than or equal to the frequency of response *k*. Because of this, if Group A is instead scored using a scoring key developed from Group B (between-group scoring), the average score for Group A is decreased. The mode from Group B is given a score of 1 and the mode from Group A is given a score of 0, resulting in lower average scores in Group A than when within-group scoring was used.

We now apply the above results to determine how group means are influenced when one of these groups is much larger in size than the other, but both groups are used to create a combined-group scoring key. If the modes in Groups A and B are the same, the use of the combined-sample scoring key will result in the same scores as within-group scoring. However, if the modes are different, the combined-sample scoring key will be most similar to the scoring key from the larger group. Suppose that Group A is much larger than Group B. Most of the time the modal response in Group A will determine the modal response in the combined sample. In this case, combined-sample scoring will result in the same average scores for Group A as within-group scoring, because the modal response will receive a score of 1 under both scoring keys. However, the combined-group scoring key will result in lower average scores for Group B than within-group scoring, because their modal response receives a score of 0 under combined-group scoring.

Most tests consist of many items. Given the effect of different modal responses on item-level scores, what will happen to total test scores if a combined sample is used to create the scoring key for the full test? The use of mode consensus scoring based on a combined sample will not influence item-level scores for items with the same modal response in different sub-groups. However, mode consensus scoring will typically decrease the average scores of the smaller group, for those items with different modal responses. Therefore, the overall effect of mode consensus scoring across the entire test may be to decrease scores for the smaller group.

Let us consider an example. If there are many more women than men in the norm group, men can receive lower average scores using a combined-group scoring key than if they were scored using a scoring key developed using only men, but they cannot receive higher average scores using the combined-group scoring key. In contrast, the larger group, in this case women, will typically receive the same scores, regardless of whether they are scored using a within-group scoring key or a combined-group scoring key. Over a large number of items and a large number of experiments, there are certain to be at least some items that are biased and which decrease the scores of smaller groups. Therefore, overall, mode consensus scoring is biased against smaller groups.

Trying to eliminate the bias of mode consensus scoring

Can the bias in mode consensus scoring be eliminated? In this section, we consider the effects of three different mode consensus scoring methods on average group scores, when the modal response is different in two groups. The first strategy would be to use mode consensus scoring as originally intended, including all sub-groups (large and small) in the norm group used to develop the scoring key. As shown above, this technique is likely to result in lower scores for the smaller groups. Although majorities are sometimes right and minorities wrong, the reverse is also possible. For example, experts can be defined as a small group of people who know more than the average person. Therefore, a scoring key that assumes *a priori* that smaller groups are always wrong is on shaky ground.

A second strategy would be to use within-group scoring exclusively: each person's score is based upon the most applicable norm group, and raw scores are scaled so that every group has the same mean score. For example, Caucasian men might be scored according to a norm group of Caucasian men; Hispanic women according to a norm group of Hispanic women, and so on. Although substituting within-group scoring for combined-group scoring eliminates the bias discussed above, within-group scoring suffers from both theoretical and practical problems. The theoretical problem is that it is questionable whether all groups are equally knowledgeable, as presupposed when group means are equated. Consider again the obvious example of one group being experts. In addition, there are two practical problems. First, when making important decisions about individuals (such as access to advanced education, employment or promotion), it may be illegal to use information about gender, ethnicity, and even age in scoring the test. Second, it is often impractical to obtain sufficient sample sizes for each norm group required. For a ten-item subscale, a sample size of only 100 is probably sufficient to obtain a scoring key that is not overly influenced by the particular individuals included in the norm group[2], but obtaining 100 people for each combination of relevant demographic variables may be difficult.

A third approach is to use a hybrid of consensus scoring and expert scoring: create a scoring key based on the consensus of a large sample of experts. Because they are experts, their modal response can be argued to be a better response, even if it disagrees with the modal response of the general population. This expert-consensus approach has been used with various scoring methods (proportion scoring, mode scoring, and others) in the areas of emotional intelligence, driving skill, general intelligence, supervisory skill of non-commissioned officers, and military leadership, and it has resulted in scores with moderate to high correlations with general consensus scoring (Legree et al., 2004). Although a promising approach in principle, expert-consensus scoring suffers from two practical problems. In some domains (such as those related to emotions) identifying genuine experts may be difficult - when experts disagree among themselves, choosing the experts may be a subjective judgment, itself a possible source of bias. The second problem is that a relatively large sample of experts may be needed, as discussed above for within-sample scoring, unless the test is very long and no subscale scores are desired.[3] In addition, expert-consensus scoring keys might not, in fact, eliminate the bias against smaller groups that general mode consensus scoring suffers from. If differences between experts correspond to differences in the general population, then the resulting mode consensus scores will be biased against whichever group was smaller among the experts.

Lack of necessary bias in proportion consensus scoring

Proportion consensus scoring is not necessarily biased against smaller groups the way mode consensus scoring is, because the use of between-group scoring keys does not necessarily reduce average scores, even when the modes are different in larger and smaller sub-groups. Let us consider again the average scores for Group A, when within- and between-group scoring is used. If

Group A is scored using a scoring key developed from Group A (within-group scoring), then in proportion consensus scoring $p_j = a_j$ and $s_j = a_j$, so that the average score is

$$y_{AA} = \sum_{j=1}^{J} a_j a_j = \sum_{j=1}^{J} a_j^2$$

If Group A is scored using a scoring key developed from Group B (between-group scoring), then $p_j = a_j$ and $s_j = b_j$, so that the average score is

$$y_{AB} = \sum_{j=1}^{J} a_j b_j$$

To compare between-group scoring with within-group scoring, we need to compare y_{AA} to y_{AB}.

We begin by considering one example in which between-group scoring results in higher-scores than within-group scoring, when the modes in the two groups are the same. Imagine that in both Groups A and B, 50% of people select response A. However, the proportion of people selecting the other options varies by group, resulting in different average scores if within-group or between-group scoring is used. In Group A, 25% of people select response B, and 12.5% select each of responses C and D. From this, we can calculate that y_{AA}= .50*.50 + .25*.25 + .125*.125 + .125*.125= .34375. In Group B, all the people who did not select A selected B. From this, we get y_{AB}= .50*.50 + .25*.50 + .125*0 + .125*0= .375. Thus, using the scoring key from Group B will result in a higher average score than using the within-group scoring key: $y_{AB} > y_{AA}$.

The higher score for between-group scoring can also be seen in figure 1. The average score for a group, $\sum_{j=1}^{J} p_j s_j$, can be visualized as the sum of the areas of J rectangles, where the width of the rectangle is given by p_j and the height is given by s_j. The first rectangle, for example, has a width of p_1 and a height of s_1. Based upon the previously stated constraints, we know that the sum of the widths of all the rectangles is 1, and the sum of the heights is also 1. Therefore, increasing the height of one rectangle requires us to decrease the height of the other rectangles. Generally, the within-group scoring key results in a high average score, because the most frequent response is given the highest score. However, by decreasing the height of the rectangles with the smallest widths, and adding that height to any rectangle with a larger width, we can increase the total area. Thus, for this particular set of data, the total area is larger when between-group scoring was used (giving a higher score to response B) than when within-group scoring was used.

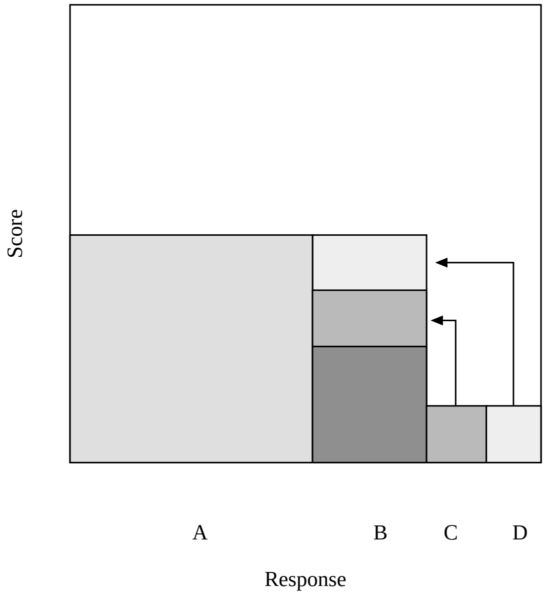

Figure 1. Increasing the score given to a non-modal response may increase the average score, which equals the sum of the areas of the rectangles

One can also imagine situations in which the average score is higher using between-group scoring, even when the group modes differ. For example, if in Group B, 40% of people select response A, and 60% select response B, y_{AB}= .50*.40 + .25*.60 + .125*0 + .125*0 = .35, which is still larger than y_{AA}. Thus, although it is generally true that a higher average group score will result if a higher score is given to the responses that are most frequent in Group A, giving a lower score for the modal response can be compensated for by giving higher scores to responses with intermediate frequencies. Because of this, between-group scoring does not *necessarily* result in lower average scores at the item level for proportion consensus scoring.

Although there is no necessary reduction in the average score of the smaller group, the reader might worry that there might still usually be a reduction when between-group scoring is used. However, there is no reason to think this would be the case. Although there is a relationship between the distribution of responses and the average group score, there is no relationship between group size and the distribution of responses. Therefore, there is no reason to think that between-group scoring will usually result in lower average item-level scores, when using proportion consensus scoring.

From this it follows that when two groups of unequal sizes are combined, there is no reason to think that proportion consensus scoring would usually result in lower scores for the smaller group at either the level of the individual item or the total test score. Thus, proportion consensus scoring is not biased against smaller groups in the way mode consensus scoring is.

There are, however, *sometimes* situations in which group size will be related to average group score using proportion consensus scoring. Consider, for example, a situation in which there is strong within-group consensus and strong between-group disagreement. The most extreme case of this would be when all members of Group A select one response, and all members of Group B select another response. In this case, if a combined-group scoring key is used, members of the larger group will necessarily obtain a higher score than members of the smaller group. Such extreme group differences are unlikely on most ability tests, but might occur in the assessment of values, beliefs, and interests. Researchers should therefore check their data for extreme group differences on those ability test items where responses might be influenced by values as well as knowledge.

Summary

Consensus scoring is an attractive method of creating a scoring key for ability items when traditional methods do not provide clear-cut answers. Many variations of consensus scoring exist. Of these, mode consensus scoring and proportion consensus scoring are the most widely used and the most psychometrically promising. However, mode consensus scoring is inherently biased against smaller groups: smaller groups will on average obtain lower scores than they would have obtained had they made up a larger proportion of the norm group, and will also obtain lower average scores than larger groups when the different groups have the same average level of knowledge. Neither within-group norming nor the use of experts to create the scoring key eliminates this bias. Therefore, mode consensus scoring should not be used to make decisions about individuals or groups. In contrast, proportion consensus scoring does not usually create bias against smaller groups, although it may do so in some extreme situations. Thus, at this time, proportion consensus scoring appears to be the consensus scoring method of choice.

Author note

We thank Glenn Geher, Michael Harwell, Peter Legree, John D. Mayer, Larry Pace, Richard D. Roberts, Takashi Yamashita and an anonymous reviewer for their comments on earlier drafts of this paper.

Notes

[1] Consensus scoring has been successfully used to score tests of social knowledge (Legree, 1995), emotional intelligence (Legree, Psotka, Tremble, & Bourne, in press; Mayer, Caruso, & Salovey, 2000; Mayer, Salovey, & Caruso, 1999; Mayer, Salovey, Caruso, & Sitarenios, 2003; Zeidner, Shani-Zinovich, Matthews, & Roberts, 2005), emotion perception (Geher, Warner, & Brown, 2001; Mayer, DiPaolo, & Salovey, 2000), driving knowledge (Legree, Martin, & Psotka, 2000), general cognitive ability (Legree et al., 2000), supervisory skills in non-commissioned officers (Heffner & Porr, 2000), and military leadership (Hedlund, Forsythe, Horvarth, Williams, Snook, & Sternberg, 2003).

[2] Let us consider two examples. Imagine a respondent who agrees with 50% of the population on each of 10 items. This person's true score is .50. Using the formula for the standard error of the mean, it can be shown that with norm samples of 100, 95% of the time, proportion consensus scoring will result in an average score over the ten items for this person that is between .47 and .53. We consider this a sufficient degree of accuracy for the scoring key. A norm sample of only 50 people is also adequate when the true score is .50: for 95% of the norm samples, the observed score will be between .46 and .54. On the other hand, adequate scoring for respondents at the extreme ends of the distribution (who are most likely to be of interest to test users) requires a larger norm group. Consider a respondent who agrees with 10% of the population on each of 10 items. This person's true score on the test is .10. With a norm sample size of 100, 95% of the time, averaging over the 10 items, this person will obtain a score between .08 and .12, using proportion consensus scoring. However, with a norm sample of only 50, sometimes the observed score is almost twice as large as other times: for 95% of the norm samples, the observed score will be between .07 and .13.

To repeat these calculations with mode consensus scoring, we would first need to calculate the probability, for a given sample size, of correctly identifying the population mode of a single item. However, this probability will be dependent upon the distribution of responses across all response options, not just the proportion of the norm group who select the modal response. For example, the probability of correctly identifying the mode will be lower for difficult items that contain an option that is selected almost as frequently as the mode (e.g., .40, .35, .15, .10) than it will be for easy items where all non-modal responses are endorsed with similar frequencies (.40, .20, .20, .20). We were unable to locate either formulas or algorithms to calculate the probability of correctly identifying the population mode. It is possible that no such formula exists, given that statisticians (even those focusing on nonparametric statistics) use the mode very rarely. Thus, at present, no general statements can be made about adequate sample sizes for mode consensus scoring.

[3] On the other hand, if subscales consist of many items, or subscales are combined to create a total test score, norm samples of much less than 100 may be sufficient. Mayer et al., (1999) and Mayer et al., (2003), for example, found very promising results for MSCEIT total scores with a sample of only 21 experts, using proportion consensus scoring.

References

Bank, L., Dishion, T.J., Skinner, M., & Patternson, G.R. (1990). Method variance in structural equation modeling: living with «glop». In G.R. Patterson (ed.): *Depression and aggression in family interaction. Advances in family research* (pp. 247-279). Hillsdale, NJ: Lawrence Erlbaum.

Brackett, M. & Salovey, P. (2006). Measuring emotional intelligence with the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT*). Psicothema, 18,* supl., 34-41.

Ekman, P. (1972). Universals and cultural differences in facial expression of emotion. In J.R. Cole (ed.): *Nebraska Symposium on Motivation,* 1971. Lincoln, Nebraska: University of Nebraska Press.

Fridlund, A. (1994). *Human facial expression: an evolutionary view*. New York: Academic.

Geher, G., Warner, R.M., & Brown, A.S. (2001). Predictive validity of the emotional accuracy research scale. *Intelligence, 29*, 373-388.

Hedlund, J., Forsythe, G.B., Horvarth, J.A., Williams, W.M., Snook, S., & Sternberg, R.J. (2003). Identifying and assessing tacit knowledge: understanding the practical intelligence of military leaders. *Leadership Quarterly, 14,* 117-140.

Heffner, T.S. & Porr, W.B. (2000, August). *Scoring situational judgment tests: a comparison of multiple standards using scenario response alternatives.* Paper presented at the Annual Conference of the American Psychological Association, Washington, DC.

Legree, P.J. (1995). Evidence for an oblique social intelligence factor established with a Likert-base testing procedures. *Intelligence, 21*, 247-266.

Legree, P.J., Martin, D.E., & Psotka, J. (2000). Measuring cognitive aptitude using unobstrusive knowledge tests: a new survey technology. *Intelligence, 28,* 291-308.

Legree, P.J., Psotka, J., Tremble, T., & Bourne, D.R. (2004). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R.D. Roberts (eds.): *International Handbook of Emotional Intelligence* (pp. 99-123). Seattle, WA: Hogrefe and Huber.

Levin, J. (1973). Bifactor analysis of a multitrait-multimethod matrix of leadership criteria in small groups. *Journal of Social Psychology, 89,* 295-299.

MacCann, C., Roberts, R.D., Matthews, G., & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance-based

Emotional Intelligence (EI) tests. *Personality and Individual Differences, 36*, 645-662.

Mayer, J.D., Caruso, D.R., & Salovey, P. (2000). Emotional intelligence meets traditional standards for an intelligence. *Intelligence, 27,* 267-298.

Mayer, J.D., DiPaolo, M., & Salovey, P. (2000). Perceiving affective content in ambiguous visual stimuli: a component of emotional intelligence. *Journal of Personality Assessment, 54,* 772-781.

Mayer, J.D. & Geher, G. (1996). Emotional intelligence and the identification of emotion. *Intelligence, 22,* 89-113.

Mayer, J.D., Salovey, P., & Caruso, D.R. (1999). *Instruction manual for the MSCEIT Mayer-Salovey-Caruso Emotional Intelligence Test, research version 1.1.* Toronto, ON: Multi-Health Systems.

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

Mayer, J.D., Salovey, P., Caruso, D.R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*, 97-105.

O'Sullivan, M. & Ekman, P. (2004). Facial expression recognition and emotional intelligence. In G. Geher (ed.): *Measuring emotional intelligence* (pp. 89-109). Hauppauge, NY: Nova Science Publishers.

Zeidner, M., Shani-Zinovich, I., Matthews, G., & Roberts, R.D. (2005). Assessing emotional intelligence in gifted and non-gifted high school students: outcomes depend on the measure, *Intelligence, 33,* 369-391.