Psychometrics and the Measurement of Emotional Intelligence

Gilles E. Gignac

It may be suggested that the measurement of emotional intelligence (EI) has been met with a non-negligible amount of scepticism and criticism within academia, with some commentators suggesting that the area has suffered from a general lack of psychometric and statistical rigour (Brody, 2004). To potentially help ameliorate this noted lack of sophistication, as well as to facilitate an understanding of many of the research strategies and findings reported in the various chapters of this book, this chapter will describe and elucidate several of the primary psychometric considerations in the evaluation of an inventory or test purported to measure a particular attribute or construct. To this effect, two central elements of psychometrics, reliability and validity, will be discussed in detail. Rather than assert a position as to whether the scores derived from putative measures of EI may or may not be associated with adequate levels of reliability and/or validity, this chapter will focus primarily on the description of contemporary approaches to the assessment of reliability and validity. However, in many cases, comments specifically relevant to the area of EI will be made within the context of reliability and/or validity assessment.

Test Score Reliability

Introduction

Overwhelmingly, the concept of reliability in psychology tends to be interpreted within the context of composite scores. In practice, a composite score usually consists of an aggregation of equally weighted smaller unit scores, where those unit scores are typically derived from item responses or subtest scores within an

G.E. Gignac (✉)
School of Psychology, University of Western Australia, Crawley, WA 6009, Australia;
Director of Research & Development, Genos Pty Ltd, Suite 1.11/365 Little Collins Street, Melbourne VIC 3000, Australia
e-mail: gilles.gignac@genos.com.au

C. Stough et al. (eds.), *Assessing Emotional Intelligence*, The Springer Series on Human Exceptionality, DOI 10.1007/978-0-387-88370-0_2,
© Springer Science+Business Media, LLC 2009

inventory. While any group of scores can technically be aggregated to form a composite score, a psychometrically defensible composite will be associated with item/subtest scores that exhibit a particular level of "inter-connectedness". Throughout the history of psychometrics, various concepts and methods have been formulated to represent and estimate the degree of inter-connectedness between the corresponding item scores.

While the various methods of reliability estimation are associated with conspicuous differences, all forms of test score reliability may be argued to be based on the notion of repeated measurements (Brennan, 2001). In its purest Classical Test Theory (CTT) form, the reliability of measurement represents the hypothetical distribution of scores expected from repeated measurements derived from the same individual, under the pretence that the individual's memory of the previous testing session is erased (from this perspective, the notion of test score reliability may be considered to be based on a "thought experiment"; Borsboom, 2005). The wider the distribution of scores (i.e., the larger the standard deviation), the less reliability one would ascribe to the scores as an indicator of a particular dimension or attribute. As the prospect of erasing the minds of individuals is not exactly practical, various other methods of estimating reliability have been devised to approximate the scores that would be expected to be derived from the "thought experiment". From this perspective, the most well-known are "parallel forms reliability" and "test–retest reliability". Within the context of reliability estimation via a single-testing session the most well-known reliability methods are "split-half reliability" and "Cronbach's alpha" (α). Less well-known methods of estimating internal consistency reliability are based directly upon latent variable model solutions. The most well-established method of estimating the internal consistency reliability of a composite score via a latent variable model solution is known as "McDonald's omega" (ω).

Prior to describing the above methods of reliability estimation in detail, it should be emphasized that reliability should not be viewed as a property of a test, per se. Instead, reliability should be interpreted as a property of scores derived from a test within a particular sample (Thompson & Vacha-Haase, 2000). This issue is not merely semantic, as the implications are directly relevant to the practice of testing and measurement in psychology. Specifically, because reliability is not a property of a test, researchers can not rely upon previous estimates of reliability to support the use of a test in their own work. Consequently, researchers are responsible for estimating and reporting the reliability of their scores based on their own data. The possibility that a particular test will yield scores of a particular level of reliability across samples and settings is a hypothesis to be tested, rather than an assumption to be made. The generalizability of a reliability estimate may be tested within a "reliability generalization" framework, a concept and method which will not be described in any further detail in this chapter (interested readers may consult Shavelson, Webb, & Rowley, 1989, for an accessible discussion of reliability generalization).

Types of Reliability Estimation

Parallel Forms Reliability

In contemporary psychometric practice, parallel forms reliability (a.k.a., alternative forms reliability) is rarely reported, despite contentions that it may be the most fundamentally sound method of estimating reliability (e.g., Brennan, 2001). Parallel forms reliability is based on the premise of creating two tests or two inventories which yield composite scores associated with the same parameters (i.e., means and variances) and are justifiably regarded to measure the same construct. In practice, participants would complete form A and form B during two different testing sessions separated by approximately two weeks (Nunnally & Bernstein, 1994). The squared correlation between the composite scores obtained from the two forms would represent an estimate of reliability of the scores derived from the inventories, individually.

The methodology of parallel forms reliability can be applied in a such a way as to offer the opportunity to identify three sources of "error" variance: (1) systematic in item content between tests (which, realistically, is expected because items are not random samples drawn from a population of items); (2) systematic differences in scoring (more common in scenarios where a rating is made by a test administrator); and (3) systematic changes in the actual attribute of interest (Nunnally & Bernstein, 1994). Thus, the capacity to segregate these three sources of measurement error via parallel forms reliability may be viewed as particularly valuable. However, the procedure is rarely observed in the applied literature. To my knowledge, there has yet to be a published instance of parallel forms reliability in the emotional intelligence literature. Thus, the temporal variation in EI (source #3) as distinct from "pure" measurement error has yet to be determined. Perhaps the primary reason why parallel forms reliability is so rarely reported in the applied literature is due to the difficulties of creating a second parallel test with the same mean and variance characteristics as the first test, not to mention the same validity. A less onerous reliability procedure that may (justifiably or unjustifiably) be viewed as sharing some properties of parallel forms reliability is known as test–retest reliability.

Test–Retest Reliability

Rather than create two separate forms considered to measure the same attribute and have participants respond to the separate forms at two different testing sessions (i.e., parallel forms reliability), an alternative reliability methodology consists of creating a single test and having participants respond to the items at two different points in time. The correlation between the corresponding time 1

and time 2 scores represents a type of reliability methodology known as "test–retest reliability". Test–retest reliability is indicated when the correlation between the scores is positive, although no widely acknowledged guidelines for interpretation appear to exist.

In its purest form, the premise of test–retest reliability may still be considered predicated upon the Classical Test Theory notion of a "thought experiment" (see, Borsboom, 2005), as the participants are assumed to have largely forgotten the questions and responses once the second testing session takes place. Such an assumption may be plausibly challenged, however, particularly given that the time interval between testing sessions may be as little as two weeks. For this reason, the utility of the test–retest method as an indicator of measurement error has been seriously challenged (e.g., Nunnally & Bernstein, 1994). Despite these criticisms, the use of the test–retest method appears to continue unabated in most disciplines in psychology, including EI. It remains to be determined what reliability related information may be drawn from this type of research.

Despite the problems associated with the interpretation of a test–retest reliability coefficient as an indicator of reliability, the observation of "stability" (as the method of test–retest reliability is often preferentially called, e.g., Matarazzo & Herman, 1984) in trait scores across time may be suggested to be important in practice. That is, if peoples' level of EI is shown to fluctuate widely across time (in the absence of any systematic treatment effects), it is doubtful that the scores could ever be found to correlate with any external attribute of interest that would be expected to be relative stable (e.g., well-being, job performance, etc.). Thus, although the supposed importance of test–retest reliability may be questioned, the importance of test–retest stability can probably not. Consequently, an examination of test–retest stability should nonetheless be considered when evaluating the scores of a psychometric inventory.

Internal Consistency Reliability

In contrast to parallel forms reliability and test–retest reliability, internal consistency reliability can be conceptualized and estimated within the context of a single administration of a single set of test items. Consequently, it is much more convenient to estimate, which may explain its popularity. The two most popular methods of estimating internal consistency reliability are the split-half method and Cronbach's alpha (α). A more sophisticated approach to internal consistency reliability has also been established within a latent variable framework, known as McDonald's omega (ω), which is beginning to gain some popularity, as it is more flexible in accommodating data that do not satisfy the rather strict assumptions associated with Cronbach's α.

Split-Half Reliability

Split-half reliability may be the simplest method of internal consistency estimation. In effect, a particular inventory is split into two halves and the summed scores from those two halves are correlated with each other. The correlation between the two summed halves may be considered conceptually equivalent to the correlation between two parallel forms. However, the correlation between the two halves would be expected to underestimate the reliability of the scores derived from the entire test. Consequently, split-half reliability is often formulated as (Nunnally & Bernstein, 1994):

$$r_{kk} \frac{2r_{12}}{1 + r_{12}}$$

where r_{kk} = the reliability of the whole test and r_{12} = the correlation between two half-tests. Thus, the greater the positive correlation between the two halves, the greater the reliability estimate.

The widely acknowledged problem with the split-half method to the estimation of reliability is that one is required to determine how the inventory will be split into two separate halves. Thus, while most investigators tend to split a scale into halves of odd and even items, there is no compelling qualitative or quantitative reason for doing so. Other seemingly justifiable splitting methods are easily conceived, which have been demonstrated to yield different estimates of reliability (Brownell, 1933).

Another problem with the split-half reliability method is pertinent to time-limited tests. Specifically, time-limited tests based on items that are ordered in terms of difficulty tend to yield upwardly biased estimates of reliability, as the correlation between particular halves may be found to be higher than would otherwise be expected had the items been administered individually (Cronbach, 1960). Given these limitations, a generalization of the split-half method has been devised, known as Cronbach's α, which represents the average reliability of all possible split-halves (Cronbach, 1951).[1]

Cronbach's Alpha (α)

Cronbach's α is the most popular approach to the estimation of internal consistency reliability (Peterson, 1994). It is typically considered to range between .00 and 1.0; however, estimates can technically be negative in the

[1] Occasionally read in the contemporary literature is internal consistency reliability based on the Kuder–Richardson 20 formula. The KR20 reliability procedure predates Cronbach's α, but was limited to dichotomously scored items from which "proportion correct" and "proportion incorrect" information could be derived for each item. When the items are of equal difficulty, a more simplified formulation can be used to estimate reliability (i.e., KR21).

event the covariances between the items are, on average, negative in direction. A relatively accessible formulation of Cronbach's α is:

$$\alpha = \frac{k^2 * \overline{COV}}{\sum S^2, COV}$$

where k = number of items used to calculate the composite score, $\overline{COV}$ = mean inter-item covariance, and $\sum S^2, COV$ = the sum of the square variance/covariance matrix (Cortina, 1993). Based on the formula above, it is apparent that reliability will increase as a function of two parameters: the number of items included in the analysis; and (2) the magnitude of the average positive association between the items. The numerator term of the formula represents the "true score variance", while the denominator represents total variance. For this reason, reliability may be referred to as the ratio of true score variance to total variance (Lord & Novick, 1968).

It should be emphasized that the correct application of Cronbach's α is based on the observation of three assumptions (Lord & Novick, 1968), which appear to be only rarely acknowledged in the literature. There is also evidence to suggest that two of the three assumptions are probably rarely satisfied in practice (Gignac, Bates, & Lang, 2007).

Firstly, it is assumed that the error variance associated with each item is not correlated with the true score variance term. This assumption is primarily of theoretical interest, and may be expected to be satisfied if the remaining two assumptions are also satisfied.

The second assumption states that each item must contribute an equal amount of variance to the true score variance term. Technically, this assumption is referred to as tau-equivalence. Effectively, the second assumption implies that the single-factor model underlying the covariances between the items is associated with a factor solution with equally sized factor loadings. While this assumption may be expected to be rarely achieved in practice, the consequence of violating the tau-equivalence assumption is not usually very consequential (Reuterberg & Gustafsson, 1992). When the tau-equivalence assumption is not satisfied, Cronbach's α will tend to underestimate the true reliability of the scores. For this reason, Cronbach's α is sometimes referred to as a "lower-bound estimate of reliability" (Novick & Lewis, 1967). The tau-equivalence assumption may be tested within a confirmatory factor analytic model, where the factor loadings are constrained to equality. It is probably safe to say that the assumption of tau-equivalence has never been tested on the items of any scale in the EI literature.

The third assumption states that the error terms ("residuals") associated with all of the items must not correlate with each other. In effect, this assumption may be said to be observed when the items conform to a well-fitting, single-factor model, as tested via confirmatory factor analysis (CFA). In the event that the single-factor model is not found to be well-fitting, which would imply some correlations between the residuals, the Cronbach's α estimate will be upwardly biased if the sum of the correlated residuals is positive (Raykov, 2001).

Fortunately, there is a formula that can calculate accurate estimates of internal consistency reliability based on data which are neither tau-equivalent nor consistent with the absence of correlated error terms. This equation has been formulated within the context of latent variable modeling and is known as McDonald's omega (ω).

MacDonald's Omega (ω)

The most popular method used to estimate the internal consistency reliability of composite scores within a factor analytic or latent variable framework is known as omega (ω), which was first formulated by McDonald (1970). For the purposes of this chapter, two omegas will be distinguished: ω_{A} and ω_{B}. The first omega (ω_{A}) is to be used in that case where there are no correlations between the error terms of the items, but where the factor loadings may not be equal across all items. In accordance with Hancock and Muller (2001), it may be formulated as:

$$\omega_{\mathrm{A}} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \delta_{ii}}$$

where λ_i = standardized factor loading and δ_{ii} = standardized error variance (i.e., $1–\lambda_i^2$). In contrast, the second omega (ω_{B}) is to be used in that case where there are correlations between the error terms of the items (Raykov, 2001). It may be formulated as:

$$\omega_{\mathrm{B}} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \delta_{ii} + 2 \sum_{1 \le i < j \le k} \delta_{ij}}$$

where λ_i and δ_{ii} are defined as above andδ_{ij} is equal to the correlations between item error terms. Cronbach's α and ω_{A} will yield the same reliability estimates only in the case where the underlying single-factor model solution is tau-equivalent and well-fitting. As the assumptions associated with Cronbach's α are likely not met in practice, it may be argued that the latent variable approach to the estimation of internal consistency reliability is decidedly the most appropriate. The explicit emphasis on first achieving a well-fitting model in CFA may also be considered an advantage. Despite the fact that the latent variable approach to the estimation of reliability has been well established for several years, only a small number of empirical investigations appear to have ever estimated ω_{A} or ω_{B} in psychology, only one of which may be related to EI (i.e., Gignac, Palmer, & Stough, 2007).

In the area of personality, Gignac, Bates, et al. (2007) demonstrated that internal consistency reliabilities associated with the subscales of the NEO-FFI were all overestimated when the correlations between the item residuals were ignored. For example, the Extraversion subscale scores were initially found to be associated with an internal consistency reliability of .83, which dropped to .77 when the correlated residuals were included into the ω_B formulation. Across all five personality dimensions, the non-biased internal consistency reliability estimates were all demonstrated to be below .80 (Gignac, Bates, et al., 2007), which is the recommended minimum for basic research (Nunnally & Bernstein, 1994).

The assumption that none of the items may correlate with each other, independently of the factor they are all hypothesized to measure, in effect states that the single-factor model underlying the plausibility of the scale must be demonstrated to be associated with adequate levels of model-fit when tested via CFA. Researchers in the area of EI (and psychology in general) appear only very rarely to first test the plausibility of the single-factor model for their data. Thus, the internal consistency reliability estimates reported in the EI literature should probably be viewed as upwardly biased.

Methods for estimating the reliability of composite scores that are derived from items that form part of a multi-dimensional model are being developed based on the same latent variable framework described above. For example, Gignac, Palmer, et al. (2007) applied ω_A to a well-fitting direct hierarchical factor model (a.k.a., nested factor model) solution based on the 20 items of the Toronto Alexithymia Scale-20 (TAS-20). The TAS-20 has been reported to measure three inter-correlated subscales, corresponding to Difficulty Identifying Feelings (DIF), Difficulty Describing Feelings (DDF), and Externally Oriented Thinking (EOT; Parker, Bagby, Taylor, Endler, & Schmitz, 1993). Based on Cronbach's α, the reliabilities for the subscale scores were estimated at .83, .81, and .65, respectively, which corresponded closely to what other investigators have reported. In contrast, the ω_A estimates derived from the well-fitting, multi-factor model solution were .56, .31, and .42 respectively. These reliability estimates are so low as to question the utility of the subscale composite scores as unique indicators of the dimensions they are purported to measure. It should be noted that the TAS-20 is not expected to be the only inventory to exhibit very low levels of unique reliability at the subscale level, when estimated from a direct hierarchical model. In fact, based on preliminary analyses (available from the author), the index scores within the WAIS-III have also been found to be associated with ω_A estimates less than .50.

Appreciating the Importance of Reliability

Several reliability estimation methods have been described above, with McDonald's ω_A and/or ω_B endorsed as likely the most appropriate method to employ in most cases. Next, an appreciation for achieving respectable levels

of reliability will be discussed based on two concepts: (1) the standard error of measurement, which will emphasize the implications of reliability for the meaningful interpretation of scores, and (2) the effects of imperfect reliability on the estimation of effect sizes in empirical studies, which will emphasize the importance of disattenuating effect sizes such as Pearson correlations and Cohen's *d*, prior to interpreting the practical significance of the results reported in an empirical study.

Standard Error of Measurement

A relatively accessible approach to the understanding of the importance of reliability may be achieved by considering the estimation of 95% confidence intervals for a particular score derived from a test. Suppose a particular intelligence test yielded scores associated with an internal consistency reliability of .70, the commonly cited demarcation criterion for minimum acceptability (Peterson, 1994). Suppose further that the mean of the scores was 100 and the SD was 15. Based on the standard error of measurement formula, one could estimate with 95% confidence the range of intelligence observed scores a person would yield under the pretence that they were tested a theoretical infinite number of times (assuming their memory was erased; or, alternatively, were administered an infinite number of parallel tests). The formula for the standard error of measurement of an observed score is:

$$\text{SEM} = \text{SD}\sqrt{1 - r_{xx}}$$

where SD = standard deviation and r_{xx} = reliability. Thus, based on a standard deviation of 15 and a reliability estimate of .70, the standard error of measurement of the observed score in this case is equal to 8.22 IQ points:

$$\text{SEM} = 15\sqrt{1 - .70}$$
$$\text{SEM} = 8.22$$

.

However, the estimate of 8.22 corresponds only to 68% of the distribution of infinite observed scores (i.e., one standard deviation above and below the point-estimate of 100). To estimate the 95% confidence interval (CI95%), the SEM must be multiplied by 1.96, which corresponds to 95% of the *z* distribution. Therefore, $\text{IQ}_{\text{CI95\%}} = 1.96 * 8.22 = 16.11$. Thus, a composite IQ score of 100, which is derived from a sample associated with a reliability estimate of .70, may be expected to range between 83.89 and 116.11, i.e., somewhere between dull and bright. Such a range should be viewed unimpressively, as the conclusion that an individual's observed score is somewhere between dull and bright can be said about nearly anyone, without even having any knowledge relevant to their completed IQ test.

Clearly, a meaningful interpretation of a point-estimate score depends upon the confidence with which an individual's observed score is represented by the point-estimate yielded from the psychometric testing. As the reliability of the scores decreases, the range in interval estimates increases. Based on the example provided above, it would be understandable if someone viewed a reliability estimate of .70 in a critical manner. Further discussion relevant to recommended standards of reliability is provided below.

While the consequences of low levels of reliability can be appreciated relatively clearly when presented in the context of the standard error of measurement, a consequence of low levels of reliability that is perhaps neglected in the literature is relevant to the issue of score meaningfulness and/or interpretability. As stated above, internal consistency reliability represents the percentage of variance in the observed scores that is true score variance. In effect, however, researchers can only be expected to formulate an understanding of the nature of an aggregated score based on how the scores relate to the actual behaviours/cognitions that yielded them. Thus, if there is a substantial difference between observed scores and the true scores, then it is difficult to imagine how an accurate understanding of the nature of the composite scores can be generated. Stated alternatively, if reliability is low, it is difficult, if not impossible, to discern which elements of the behaviours or cognitions are the contributors to the true score variance associated with the composite scores. It is likely for this reason that Cliff (1983) asserted that latent variables defined by only three indicators with factor loadings of .70 or less remain very ambiguous. As factor analysis and internal consistency reliability have direct corollaries, Cliff's (1984) assertion may be viewed as relevant to any group of composite scores associated with a reliability of .75 or less.[2] Thus, both applied psychometrists and pure researchers should have an interest in achieving respectable standards for reliability with their scores, irrespective of the fact that pure researchers may have legitimate recourse to statistical analyses that can disattenuate the effects of imperfect reliability (see next section). Ultimately, in the absence of meaningfulness, it is difficult to discern the utility of reporting an effect between two variables.

Reliability and Effect Size

While reliability and validity are generally regarded as separate psychometric concepts, they are inextricably interwoven, because typical validity studies report effect sizes (e.g., correlations, Cohen's *d*, etc.) to support arguments of validity.[3] That is, it is well established that the magnitude of effect sizes will be

[2] $\omega_A = \frac{\sum (.70 + .70 + .70)^2}{\sum (.70 + .70 + .70)^2 + (.51 + .51 + .51)} = \frac{2.10^2}{2.10^2 + 1.53} = \frac{4.41}{5.94} = .74$

[3] Of course, the interface between validity and reliability is further blurred by the close correspondence between factorial validity and internal consistency reliability.

attenuated as a function of the reliabilities of the scores associated with the variables included in the analysis (Baugh, 2002).

Consider, for example, an estimated correlation of .35 between a predictor and criterion associated with reliabilities of .70 and .80, respectively. The maximum correlation that can be observed between these predictor and criterion scores is not |1.0|. Instead, the maximum correlation is equal to:

$$\begin{aligned} r_{\max} &= \sqrt{r_{xx} r_{yy}} \\ r_{\max} &= \sqrt{(.70)(.80)} \\ r_{\max} &= \sqrt{.56} \\ r_{\max} &= .75 \end{aligned}.$$

Consequently, it has been recommended that observed correlations be disattenuated for the purposes of interpretation, which is achieved by dividing the observed correlation by the corresponding maximum correlation (Nunnally & Bernstein, 1994). Thus, for this example,

$$r' = \frac{r_{\text{obs}}}{r_{\max}} = \frac{.35}{.75} = .47.$$

Some debate has surrounded the appropriateness of the "double correction", i.e., where the reliabilities of both the predictor scores and the criterion scores are used to disattenuate the effect size (see Muchinsky, 1996). It has been recommended that in applied research contexts (e.g., personnel selection, clinical diagnosis, etc.) only the single correction be used, where the reliability of only the criterion is taken into consideration. The reason why such a recommendation is sensible is based on the fact that individuals and/or organizations, in practice, can operate only at the level of predictor scores composed of both true score variance plus error variance. That is, they do not have access to true scores – only observed scores. Consider a human resource department that has estimated a group of prospective employees' IQ. That department is left with using the IQ scores derived from the IQ testing, which will be invariably contaminated by measurement error. Thus, there is no recourse to disattenuation formula in applied settings (with respect to predictor variables).

It will be noted that the disattenuation effects observed with correlations can also be observed for effect sizes based on mean differences (e.g., Cohen's *d*). The reliability levels of the scores do not have an effect on the point estimates of the means. Instead, reliability has an effect on the corresponding standard deviations. Specifically, when scores are less than perfectly reliable, the standard deviations tend to be larger than would otherwise be the case. Thus, the

denominator of the Cohen's *d* formula needs to be corrected to obtain the dissatenuated standardized mean difference:

$$d' = \frac{\overline{X}_1 - \overline{X}_2}{SD_{\text{pooled}}(\sqrt{r_{xx}})}.$$

It should also be made clear that the effect sizes reported in structural equation modelling studies are "automatically" disattenuated for imperfect reliability based on the same Classical Test Theory principle applied above (Fan, 2003). Thus, ostensibly large effect sizes reported in a SEM study may have been "achieved" based on latent variables with underlying composite score reliability of questionable respectability (Cohen, Cohen, Teresi, Marchi, & Velez, 1990; Gignac, 2007).

Given the above, validity research reported in EI empirical studies should be evaluated within the context of the reliability of the scores used in the analysis and whether the reported effects have been disattenuated or not. In the context of convergent validity, the absence of an effect may be due to unacceptably low levels of reliability, rather than absence of a true effect. Conversely, evidence in favour of discriminant validity should also be interpreted within the context of the reliability of the scores, as unacceptably low levels of reliability may dictate the absence of an association between two variables.

Recommended Standards for Reliability

The most frequently cited recommendation for minimum levels of internal consistency reliability is .70 (Peterson, 1994). From a reliability index perspective, a recommendation of .70 corresponds to a correlation of .84 between observed scores and true scores.[4] Consequently, a recommendation of .70 would suggest that a minimum of 70% of the observed score variance must be true score variance. However, it is at best misleading to suggest that Nunnally (1978) recommended .70 as a minimum demarcation criterion for internal consistency reliability. As pointed out by Lance, Butts, and Michels (2006), Nunnally (1978) recommended .70 only for early stage research. For basic research, Nunnally (1978) recommended a criterion of .80, and for clinical decision making a minium reliability level of .90+ was encouraged. Given the relatively early stage at which emotional intelligence research may be regarded, a minimum reliability criterion of .70 may be considered acceptable. However, if emotional intelligence research is to be considered basic research rather than exploratory, efforts should be made to improve the reliability of the scores so as

[4] Some investigators have erroneously equated true scores with constructs scores (e.g., Schmidt & Hunter, 1999). It should be noted that scores devoid of measurement error (i.e., true scores) are not necessarily scores associated with any construct validity (Borsboom & Mellenbergh, 2002).

to ensure levels above .80. The application of EI to clinical or workplace settings may be expected to require minimum reliability levels of .90.

It has been observed that self-report measures of emotional intelligence tend to be associated with relatively adequate levels of internal consistency reliability, while ability based subtests of EI, such as the MEIS/MSCEIT, tend to struggle to achieve even minimum standards of reliability for exploratory research (Matthews, Zeidner, & Roberts, 2002). Proponents of the MSCEIT have countered that while the individual subtests of the MSCEIT may suffer from a lack of reliability, the reliabilities do achieve respectable standards at the branch and total score level (Mayer, Salovey, Caruso, & Sitarenios, 2003).

There is reason to seriously question such assertions, however. Consider the magnitude of the inter-subscale correlations within the MSCEIT V2.0 reported in Mayer et al. (2003). For example, the correlation between the Faces and Pictures subscales (which form the Perceiving branch score) was reported to equal .347. Summed together, the Faces and Pictures subscales form the Perceiving branch score. Based on the split-half reliability formula (see above), a correlation of .347 amounts to an internal consistency reliability of .51, which is substantially lower than the reliability estimate of .91 reported for the Perceiving branch by Mayer et al. (2003). What could be the explanation for the substantial difference in the split-half reliability estimate calculated above and the Cronbach's α estimate reported by Mayer et al. (2003)?[5]

It is likely that Mayer et al. calculated their branch level reliability estimates based on correlation matrices that included all of the items of the Faces subscale and all of the items of the Pictures subscales. Such a procedure may seem appropriate; however, it is important to consider the assumptions associated with the estimation of internal consistency reliability, as described in detail above – in particular, the assumption that the item error terms ("residuals") can not correlate with each other. In practice, this assumption implies that the items used to calculate the reliability estimate must conform to a well-fitting, single-factor model as determined by confirmatory factor analysis. Thus, no two items can share unique variance with each other.

To appreciate the importance of the well-fitting, single-factor model assumption, consider that Green, Lissitz, and Mulaik (1977) were able to demonstrate that a composite score associated with data that conformed to a completely orthogonal five-factor model were found to be associated with a Cronbach's α of .81, when reliability was estimated at the item level. In contrast, the corresponding Cronbach's α based on the inter-correlations of the five corresponding subscales would be .00, as the five factors were simulated to be completely orthogonal. Ultimately, Cronbach's α should never be viewed as an index of homogeneity (i.e., unidimensionality). Instead, the accurate application of Cronbach's α assumes the data have already been demonstrated to be consistent

[5] While it is true that split-half reliability estimates and Cronbach's α estimates will not usually yield the same values, such a large discrepancy can not reasonably be expected to be due to the different formulations.

with a well-fitting, single-factor model. To my knowledge, this has never been demonstrated with any of the MSCEIT subscales, not to mention any of the branches or the total scale.

When the internal consistency reliability estimate of .51 was calculated above for the Perceiving branch score based on the split-half reliability formula, the corresponding CFA model in such a case is necessarily impossible to disconfirm, as there are only two "items" (subtests, more precisely) and, consequently, a single correlation. Thus, the assumption of no correlated error terms (or a well-fitting, single-factor model) is necessarily satisfied. In contrast, the estimate of .91 reported by Mayer et al. is assumed to be based on a corresponding CFA model of 50 items (Faces = 20 items; Pictures = 30 items). In the event that any of the 50 items were to correlate positively with each other above and beyond the shared variance accounted for by the global factor, the .91 estimate would be an overestimate. It is suggested, here, that the more accurate estimate of the internal consistency reliability of the Perceiving branch scores is closer to .51 rather than .91. Based on the same split-half reliability procedure used above and the inter-subscale correlations reported in Table 2 of Mayer et al. (2003), the branch level score reliabilities for the Facilitating, Understanding, and Managing branches were estimated at .52, .74, and .73, respectively. Thus, all below the .80 recommendation for basic research (Nunnally & Bernstein, 1994). For an accessible demonstration of the application of CFA in the estimation of internal consistency reliability (ω_A and ω_B), readers are referred to Gignac, Bates, et al. (2007).

Internal Consistency Reliability Versus Test–Retest Reliability

It is important to distinguish parallel forms reliability (or test–retest reliability) from internal consistency by noting that the two types of reliability have no necessary association with each other. That is, composite scores may be demonstrated to be associated with very high levels of parallel forms reliability (or test–retest reliability) but very low levels of internal consistency reliability. Consider the three following variables: height, intelligence, and extraversion. The correlations between these variables measured in adults would all be expected to be less than .20, which would preclude any meaningful aggregation of the scores (and a very low level of internal consistency reliability). However, these same aggregated scores would be expected to have very high levels of test–retest reliability, because all three of the scores would not be expected to change over a time period of, say, 2 weeks. Because the individual variable scores would not be expected to change, the corresponding composite scores would also not be expected to change, resulting in a substantial level of test–retest reliability. Thus, evidence in the EI literature that suggests substantial test–retest reliability for the MEIS or MSCEIT (e.g., Brackett & Mayer, 2003), or any other instrument for that matter, should not be interpreted as in any way indicative of substantial internal consistency reliability.

Validity

Introduction

The process of measurement in psychology is generally consistent with the ascription of numbers to attributes according to a rule (Stevens, 1946). In contrast, the purpose of measurement in psychology is generally to "make inferences from observed test scores to unobserved constructs..." (Sireci, 1998, p. 84). The enterprise of validity research is relevant to the evaluation of the plausibility of those inferences.

As was the case for reliability described above, validity is not a property that can be ascribed to a test, per se. Instead, it is the interpretation of a score derived from a test that may be declared valid. Thus, the scores derived from a test may be considered valid in one case and invalid in another. Ultimately, the user of a test is left with the responsibility of justifying the use of the scores in a particular context. Many approaches to the assessment of validity have been devised over the years, the most common of which are face validity, content validity, factorial validity, predictive validity, incremental predictive validity, concurrent validity, discriminant validity and multitrait–multimethod (MTMM) validity.

Face Validity

Face validity is arguable the least sophisticated approach to the assessment of test validity. It refers to the degree to which, at least superficially (i.e., "on the face of it"), the items within an inventory appear to measure the attribute or construct of interest. Face validity is not typically estimated with a numerical coefficient or index. Consequently, the broader research and test taking community generally makes an informal assessment of the face validity of a test.

Unlike all other forms of validity, high levels of face validity have not categorically been regarded as a positive attribute. For example, Cattell and Warburton (1967) expressed reservations about inventories with high levels of face validity, because they believed that there was a positive correlation between high face validity and the probability of high levels of simulation (i.e., "faking good") on the part of test takers. That is, for the items of a test to be associated with high face validity, the items must be considered to be measuring the attribute of interest in an obvious manner. However, if individuals can easily discern the purpose of an item, they will be in a better position to respond in such a way as to present themselves in an unrealistically positive manner. Thus, argued Cattell and Warburton (1967, p. 35), face validity "defeats the real art of the psychologist, which is to produce the find of test that disguises (from the subject) what it measures." As the items from a typical self-report measure of EI appear to be associated with a high level of face validity, they would be appraised critically from the Cattell/Warburton perspective.

However, it may be beneficial to distinguish between different purposes for administering an emotional intelligence questionnaire or test. Specifically, for the purposes of personnel selection, it may be beneficial for the items of an inventory to be associated with low levels of face validity, as this would militate against the possibility of socially desirable responding. In contrast, if the purpose of the psychometric testing was to provide individuals the opportunity to learn more about themselves, then high levels of face validity may be regarded as advantageous.

Content Validity

Content validity is indicated when the items within a test or inventory may be justifiably contended to be an accurate representation of the entire domain of interest. The concept of content validity initially emerged from educational psychology. Achievement tests, for example, can be evaluated for content validity relatively straightforwardly, as the boundaries of a particular achievement test can be expected to be easily described and agreed upon. For instance, an exam for an undergraduate course in statistics should encompass the material that was covered in the lectures and labs throughout the semester. Questions that are included in an exam that were not covered in the lectures or labs would clearly compromise the content validity of the achievement test. In other cases, the exam may be based on a disproportionately large number of items from only three or four lectures, rather than spread relatively evenly across all lectures/labs. Again, the content validity of the test would be in question in this case.

Outside the domain of achievement testing, the prospect of evaluating content validity is much more difficult, as experts in the field can be expected to disagree on the theoretical boundaries of the construct. The construct of emotional intelligence is certainly no exception. For example, should the attribute of "empathy" be included in the measurement of EI? On what basis might it be included or excluded? Given that there are no widely acknowledged empirical protocols to determine whether a facet should or should not be viewed as within the boundaries of a construct, the issue is generally left to theoretical argumentation. EI is not the only construct in psychology that may have difficulty in specifying clearly its construct related boundaries.

Consider the construct of personality, which has been suggested to encapsulate the ways "individuals differ in their enduring emotional, interpersonal, experiential, attitudinal, and motivational styles" (McCrae & John, 1992, p. 175). On what basis were these domains judged to be the domain of personality rather than another individual differences attribute? If one were to suggest that the area encapsulated by personality may be overexpansive, on what basis may this assertion be supported? The issue of content validity in the area of personality has implications for the possibility that emotional intelligence may be redundant with personality, as suggested by several commentators

(e.g., Landy, 2005). For example, consider the Openness to Experience personality dimension within the NEO PI-R, which incorporates a facet called "openness to feelings". If a substantial correlation between this personality facet and an emotional intelligence subscale were observed, would this indicate construct redundancy for EI? Perhaps it could be countered that "openness to feelings" is better conceived as a facet of EI rather than personality. On what basis might this contention be convincingly refuted? This rather thorny issue relevant to the content validity of personalty and emotional intelligence is not intended to be resolved here. Instead, an appreciation of the nature of the concept of content validity and some of its implications are described.

Readers may have noted that face validity and content validity have been referred to as a property of the test, rather than a property of the scores derived from a test. It is effectively impossible to discuss face and content validity as a property of scores, as they are obviously properties of the items which make-up the test. This issue has not gone unnoticed in the validity literature, and for this reason, some experts have argued that neither face validity nor content validity are truly justifiable elements of psychometric validity (e.g., Guion, 1977).

Factorial Validity

The term "factorial validity" is not frequently observed in the literature, despite the fact that it plays a central role in the validity assessment of scores derived from a measure. As Nunnally (1978, pp. 112–113) contended, "...factor analysis is intimately involved with questions of validity.... Factor analysis is at the heart of the measurement of psychological constructs." Guilford (1946, p. 428) considered factorial validity more central to any other type of validity evidence, as it addressed the question, "'What does this test measure?', rather than, 'Does this test measure what it is supposed to measure?'". While Guilford's (1946) assertion may be criticised, the point to be taken from the above passages is that factorial validity is crucially important to the validity enterprise, as it helps determine what composite scores derived from an inventory measure from a dimensional perspective, or more specifically, how many dimensions are measured by the scores of an inventory? Secondly, by interpreting the loadings within the factor solution (i.e., strength and direction), the nature of those dimensions may be discerned.

The utility of the factorial validity research strategy is relevant to both theory and practice for the domain of EI. Theoretically, the demonstration of factorial validity is important, as the various models of EI postulate the existence of narrower dimensions, such as emotional management and emotional perception, for example, in addition to a global emotional intelligence factor. From a more practical perspective, factorial validity results help an investigator or test developer determine which items should be used to define each subscale.

While the term "factorial validity" is not very frequently used in the literature, factorial validity studies are, in fact, very commonly published in academic journals. There are two primary types of factorial validity studies: exploratory (unrestricted) and confirmatory (restricted). The more impressive factorial validity evidence may be derived from confirmatory factor analysis (CFA), as simple structure is specified and tested statistically within a restricted factor analytic framework. Consequently, largely arbitrary decisions such as rotation (oblique vs. orthogonal) and determining salient vs. non-salient loadings in EFA are obviated in CFA.

Effectively, if a developer of an emotional intelligence inventory asserts that the scores from that inventory measures a global emotional intelligence factor and four subfactors, factor analysis can help test the plausibility of such an assertion. In the event that a CFA model consistent with the developers (or EI model proponents) assertion is found to be associated with adequate model-fit, the test developers' assertion may be considered at least partly supported. The problem of equivalent models and non-equivalent models does seriously undermine assertions of factorial validity (see Tomarken & Waller, 2003), particularly in the context of data that are associated with a relatively strong global factor. In such cases, a large number of different models may be found to be associated with acceptable levels of CFA model-fit. Consequently, factorial validity evidence is not sufficient to justify the use of inventory for any particular purpose. Invariably, factorial validity evidence must be complimented by other types of validity evidence, such as convergent and discriminant validity (described below). However, because factorial validity evidence helps determine how to score an inventory, it must be conducted prior to the assessment of other quantitatively based methods of validity research. Consequently, factorial validity evidence is generally sought prior to convergent or discriminant validity evidence. That is, prior to correlating composites scores derived from the inventory with criteria, the researcher must know how to aggregate the various items together to represent the various factors/constructs.

It will be noted that factorial validity has much in common with internal consistency reliability. This argument may be appreciated from at least two perspectives. First, data for a factorial validity study can be obtained from a single administration. Secondly, as discussed above, sophisticated approaches to the estimation of internal consistency reliability (i.e., ω_A and ω_B) are based on factor solutions. In effect, factorial validity may be regarded to be at the interface between validity and reliability, which seriously undermines the notion that reliability and validity are separate concepts.

Predictive Validity

Predictive validity may be the oldest type of research method to help justify the valid use of scores derived from a measure. Perhaps not coincidently, it is also

based on a straightforward methodology. Specifically, the predictor variable is measured across all participants, then, at a later date (ideally), the criterion variable is measured across all of the same participants. If the correlation coefficient between the predictor and criterion scores is statistically significant (or preferably practically significant), the scores from the measure may be said to be associated with predictive validity. In educational psychology settings, predictive validity studies were devised to determine whether intellectual intelligence tests could predict future academic performance. In such cases, an understanding of the nature of intellectual intelligence was irrelevant. That is, so long as the scores from the measure predicted the criterion of interest, the scores from the measure were deemed valid. For this reason, McGrath (2005) has distinguished between research strategies focused on predictive accuracy versus those that are focused on representational accuracy, as the two are not necessarily compatible.

Typically, predictive validity studies are based on statistical techniques such as correlations or multiple regression, where the percentage of variance accounted for in the dependent variable (e.g., performance) is reported as r^2 or R^2. Precisely how large an r^2 value has to be to indicate predictive validity is a source of some contention. Judgements based purely on statistical significance testing have been thoroughly criticised (see Harlow, Mulaik, & Steiger, 1997), as large sample sizes may be expected to detect miniscule effect sizes. Cohen (1992) has suggested some guidelines for interpreting effect sizes, which have become well-cited in the literature. In the context of correlations, Cohen (1992) suggested that a small, medium, and large correlation may be considered to be equal to .10, .30, and .50, respectively. Cohen's (1992) guidelines were based on his experience of doing research and reading other published articles. In contrast, Hemphill (2003) conducted a quantitative review of meta-analyses to ascertain the range of effect sizes actually reported in the literature. Hemphill (2003) found that the lower third, middle third, and upper third correlation sizes were equal to $<.20$, .20–.30, and $<.30$. Thus, the Hemphill (2003) guidelines differ non-negligibly from Cohen's (1992) guidelines. In particular, a large effect is considered to be a correlation of .30 or greater, rather than a correlation of .50 or greater.

Bold claims have been made about the predictive validity associated with emotional intelligence as a predictor of job performance, many of which have been criticised as outlandish, even by proponents of the EI construct (see Mayer, Salovey, & Caruso, 2000, for example). Further, it has been argued that in the case of emotional intelligence, evidence of *incremental* predictive validity must be demonstrated rather than simply predictive validity (Zeidner, Matthews, & Roberts, 2001).

Incremental Predictive Validity

Incremental predictive validity is a conceptually identical to predictive validity, with the exception that the predictor(s) of interest must demonstrate some

unique capacity in predicting an external criterion (e.g., job performance). Further, the data are generally subjected to hierarchical multiple regression, where a statistically significant increase in percentage of variance accounted for in the dependent variable is expected to be observed based on the addition of the variables of interest to the regression model. As was the case with predictive validity, it is more impressive when the criterion is measured at a time later than when the predictor variables were measured, so as to militate against the possibility of method effects (Spector, 1994). Landy (2005) has argued that the issue of incremental predictive validity is of central importance to the case of emotional intelligence for two reasons. Firstly, parsimony in the number of constructs should be sought in any science; and, secondly, the empirical evidence based on scores derivable from putative emotional intelligence measures have not yet demonstrated incremental predictive validity above well-known measures of personality and/or intellectual intelligence.

One particularly common problem in the EI incremental predictive validity studies is the failure to take into consideration the reliability of the scores included in the regression analysis. Analytical techniques such as multiple regression and path analysis assume that the variables used to represent the attributes or constructs are measured without error (Pedhazur, 1997). Failure to take this assumption into consideration can have serious interpretative problems for multiple regression and path analysis, particularly if the variables have been measured with differential levels of reliability (as would likely be the case in practice). As demonstrated in a section above, the maximum correlation between two composite scores is not necessarily equal to 1.0. Rather, the maximum correlation is equal to the square root of the product of their reliabilities. If the composite scores included in the analysis are associated with different levels of reliability, then the potential effects of independent variables on the dependent variable will be, to some degree, contingent upon the reliabilities of the scores, rather than their true association.

One approach to overcome the reliability issue in multiple regression or path analysis is to disattenuate all of the correlations within the corresponding correlation matrix prior to performing the analysis, based on the same disattenuating procedure described above for a single bivariate correlation. While such a procedure does have some appeal, it is very rarely observed in the literature possibly for several reasons, one of which may be because structural equation modelling (SEM) is a statistical technique that can decompose true score variance from error variance, allowing for the estimation of effect sizes that are not attenuated by measurement error.

However, even those empirical EI studies that have disattenuated the effects obtained in their analyses, either through the classical disattenuation procedure or the more sophisticated SEM approach, tend to be associated with other serious limitations. Perhaps the most common limitation are the measures chosen to represent the control variables. That is, while it is widely acknowledged that emotional intelligence may not be associated with incremental predictive validity due to its shared variance with self-report personality

and/or intellectual intelligence, the studies that test incremental predictive validity hypotheses tend to only use incomplete measures of personality and/or intellectual intelligence (Landy, 2005).

Hunsley and Meyer (2003) have suggested that a semi-partial correlation of .15–.20 be considered supportive of incremental predictive validity. Thus, the predictor(s) must be demonstrated to share a minimum of between 2.3 and 4.0% unique variance with the criterion, independently of the association between emotional intelligence and control variables such as personality and intellectual intelligence. The fact that Hunsley and Meyer (2003) suggested the use of a semi-partial correlation should not go unnoticed, as it is the appropriate analysis in the incremental predictive validity case, despite the fact that most (if not all) of the incremental predictive validity research in the area of emotional intelligence has made use of either partial correlations or multiple regression.

In effect, the absence of a statistically significant beta weight associated with an independent variable within a multiple regression equation does not necessarily preclude the possibility that that independent variable may share some unique variance with a criterion, independently of the shared variance between that independent variable and the control variables, as estimated via a semi-partial correlation. Multiple regression and semi-partial correlation are not the same statistical analyses. When independent variables are entered into a multiple regression, the analysis will attempt to build a regression equation that will maximally predict the dependent variable, based on the estimation of unique beta weights for each independent variable (Pedhazur, 1997). In contrast, a semi-partial correlation is simply a Pearson correlation between a dependent variable and one independent variable that has been residualized from one or more other independent (control) variables. In the semi-partial correlation case, there is no attempt to build a regression equation based on all of the independent variables to maximally predict the dependent variable. Ultimately, the performance of a multiple regression analysis is an implicit or explicit attempt to build a model to represent the associations between a number of independent variables and one dependent variable. In contrast, in no justifiable way can a semi-partial correlation be said to represent a theoretically relevant model.

Incremental predictive validity hypotheses are also frequently operationalized statistically within the context of mediation analyses or partial correlations. Again, just as multiple regression and semi-partial correlation are not the same analysis, mediation via multiple regression and partial correlation are also not the same analysis. A partial correlation is a Pearson correlation between a residualized independent variable and a residualized dependent variable, where the independent variable and the dependent variable have both been residualized against one or more control variables. In contrast, a mediation analysis via multiple regression is a model that combines both bivariate regression and multiple regression, from which indirect and direct effects between the independent variable and the dependent variable can be estimated (Alwin & Hauser, 1975).

The distinctions discussed above between multiple regression, semi-partial correlation, mediation, and partial correlation are not simply pedantic, as the coefficients derived from the analyses will very likely yield at least numerically different results, which may occasionally result in substantively different conclusions (Werts & Watley, 1968). The majority of the incremental predictive validity research in the area of EI appears to make use of multiple regression to determine whether EI can be demonstrated to be associated with incremental predictive validity. Unfortunately, those studies rarely report the corresponding semi-partial correlations to determine EI's unique capacity to correlate with the criterion. Thus, the incremental predictive validity research reported in the EI literature must be interpreted cautiously on this account, as well as the fact that researchers rarely first disattenuate the correlations for imperfect reliability prior to conducting the analyses.

In contrast to multiple regression and Pearson correlations, which assume the independent and dependent variables are measured without error, the data analytic strategy of structural equation modelling (SEM) can accommodate observed variables associated with some level of measurement error, as the observed variables are typically modelled with other observed variables to form latent variables, which are devoid of measurement error (Bollen, 1989). Consequently, the effects (i.e., correlations or beta-weights) obtained between latent variables in SEM are not attenuated due to imperfect reliability.

Only a minority of the incremental predictive validity research in the EI literature to-date has used SEM. While there are clear advantages of using SEM to test empirical hypotheses in the area of individual differences, its application in the area of EI raises an interesting question relevant to the adequacy of models of personality, which may be considered relevant to whether mixed-model measures of EI may justifiably contend to be redundant with self-report measures of personality. This issue will be addressed more fully in the section on discriminant validity (below); however, it will be noted that the implications are equally relevant to the incremental predictive validity research strategy (discussed here).

Concurrent Validity

The concurrent validity research strategy is based on the attempt to demonstrate a theoretically justifiable empirical association between the scores from one measure with those of another measure (or variable, e.g., current salary), where the scores from both measures are collected during the same testing session. The fact that the data are collected during the same testing session is what distinguishes concurrent validity from predictive validity. Collectively, concurrent validity and predictive validity are known as convergent validity or criterion-related validity.

Typical concurrent validity studies in the area of EI consist of collecting the scores from one putative measure of EI with those of another putative measure of EI. This instance of concurrent validity is not regarded highly, as the observation of a positive correlation between the scores of the two inventories does not necessarily imply that either of the inventories yields valid scores of EI, as it is certainly possible that neither of the inventories is useful at measuring EI. Some experts in psychometrics have suggested that the only justifiable instance of correlating two inventories together to demonstrate concurrent validity for the scores of one of the inventories is the case where a short form is correlated with its corresponding long form (e.g., Anastasi, 1996).

More impressive instances of concurrent validity are those where the scores of a putative inventory are correlated with theoretically relevant variables such as age, salary, achievement, performance, mental illness, etc. – that is, scores that are less likely to share method variance with the scores derived from the inventory of interest. In the area of EI, one particularly important instance of concurrent validity that has been argued to be crucial to establish is a positive association between EI scores and age (Mayer, Caruso, & Salovey, 2000). To date, the research that has tested the "age hypothesis" has been decidedly mixed. Even in those instances where a positive correlation is observed, the correlation is not particularly large ($<.15$). An issue that has not been addressed very well in this area is the possibility that emotional intelligence may be exclusively associated with crystallized intelligence (Gc), suggesting that EI and intelligence studies need to model intellectual intelligence via SEM for the purposes of representing unique sources of intelligence. It is possible that ability based EI measures may yield appreciable associations with an orthogonal Gc factor of intellectual intelligence. The failure to measure and model intellectual intelligence as a multi-factor construct is a general limitation to all EI research that has attempted to incorporate intellectual intelligence as either a concurrent validity relevant variable or a control relevant variable.

A problem in the area of EI and validity assessment literature relevant to construct validity is the possibility that EI may be susceptible to the "jingle fallacy", which represents the case where two measures that are purported to measure the same construct are in fact measuring different constructs (Block, 2000). The jingle fallacy may be argued to be relevant to the area of EI, particularly within the context of ability-based model measures of EI and mixed-model measures of EI. That is, measures from both models are often referred to as measures of "emotional intelligence", yet the reported correlations between ability-based model measures and self-report measures is so low as to suggest that they measure largely different attributes. Consequently, while two measures may sound similar ("jingle"), they can nonetheless be demonstrated to measure largely unique sources of variance. Concurrent validity research is especially important to confirming or disconfirming jingle fallacies.

A particularly problematic issue in conducting convergent validity research in psychology is the specification of how large an effect needs to be to support an argument in favour of validity. Rarely do psychology researchers make such

specifications, and the area of EI is certainly no exception. For example, with respect to the hypothesized correlation between age and EI, how large of a correlation would need to be observed to support the hypothesis? Would a correlation of .15 be sufficient? This unresolved issue is especially problematic in light of Meehl's sixth law of soft psychology: "Everything correlates with everything." Thus, it may be contended that the observation of a statistically significant correlation between the scores derived from two measures should not necessarily be viewed as evidence in favour of an argument postulating concurrent validity (e.g., ability-based EI correlating with mixed-model EI). Instead, to be especially compelling, the magnitude of the correlation should be specified and confirmed by empirical results. While exact specifications may be unrealistic in most any area of psychology, suggestions of small, moderate, or large correlations should probably be made.

In the context of ability-based measures and mixed-model measures, it is suggested here that a disattenuated correlation of at least .50 should be observed to support contentions of concurrent validity. An interesting comparison can be made by referring to the intellectual intelligence literature, where several studies have demonstrated a correlation of approximately .30 between ability-based intellectual intelligence and self-reported intellectual intelligence. Based on this research, it is doubtful that ability-based measures of EI and self-report measures of EI will ever be demonstrated to converge sufficiently to support concurrent validity, if mixed-model measures of EI, as measured via self-report, are measuring anything akin to ability EI. Interestingly, intellectual intelligence researchers appear not to interpret the coefficient of .30 as evidence of concurrent validity.

It will be noted that some researchers have acknowledged the distinction between ability-based EI and mixed-model EI, such that mixed-model EI measures are referred to as "trait EI". It is argued here that this does not solve the jingle-fallacy problem, as the acronym EI incorporates the word "intelligence". Ultimately, if a mixed-model or trait-based measure is not clearly measuring a cognitive ability, it serves no credible benefit to refer to the measure in any way as an "intelligence".

Finally, it will be noted that the confirmation of the jingle fallacy in the area of EI does not necessarily imply that self-report measures relevant to emotions are necessarily devoid of any utility or validity. There does remain the plausible possibility that self-report measures relevant to emotions may represent attributes relevant to typical performance rather than maximal performance, which may prove to serve greater utility in applied settings.

Discriminant Validity

Discriminant validity is the opposite of convergent validity. Thus, in contrast to hypothesizing the existence of a correlation between EI scores and a criterion,

discriminant validity would be observed when the scores from an EI inventory are found not to correlate with a criterion that is theoretically postulated to be unrelated to EI. To a non-negligible degree, the establishment of discriminant validity can facilitate an understanding of the nature of the underlying construct, as it can be equally as informative to learn with what the inventory scores do correlate as it is to learn with what they do not correlate. In fact, some methodologically oriented experts have recommended that factor analyses be performed with the inclusion of one variable with which one or more of the factors should not be found to be defined to help understand the nature of the factor(s) (Mulaik, 2007, personal communication).

In addition to facilitating an understanding of the nature of the underlying construct an aggregate of scores may represent, discriminant validity also plays a central role in evaluating "jangle fallacies". As described above, the "jingle fallacy" is observed when the scores from two measures purported to represent the same or similar constructs in fact measure different constructs ("jingle" because the name of the measures is the same). In contrast, the "jangle fallacy" is observed when the scores from two measures purported to represent two different constructs in fact measure the same construct ("jangle" because the name of the measures is different).

The jangle-fallacy may be observed in the case of EI, as the developers of both ability-based measures and mixed-model measure contend that their inventories are not redundant with intellectual intelligence and/or personality. However, while EI inventories may be labelled by their creators with a different name ("emotional intelligence"), the reliable variance derivable from the putative EI inventories may in fact be measuring intellectual intelligence and/or personality.

While the proliferation of constructs and inventories and psychology should probably not be condoned in the absence of any unique construct validity, it may be argued that it is unreasonable to insist that mixed-model measures of EI demonstrate a non-negligible amount of unique construct validity, independently of five factor model of personality, as has been suggested by others (e.g., McCrae, 2000). The justification of this argument is predicated upon the fact that there is yet to be a single CFA study that has demonstrated the plausibility of the five-factor model of personality (or any other theorized model for that matter; see review in Gignac, Bates, et al., 2007). While many attempts have been made, they have invariably failed to be associated with adequate model fit, suggesting that the five-factor model is implausible. In contrast to personality, there have been several publications documenting the plausibility of EI models based on putative measures of EI via CFA (see various chapter of this volume). Consequently, it may be argued to be unjustified to compare measures of EI associated with well-fitting models against popular measures of personality, all of which have been demonstrated to be inadequate via CFA. Surely advocates of personality need to first demonstrate the plausibility of their models, prior to making assertions that other well-fitting models are redundant with personality. Gignac, Bates, et al. (2007) have suggested that the dimensions of the FFM may be excessively complex and over-expansive, as even the individual

dimensions of FFM are associated with poor CFA fit when tested individually. Further, Gignac (2006) obtained a multiple *R* of .93 by regressing the Depression facet from the NEO PI-R onto 11 other NEO PI-R facets, suggesting that at least one of the facets within the NEO PI-R failed to demonstrate discriminant validity from the NEO PI-R! Thus, a substantial amount of psychometric work associated with well-known measures of personality may be required, prior to suggesting these measures as some sort of "gold-standard" with which newly developed measures should be compared.

With respect to ability-based models of EI, contentions have been made that they may lack sufficient discrimination from intellectual intelligence (Landy, 2005). While there are a number of studies that have attempted confirm and/or disconfirm the discriminant validity of the MSCEIT, most, if not all, of these studies suffer from the limited manner in which intellectual intelligence was measured (Landy, 2005).

An example of where such a criticism would apply is the study by Mayer, Caruso, et al. (2000), where an attempt was made to demonstrate that the MEIS was correlated with intelligence, which would support concurrent validity. However, it was also expected that the correlation would not be so strong as to contra-indicate discriminant validity. To test this hypothesis, Mayer, Caruso, et al. (2000) correlated total MEIS (the precursor to the MSCEIT) scores with a single subtest of intelligence (Vocabulary),[6] which yielded a correlation of .36. Obviously, a single subtest is not a comprehensive measure of intelligence. It should not even be considered comprehensive measure of crystallized intelligence. A comprehensive measure of intelligence would require a minimum of 7–9 subtests selected from a diverse group of intelligence sub-factors to allow for the modelling of a general factor, in conjunction with possible sub-factors such as Gc, Gf, and WM (each defined by 3–4 subtests). Consequently, the reported correlation of .36 would certainly be expected to be an underestimate of the true correlation between the MEIS and intellectual intelligence. How large the disattenuated association may be between the MEIS/MSCEIT and an intellectual intelligence battery as estimated via SEM does not yet appear to have been determined.

The Mayer, Caruso, et al. (2000) study may also be criticised for not taking reliability into consideration. That is, observed correlations based on imperfectly reliable scores will be attenuated (see above section on Reliability). For this reason, it may be argued that it is imperative to report disattenuated correlations for the purposes of confirming or disconfirming discriminant validity contentions (as well as convergent validity). Such an effect may be accomplished by using the Classical Test Theory disattenuation formula first proposed by Spearman (1904). Alternatively, the correlation between EI and the criterion may be estimated within a structural equation modelling (SEM)

[6] Technically, the Vocabulary measure consisted of only 60% of the items of a full Vocabulary subtest, as only 30 of the 50 items from the standard Vocabulary subtest were chosen in the Mayer, Caruso, and Salovey (2000) study.

framework, where the measures would be represented by latent variables that are devoid of measurement error (Bollen, 1989). For this reason, the correlations between latent variables are not attenuated by measurement error.

Multitrait-Multimethod (MTMM) Validity

Multitrait-Multimethod (MTMM) validity is not typically referred to as a type of validity, per se. Further, it is not typically applied within the context of attempting to confirm the validity of the scores from a single test or inventory. Instead, MTMM research strategies are more specifically concerned with the possible confirmation or disconfirmation of the plausibility of a construct (Campbell & Fiske, 1959). All validity research may be regarded as series of research strategies that, when the results are interpreted as a whole, may be supportive of the plausibility of a postulated construct (Angoff, 1988). The MTMM research strategy may be applied in such a way as to incorporate all forms of quantitative validity research strategies. Consequently, it may be argued that the enterprise of validity is culminated within the MTMM research strategy.

At its most basic level, MTMM validity is indicated when the scores derived from two different methods of measurement, which are putatively measuring the same construct, are demonstrated to correlate positively (in this basic case, it may be more appropriate to use the term Single Trait-Multimethod (STMM) validity). In the case of emotional intelligence research, MTMM validity would be indicated in the event that an ability-based measure of EI was found to correlate positively with a self-report measure of ability-based EI. A more impressive indication of MTMM validity would be observed in the event that the sub-factors of the ability-based measure of EI were found to have correlated more strongly with the self-report based inventory congruent sub-factors, in comparison to the remaining "heterogenous" sub-factors. Within this context, the sub-factor correlations between factors purported to measure the same trait may be referred to as intra-group correlations. In contrast, the sub-factor correlations between different traits may be referred to as extra-group correlations. The initial methods proposed to evaluate the pattern of relations between measures within a MTMM approach to validity testing suffered from either a lack of statistical significance testing or elegance (or both). A method based on confirmatory factor analysis, however, is both elegant and statistically useful from a validity confirmation or disconfirmation perspective (e.g., Marsh & Byrne, 1993). A fictitious example of a MTMM CFA model created to test the plausibility of an emotional intelligence construct is presented in Fig. 1. It can be observed that there are three "trait" latent variables: Emotional Perception, Emotional Expression, and Emotional Management. Each of the three traits is defined by four measured variables: a self-report measure, an other-report measure, an ability-based measure, and a physiological measure. To account for the expectation that measures derived from the same method

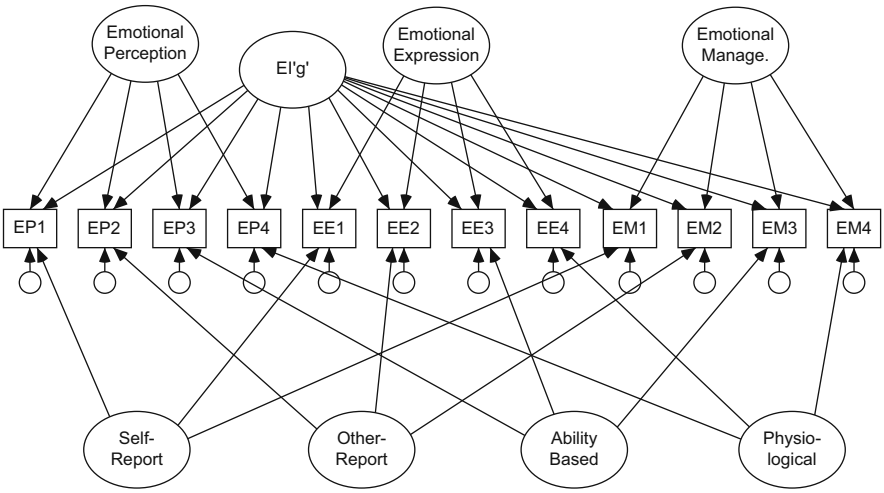

Fig. 1 An example of a MTMM CFA model within the context of emotional intelligence research

would likely share unique variance, an additional four latent variables have been included in the model to account for the methods used to measure the three emotional intelligence traits. In the event that this CFA model were to yield a factor solution with statistically significant positive loadings and adequate model fit, evidence in favour of MTMM validity would be indicated. The model could be expanded to include external criteria variables to test hypotheses relevant to concurrent validity, predictive validity, and discriminant validity.

A Note on the Association Between Reliability and Validity

It has been asserted that of the two primary psychometric properties of scores, validity evidence should be considered more valuable or impressive than evidence for reliability. While this statement will not be challenged here, it should nonetheless be made clear that reliability is a necessary but not sufficient condition for validity. Thus, in the absence of reliability, there is no possibility for validity. Consequently, it may be considered feasible to undertake criterion-related validity research, only once comprehensive reliability and factorial validity assessments have been completed.

Conclusion

In this chapter, the topics of reliability and validity in relation to the evaluation of psychometric scores were introduced and described in some detail. Test–retest reliability was described; however, its actual utility as an indicator

of reliability was questioned, although its value as an indicator of stability was supported. Internal consistency reliability was discussed; however, the application of the ubiquitous Cronbach's alpha was criticized in favour of a more modern reliability estimation technique known as omega, which can accommodate more realistic assumptions associated with data typically obtained in practice. All common forms of validity were discussed, from face validity to MTMM validity. Some problems in the interpretation of discriminant validity were noted, particularly as it relates to whether observed coefficients of .00 must be observed to indicate discriminant validity (within sampling fluctuations), or whether the observation of small correlations may also be indicative of discriminant validity. Otherwise, the testing and interpretation of validity study results appear to be relatively straightforward, although, admittedly, high quality validity studies (e.g., MTMM studies) are difficult to resource and implement in practice.

In conclusion, the contents of this chapter may be considered a relatively comprehensive review of a number of well-established psychometric considerations in the evaluation of scores derivable from a psychometric measure. However, this chapter should not be considered an exhaustive treatment of the area, as topics such as Item Response Theory (IRT), reliability generalization, and differential item functioning, for example, were not treated in any detail, if at all. Readers interested in learning more about psychometrics may consider consulting a classic text such as Nunnally and Bernstein (1994).

It should be made clear that the psychometric principles described in this chapter can by applied to virtually any discipline in psychology that uses psychometric measurements of some description, rather than only the area of emotional intelligence. Thus, although the proposed measures of emotional intelligence should be evaluated within the context of the reliability and validity considerations described in this chapter, the other measures with which EI is "competing", both in terms of construct space and commercial application space, should be evaluated just as rigorously.

References

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, *40*, 37–47.

Anastasi, A. (1996). *Psychological testing* (7th ed.). New York: Macmillan.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 9–13). Hillsdale, NJ: Lawrence Elrbaum.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement, 62*, 254–263.

Block, J. (2000). Three tasks for personality psychology. In L. R. Bergman, R. B. Cairns, L. G. Nilsson, & L. Nystedt (Eds.). *Developmental science and the holistic approach*, (pp. 155–164). Mahwah, NJ: Lawrence Erlbaum Associates.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley-Interscience.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.
Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*, 505–514.
Brackett, M. A., & Mayer, J. D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin*, *29*, 1147–1158.
Brennan, R. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295–317.
Brody, N. (2004). Emotional intelligence: Science and myth (book review). *Personality and Individual Differences*, *32*, 109–111.
Brownell, W. A. (1933). On the accuracy with which reliability may be measured by correlating test halves. *Journal of Experimental Education*, *1*, 204–215.
Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.
Cattell, R. B., & Warburton, F. W. (1967). *Objective personality and motivation tests: A theoretical introduction and practical compendium*. Champaign, IL: University of Illinois Press.
Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavior Research, 18*, 115–126.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.
Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal modeling. *Applied Psychological Measurement*, *14*, 183–196.
Cortina, J. M. (1993). What is Coefficient Alpha? An examination of theory and applications. *Journal of Applied psychology, 78*, 98–104.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper & Brothers.
Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement*, *63*, 915–930.
Gignac, G. E. (2006). Testing jingle-jangle fallacies in a crowded market of over-expansive constructs: The case of emotional intelligence. In C. Stough, D. Saklofske, & K. Hansen (Eds.), *Research on emotional intelligence: International symposium 2005* (pp. 3–13). Melbourne: Tertiary Press.
Gignac, G. E. (2007). Working memory and fluid intelligence are both identical to g?! Reanalyses and critical evaluation. *Psychological Science, 49*, 187–207.
Gignac, G. E., Bates, T. C., & Lang, K. (2007). Implications relevant to CFA model misfit, reliability, and the five factor model as measured by the NEO-FFI. *Personality and Individual Differences*, *43*, 1051–1062.
Gignac, G. E., Palmer, B., & Stough, C. (2007). A confirmatory factor analytic investigation of the TAS-20: Corroboration of a five-factor model and suggestions for improvement. *Journal of Personality Assessment*, *89*, 247–257.
Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827–833.
Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*, 427–439.
Guin, R. M. (1977). Content validity – The source of my discontent. *Applied Psychological Measurement, 1*, 1–10.
Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sorebom (Eds.), *Structural equation*

modeling: Present and future – A festschrift in honor of Karl Jöreskog (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*, 78–80.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–455.

Lance, C. E., Butts, M. M., Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202–220.

Landy, F. J. (2005). Some historical and scientific issues related to research on emotional intelligence. *Journal of Organizational Behavior*, *26*, 411–424.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Marsh, H. W., & Byrne, B. W. (1993). Confirmatory factor analysis of multitrait-multimethod self-concept data: Between-group and within-group invariance constraints. *Multivariate Behavioral Research, 28*, 313–449.

Matarazzo, J. D., & Herman, D. O. (1984). Base rate data for the WAIS-R: Test–retest stability and VIQ–PIQ differences. *Journal of Clinical and Experimental Neuropsychology*, *6*, 351–366.

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

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

Mayer, J. D., Salovey, P., & Caruso, D. (2000). Models of emotional intelligence. In R. J. Sternberg (Ed.), *The handbook of intelligence* (pp. 396–420). New York: Cambridge University Press.

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

McCrae, R. R. (2000). Emotional intelligence from the perspective of the five-factor model of personality. In R. Bar-On & J. D. A. Parker (Eds.), *Handbook of emotional intelligence* (pp. 263–276). San Francisco: Jossey-Bass.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*, 175–215.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Statistical and Mathematical Psychology, 23*, 1–21.

McGrath, R. E. (2005). Conceptual complexity and construct validity. *Journal of Personality Assessment*, *85*, 112–124.

Muchinsky P. M. (1996). The correction for attenuation. *Educational & Psychological Measurement,* 56, 63–75.

Novick, M. R., & Lewis, C. L. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1–13.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

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

Parker, J. D. A., Bagby, R. M., Taylor, G. J., Endler, N. S., & Schmitz, P. (1993). Factorial validity of the twenty-item Toronto Alexithymia Scale. *European Journal of Personality*, *7*, 221–232.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research*. Orlando, FL: Harcourt Brace.

Peterson, R. A. (1994). A meta-analysis of Cronbach's alpha. *Journal of Consumer Research*, *21*, 381–391.
Raykov, T. (2001). Bias in coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*(1), 69–76.
Reuterberg, S. E., & Gustafsson, J.-E. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement, 52*, 795–811.
Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, *27*, 183–198.
Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922–932.
Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*(4), 299–312.
Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.
Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. *Journal of Organizational Behavior*, *15*, 385–392.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.
Thompson, B., & Vacha-Haase, T. (2000). Psychometrics and datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174–195.
Tomarken, A. J., & Waller, N. G. (2003). The problems with "well-fitting" models. *Journal of Abnormal Psychology*, *112*, 578–598.
Werts, C. E., & Watley, D. J. (1968). Analyzing school effects: How to use the same data to support different hypotheses. *American Educational Research Journal*, *5*, 585–598.
Zeidner, M., Matthews, G., & Roberts, R. D. (2001). Slow down, you move too fast: Emotional intelligence remains an "elusive" construct. *Emotion*, *1*, 265–275.