Learning Accelerator Research Paper

Development of a Forced-Choice Measure of Typical-Performance Emotional Intelligence

Cristina Anguiano-Carrasco

Carolyn MacCann

Mattis Geiger

Jacob M. Seybert

Richard D. Roberts

2015

Anguiano-Carrasco, C., MacCann, C., Geiger, M., Seybert, J.M. & Roberts, R.D. (2015). Development of a Forced-Choice Measure of Typical-Performance Emotional Intelligence. *Journal of Psychoeducational Assessment, 33*, 83-97 (DOI: 10.1177/0734282914550387)

LA015039

Development of a Forced-Choice Measure of Typical-Performance Emotional Intelligence

Cristina Anguiano-Carrasco, Carolyn MacCann, Mattis Geiger, Jacob M. Seybert, and Richard D. Roberts

Abstract

Self-report ratings of emotional intelligence (EI) can be faked in high-stakes situations. Although forced-choice administration can prevent response distortion, it produces ipsative scores when scored conventionally. This study (n = 486) develops an 18-item EI rating scale assessing emotion perception, understanding, and management. We compare validity evidence for: (a) a singlestimulus rating scale; and (b) a forced-choice assessment scored with conventional methods versus item response theory (IRT) methods. The single-stimulus items showed acceptable fit to a three-factor solution, and the forced-choice items showed acceptable fit to the IRT solution. Correlations with criterion variables (ability and self-reported EI, Big Five personality, loneliness, life satisfaction, and GPA) were obtained for 283 participants. Correlations were in the expected direction for the single-stimulus and the IRT-based forced-choice scores. In contrast, the conventionally scored forced-choice test showed the expected correlations for emotion management, but not for emotion perception nor understanding. Results suggest that IRT-based methods for scoring forced-choice assessments produce equivalent validity to singlestimulus rating scales. As such, IRT-based scores on forced-choice assessments may allow EI tests to be used for high-stakes applications, where faking is a concern.

Keywords

item response theory (IRT), emotional intelligence, forced-choice assessment, Thurstone's law of comparative judgment

There are two ways that emotional intelligence (EI) can be measured. First, *rating scales*, associated with typical performance, require participants to rate their agreement with items such as "I know why my emotions change." Second, *ability scales*, associated with maximum performance, require participants to process emotion-related stimuli and make a judgment (e.g., the extent of emotion expressed in a particular facial expression; Mayer, Roberts, & Barsade, 2008). One criticism of rating scales is that they are prone to response distortion. Test-takers are able to fake high scores if they choose to and may also unintentionally inflate their scores if they have an overly positive view of themselves as highly emotionally intelligent (Grubb & McDaniel, 2007). Moreover, such faking can reduce the correlation between EI scores and criteria (Choi, Kluemper, & Sauley, 2011). This is a clear problem for the use of EI rating scales for various applications (e.g., selection), as well as for research.

One solution to the issue of faking on rating scales is the use of forced-choice responses. In forced-choice items test-takers must choose between several positive statements representing different constructs and therefore cannot simply rate themselves highly on everything (e.g., "Which one is more like you: *I know why my emotions change* OR *I manage my emotions well*"). However, conventionally derived scores from forced-choice assessments are ipsative such that they cannot be compared across different people (Brown & Maydeu-Olivares, 2013). Recently, Brown and Maydeu-Olivares (2011, 2012) have proposed that item response theory (IRT) methods for scoring forced-choice assessments may solve the problem of ipsativity.

The goal of this study is to apply these IRT-based scoring methods to a forced-choice assessment measuring three key branches of EI (emotion perception, emotion understanding, and emotion management). We will compare three versions of this assessment: (a) a single-stimulus rating scale where participants rate each item independently, (b) a forced-choice version scored with these new IRT methods, and (c) a forced-choice version scored conventionally (such that it will yield ipsative scores).

Background to Assessment of EI

The dominant theoretical model of EI is a four-branch hierarchy where skills in the higher branches are dependent on skills in the lower branches (Mayer et al., 2008). The four branches are (a) *emotion perception* (the ability to identify and express emotions; e.g., recognizing a facial expression represents happiness), (b) *emotion facilitation of thought* (using emotions to aid in non-emotional task; e.g., deciding that a feeling of excitement would best be used to call up an old friend rather than assemble a complicated piece of machinery), (c) *emotion understanding* (understanding the way that emotions combine and progress over time; e.g., knowing that irritation might intensify into frustration if no solution is found), and (d) *emotion management* (reflexively managing the emotions of oneself and others; e.g., knowing the strategies to cheer up a friend who has had a stressful day at work; Mayer et al., 2008).

Typical Versus Maximum Performance Measures of EI

This four-branch hierarchical model defines EI as a set of abilities and therefore was initially operationalized by maximum performance scales (Mayer, Caruso, & Salovey, 1999; Roberts, Zeidner, & Matthews, 2001). However, several research teams have also used this four-branch hierarchical model as the theoretical basis for rating scales assessing EI (e.g., Brackett & Mayer, 2003; Schutte et al., 1998). Such ability model rating scales can be distinguished from the larger number of rating scales that are based on mixed models of EI, such as the Emotional Quotient Inventory (EQ-i; Bar-On, 2000) and the Trait Emotional Intelligence Questionnaire (TEIQue; Petrides, 2009). There are thus three ways EI is measured: (a) Maximum performance measures based on the ability model of EI, (b) rating scales based on the ability model of EI (what we will henceforth refer to as typical-performance EI), and (c) rating scales of alternative "mixed" models of EI (which tend to cover a broader content area than ability models, including aspects of motivation, character, and problem solving). These three types of measures show different relationships with personality, intelligence, and valued outcomes such as well-being and workplace performance (Davis & Humphrey, 2012; Joseph & Newman, 2010b).

The current study focuses on the second type of measure (typical-performance EI) based on the four-branch ability EI model. Scales measuring this construct are known to predict higher life satisfaction, lower loneliness, and greater academic achievement (Gardner & Qualter, 2010; MacCann & Burrows, 2013; Schutte et al., 1998). We will compare how strongly the three different types of EI scores (rating scales, conventionally scored forced-choice, and IRT-scored forced-choice) correlate with these outcomes. Furthermore, we will also compare how these three types of EI scores relate to the Big Five domains of personality. Meta-analytic estimates show that typical EI has positive correlations with Conscientiousness, Agreeableness, Extraversion, and Openness (ranging from .24 for Openness to .32 for Conscientiousness) and a negative correlation with Neuroticism (−.34; Joseph & Newman, 2010b).

Evidence for a Three-Branch EI Model

Recently, some researchers have suggested that the second branch of EI (using emotions to facilitate other tasks) is conceptually redundant with other branches and has measurement problems (Allen, MacCann, Matthews, & Roberts, 2014; Joseph & Newman, 2010b; MacCann, Joseph, Newman, & Roberts, 2014). Instead, a three-branch model of EI with emotion perception, understanding, and management may be a more appropriate way to conceptualize EI. We use this three-branch model as the basis for a new scale that we administer in both traditional format (single-stimulus rating scale items) and forced-choice format.

Introduction to Forced-Choice Assessment

The predominant approach to measuring typical-performance EI is through the use of singlestimulus items, where respondents are instructed to rate their level of agreement using a Likerttype rating scale. Because these items are often transparent in terms of their intended constructs, they are easily susceptible to response distortions such as faking and socially desirable responding. One alternative approach is the use of forced-choice item types, where statements are grouped in blocks, and the test-taker is instructed to make selections from among those within each block. These instructions can range from simply selecting the statement which is "most like" the respondent, to a partial ranking by selecting the "most like" and "least like," and the complete ranking of the full set. An example multidimensional forced-choice item triad is presented below, where
Statement B has been selected as "most like" and Statement C is selected as "least like."

(A) I can tell when my mood is low. [Emotion Perception]

Most (B) I know what can make me sad. [Emotion Understanding]

Least (C) I never let myself get too upset about trivial things. [Emotion Management]

Traditional approaches to scoring these measures involve assigning scale score points based on the relative position of the selected statements. In the example provided above, one scoring strategy would be to assign the dimension associated with the "most like" statement +1, the statement's dimension selected as "least like" −1, and the dimension not selected assigned a score of zero. Such an approach eliminates the ability of respondents to endorse every statement, reducing the ability of response distortions to uniformly influence scale scores.

Approaches to scoring forced-choice data using summative techniques result in *ipsative* or *partially ipsative* scale scores, where scores on attributes can only be interpreted relative to other scores within the individual (Baron, 1996). This ipsativity reduces the usefulness of scores when the purpose of testing is making comparisons across individuals. Maydeu-Olivares and Brown developed a method for collecting and scoring forced-choice responses in accordance with Thurstone's (1927) law of comparative judgment (Brown, 2010; Brown & Maydeu-Olivares, 2011, 2013; Maydeu-Olivares, 2001; Maydeu-Olivares & Brown, 2010). Thurstone proposed that when comparing stimuli (e.g., statements), an individual judges the *utility* of each stimulus, determining how closely it resembles his or her typical preference or behavior. The individual then selects the stimulus with the largest utility among the available options. Following this conceptual framework, Maydeu-Olivares and Brown developed a strategy to estimate normative trait scores across multiple dimensions. Specifically, Brown and Maydeu-Olivares (2011) proposed scoring binary responses derived from fully and partially ranked data using a multidimensional normal ogive model, with local dependencies due to statements appearing in the multiple pairs associated with each tetrad having constrained (equal) factor loadings. Brown and MaydeuOlivares (2013) provided Mplus (Muthén & Muthén, 1998-2010) syntax to compute item loadings, item thresholds, and factor scores, which are akin, respectively, to item discrimination, item extremity, and person parameters (trait scores) in traditional IRT terminology. For details on this confirmatory factor analysis (CFA) procedure, readers are encouraged to consult Brown and Maydeu-Olivares (2011, 2013).

Forced-Choice Assessment of EI

To our knowledge, there has only been one previous article reporting a forced-choice assessment of EI (Wong, Law, & Wong, 2004). Based on concerns about response distortion, Wong et al. created two forced-choice EI tests. Items in the first test contained two responses to a scenario (one representing high EI and one representing low EI). Items in the second test asked participants to choose between two statements (one representing high EI and one representing high cognitive ability). Both tests showed similar criterion correlations to a single-stimulus rating scale of EI (the Wong-Law EI Scale [WLEIS]). Wong et al. demonstrated that forced-choice EI assessments can show similar prediction to standard rating scale tests. However, it is not clear whether response distortion would in fact be lower for the first type of item (where the choice is between a high-EI and low-EI option). In addition, the confounding of low self-perceived cognitive ability with high self-perceived EI in the second type of item may make this problematic for applications having any meaningful stakes (e.g., both high cognitive ability and high EI are likely required for job selection). Given the transparency evident in the methodology adopted by Wong and colleagues, it is also likely that responses to them could be easily coached. Moreover, because scores are based on the conventional approach and thus ipsative, comparisons cannot be made across people.

In contrast to Wong et al.'s (2004) approach, the forced-choice format used in the current study consists of triads of statements (one from each of emotion perception, understanding, and management) where participants must endorse one statement as "most like me" and one statement as "least like me." This produces a rank-ordering of the three statements for each triad. Moreover, the current study considers three different sets of typical-performance EI scores assessing the three major components of EI: (a) a single-stimulus rating scale, (b) a conventionally scored forced-choice assessment, and (c) the same forced-choice assessment scored using Brown and Maydeu-Olivares's (2013) IRT methods. We will compare the correlation of these three sets of scores to known correlates of self-estimated ability EI. For all three sets of EI scores, we expect (a) high positive correlations with typical-performance EI and lower positive correlations with ability EI; (b) positive correlations with Conscientiousness, Agreeableness, Extraversion, and Openness, and a negative correlation with Neuroticism; and (c) positive correlations with life satisfaction and

GPA, and negative correlations with loneliness. We further expect that the IRT method of scoring will produce more valid results than the conventional scoring method (i.e., correlations are more likely to be in the expected direction), with validity evidence comparable to the single-stimulus rating scale.

Method

Participants and Procedure

Participants were recruited from the United States using Amazon's Mechanical Turk, and paid US$8 for their participation. Participants' responses were only used if they passed two data check items embedded within other rating scales: For example, "For data checking purposes, please select point 6 (Strongly Agree) for this item." After exclusion criteria were applied, 283 participants completed all tests in the battery outlined below (45.4% female, aged 18-74 with a mean age of 33.84 years, *SD* = 11.44; 78% White, 5.7% Hispanic, 7.1% Asian, 7.4% Black, 1.8% Other or unspecified). An additional 203 people completed only the forced-choice and rating scale test of EI (39.4% female, aged 19-64 with a mean age of 32.02 years, *SD* = 9.89; 77.8% White, 6.4% Hispanic, 3% Asian, 7.8% African American, 0.5% Other or unspecified). Participants also reported some basic demographic information as well as their high school GPA (not available for 3 participants). GPA was reported as the categories: 3.50 to 4.00; 3.00 to 3.49; 2.50 to 2.99; 2.00 to 2.49; 1.00 to 1.99; and less than 1.00, and re-coded as the mean value for each category (i.e., 3.5 to 4 = 3.75; 3 to 3.49 = 3.25, etc.). All tests and protocols were approved by the Educational Testing Service human ethics and fairness review committee.

Test Battery

Three-Branch Emotional Intelligence Rating Scale Assessment (TEIRA). Participants completed 24 items that assessed emotion perception (8 items), emotion understanding (8 items), and emotion management (8 items). Items were drawn from the Perception and Introspection of Emotions Schedule (PIES; Roberts, Schulze, & Sattler, 2005), with minor modifications. For example, we rephrased reverse-keyed items so that all items were positively keyed (e.g., "I have problems dealing with my anger" was changed to "I have no problems dealing with my anger"). All items were rated on the following 6-point scale: 1 = *strongly disagree*, 2 = *disagree*, 3 = *somewhat disagree*, 4 = *somewhat agree*, 5 = *agree*, and 6 = *strongly agree*. Example items are given in Table 1.

Table 1. Standardized Factor Loadings From the Confirmatory Factor Analysis of TEIRA Items.

No.	Item	F1	F2	F3
4	asked, I could gauge my level of sadness at any given time.	.57		
7	I can tell when my mood is low.	.64		
13	ly heart beats faster when I am excited about something.	.66		
16	I can tell when I am surprised by something.	.78		
19	asked, I could gauge my level of happiness at any given time.	.73		
22	I can tell when my mood is good.	.33		
5	When I am in a bad mood, I usually know why.		.70	
8	I know what can make me sad.		.67	
14	When I am surprised, I can usually tell people why.		.31	
17	When I am content I usually know why.		.77	
20	If I am in a good mood, I can identify what led to it.		.73	
23	I know what can put me in high spirits.		.80	
6	I can keep a cool head in distressing situations.			.55
9	I never let myself get too upset about trivial things.			.51
15	I know how to make a positive feeling last.			.71
18	I can easily keep myself in a good mood.			.73
21	I can control my laughter when I need to.			.76
24	I am good at cheering other people up.			.78

Note. Analysis undertaken in MPlus using WLSM estimator. F1 = Perception; F2 = Understanding; F3 = Management; TEIRA = Three-Branch Emotional Intelligence Rating Scale Assessment.

Three-Branch Emotional Intelligence Forced-Choice Assessment (TEIFA). This forced-choice assessment used the same 24 statements as the TEIRA, randomly grouped into 8 triads representing one statement of each subscale in each triad. In each of the 8 triads, test-takers had to choose which statement was *most like them* and *least like them*. For example, "A: When I get scared, I feel it physically; B: When I am in a bad mood, I can identify what led to it; C: It is easy for me to calm down after a heated argument."

Self-Report Emotional Intelligence Scale (SREIS). We used a subset of 12 items from the 19-item SREIS (Brackett, Rivers, Shiffman, Lerner, & Salovey, 2006), including only items assessing the following three subscales: Perception (4 items; e.g., "I am aware of the nonverbal messages other people send"), Understanding (4 items; e.g., "I have a rich vocabulary to describe my emotions"), and Managing Emotion (self; 4 items; e.g., "I know how to keep calm in difficult or stressful situations"). These items were intermixed with items from the TEIRA and were rated on the same 6-point rating scale.

Situational Test of Emotional Understanding–Brief (STEU-B). This 19-item multiple-choice assessment of emotion understanding (Allen, Weissman, Hellwig, MacCann, & Roberts, 2014) is a short form of the longer assessment described in MacCann and Roberts (2008). In each item, the test-taker is required to choose which of five emotions is most likely to result from an emotional situation. For example, "Xavier completes a difficult task on time and under budget. Xavier is most likely to feel? (a) Surprise; (b) Pride; (c) Relief; (d) Hope; (e) Joy" (right answer = b). Items are scored dichotomously according to a rubric derived from appraisal theory.

Situational Test of Emotional Management–Brief (STEM-B). This 18-item multiple-choice test (Allen, Rahman, Weissman, MacCann, & Roberts, 2014) is a short form of the longer test described in MacCann and Roberts (2008). In each item, the test-taker is required to select the most effective response to manage an emotional situation. For example, "Clayton has been overseas for a long time and returns to visit his family. So much has changed that Clayton feels left out. What action would be the most effective for Clayton? (a) Nothing—it will sort itself out soon enough; (b) Tell his family he feels left out; (c) Spend time listening and getting involved again; (d) Reflect that relationships can change with time." (best answer = c). Items are scored according to expert judgment, with partial scoring allowed.

Satisfaction With Life Scale (SWLS). This five-item rating scale (Diener, Emmons, Larsen, & Griffin, 1985) assesses satisfaction with life. All items are rated on the following 7-point scale: 1 = *strongly disagree*, 2 = *disagree*, 3 = *slightly disagree*, 4 = *neither agree nor disagree*, 5 = *slightly agree*, 6 = *agree*, 7 = *strongly agree*. For example, "I am satisfied with my life."

Loneliness. This eight-item rating scale (Hays & Dimatteo, 1987) assesses feelings of loneliness. Items are rated on the following 4-point scale: 1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = often. For example, "I feel left out."

Big Five Inventory (BFI-44). This 44-item rating scale (John & Srivastava, 1999) assesses the five broad personality traits of Neuroticism (8 items; e.g., "is depressed, blue"),

Extraversion (8 items; e.g., "is outgoing, sociable"), Openness to experience (10 items; e.g., "is inventive"), Agreeableness (9 items; e.g., "is generally trusting"), and Conscientiousness (9 items; e.g., "is a reliable worker"). There are 16 reverse-keyed items. Items are rated on the following 5-point rating scale: 1 = *disagree strongly*, 2 = *disagree a little*, 3 = *neither agree nor disagree*, 4 = *agree a little*, 5 = *agree strongly*.

Data Analysis

The three-factor structure of the TEIRA was tested to ensure that each of the three-component triads included in the TEIFA represented three separate factors. The SWLS, the Loneliness scale, the BFI-44, and the TEIRA were analyzed using the Graded Response Model and the estimation maximization algorithm (Samejima, 1969). For the BFI-44 and the TEIRA, a confirmatory multidimensional IRT approach was applied. All IRT modeling was done using the "MIRT" package (Chalmers, 2012) in R v3.0.2. Finally, TEIFA scores were computed using conventional scoring model (TEIFA) and Thurstonian IRT model (TEIFA-IRT) using the syntax creator provided by Brown and Maydeu-Olivares (2012).

Results

Structural Analysis of the TEIRA

A three-factor CFA was conducted on the 24 items of the TEIRA using MPlus with a WLSM (Weighted least squares, robust standard errors, & mean adjusted chi-square test statistic) estimator. Loadings for two of the emotion perception items were very low (<.10), and these were removed from further analyses. These two items contained content relating to perceiving the physiological manifestations of emotion. Two items were also removed from each of the understanding and management branches, as these items were used to form forced-choice triads with the two low-loading emotion perception items (i.e., we decided to remove all items that could not be used in the forced-choice version from the CFA of the rating scale items).

Fit indices for the 18-item three-factor model were as follows: χ^2 = 2420.753, *df* = 132, comparative fit index (CFI) = .913, Tucker–Lewis index (TLI) = .899, root mean square error approximation (RMSEA) = .189 (90% confidence interval [CI] = [.182, .196]), and WRMSR (Weighted root mean square residual) = 2.242. That is, the CFI and TLI indicated reasonable fit but the RMSEA and WRMSR did not. No cross-loadings or correlated error was modeled. Standardized factor loadings for the 18-item CFA are shown in Table 1. All loadings were salient, ranging from .31 to .80. Factor inter-correlations were high: .91 for Perception and Understanding, .94 for Perception and Management, and .98 for Understanding and Management.

Thurstonian IRT Analysis of the TEIFA

Using the Excel Macro provided by Brown and Maydeu-Olivares (2012), the MPlus syntax was written. After fixing the factor inter-correlations based on the CFA, the model converged showing an acceptable model fit: χ^2 = 240.004, *df* = 123; CFI = .933; RMSEA = .044 (90% CI = [.036, .053]).

Reliability and Descriptive Statistics for All Variables

Table 2 shows the reliability and descriptive statistics for each of the scales across the sample, and by each gender (along with effect size for the gender differences). Reliabilities of the TEIFA and TEIFA-IRT are not reported as reliability for forced-choice tests is artificially high (i.e., scale

Table 2. Reliability and Descriptive Statistics for All Study Measures.

			All		Males		Females		
	No. items	α	*M*	*SD*	*M*	*SD*	*M*	*SD*	Sex *d*
Emotional intelligence									
TEIFA: Perception[a]	6	—	0.67	2.56	0.50	2.57	0.90	2.55	−0.15
TEIFA: Understanding[a]	6	—	0.50	2.22	0.17	2.17	0.96	2.22	−0.35
TEIFA: Management[a]	6	—	−1.18	3.19	−0.67	3.15	−1.87	3.12	0.37
TEIFA-IRT: Perception[a]	6	—	0.01	0.71	0.06	0.68	0.06	0.68	0.00
TEIFA-IRT: Understanding[a]	6	—	−0.03	0.62	−0.05	0.53	−0.05	0.53	0.00
TEIFA-IRT: Management[a]	6	—	0.00	0.79	0.06	0.74	0.07	0.74	−0.01
TEIRA: Perception	6	.79	54.95	6.43	54.84	6.61	54.82	6.24	0.01
TEIRA: Understanding	6	.87	54.00	7.90	53.72	8.06	53.30	8.07	0.05
TEIRA: Management	6	.90	49.75	10.10	51.57	9.44	46.90	11.38	0.46
SREIS: Perception	4	.82	18.04	3.14	17.90	3.17	18.10	3.16	−0.06
SREIS: Understanding	4	.90	16.31	4.21	16.16	3.94	16.54	4.44	−0.09
SREIS: Management (Self)	4	.84	16.88	3.94	17.37	3.69	16.00	4.12	0.34
STEU-B (Understanding)	19	.54	0.63	0.14	0.61	0.15	0.65	0.15	−0.28
STEM-B (Management)	18	.71	0.69	0.14	0.56	0.14	0.62	0.13	−0.42
Personality									
Openness	10	.87	37.37	7.30	36.94	7.35	37.42	7.29	−0.06
Conscientiousness	9	.89	35.10	7.10	34.47	6.68	35.40	7.53	−0.13
Extraversion	8	.91	23.08	7.95	23.40	7.82	22.73	7.73	0.08
Agreeableness	9	.85	34.80	6.46	33.86	6.35	35.63	6.45	−0.27
Neuroticism	8	.91	21.02	7.81	19.74	7.33	22.82	7.83	−0.39
Criterion variables									
Life satisfaction	6	.89	27.45	8.22	26.25	8.33	28.84	7.56	−0.31
Loneliness	8	.63	18.83	8.22	18.82	3.92	18.91	4.24	−0.01
High school GPA	—	—	3.34	0.45	3.30	0.47	3.36	0.42	−0.13

Note. n = 486 for TEIFA and TEIRA; n = 283 for other measures. TEIFA = Three-Branch Emotional Intelligence Forced-Choice Assessment (forced-choice scored conventionally); TEIFA-IRT = forced-choice scored with Thurstonian IRT; TEIRA = Three-Branch Emotional Intelligence Rating Scale Assessment (self-report rating scales); SREIS = Self-Report Emotional Intelligence Scale; STEM-B = Situational Test of Emotional Management–Brief; STEU-B = Situational Test of Emotional Understanding–Brief. [a]Internal consistency-reliability of forced-choice items is not reported, in line with Tenopyr (1988).

reliabilities are highly inter-dependent meaning that one scale being reliable artificially increases all the other scales reliabilities; Tenopyr, 1988). All reliabilities are acceptable for research purposes, although the STEU-B and Loneliness reliability estimates were low (α = .54 and .63, respectively). The correlations obtained with these scales should be interpreted with caution.

Validity of the TEIRA, TEIFA, and TEIFA-IRT

Correlations among the TEIRA, TEIFA, and TEIFA-IRT. Correlations among the TEIRA, TEIFA, and TEIFA-IRT are shown in Table 3. Across all three EI branches, the TEIFA-IRT and TEIRA subscales showed significant positive correlations. However, the TEIFA Perception score showed significant (and very large) negative correlations with the Perception scores obtained from the TEIFA-IRT and TEIRA. In addition, the TEIFA Understanding score was unrelated to the Understanding scores

on the TEIFA-IRT and the TEIRA. Only the TEIFA Management scores were positively related to the Management scores in TEIFA-IRT and TEIRA. These results

Table 3. Correlations Among the Different Scoring Models of the Three-Branch EI Assessments.

	Perception			Understanding			Management		
	TEIFA	TEIFA-IRT	TEIRA	TEIFA	TEIFA-IRT	TEIRA	TEIFA	TEIFA-IRT	TEIRA
Perception: TEIFA		−.79*	−.45*	−.12**	−.47**	−.19**	−.72**	−.80**	−.36**
Perception: TEIFA-IRT			.42*	.16**	.62**	.20**	.53**	.57**	.35**
Perception: TEIRA				−.15**	.30**	.63**	.46**	.47**	.72**
Understanding: TEIFA					.51*	.05	.05	−.42**	−.60**
Understanding: TEIFA-IRT						.18**	.26**	.44**	.31**
Understanding: TEIRA							.12**	.16**	.79**
Management: TEIFA								.93**	.26**
Management: TEIFA-IRT									.31**

Note. EI = emotional intelligence; TEIFA = Three-Branch Emotional Intelligence Forced-Choice Assessment (forced-choice scored conventionally); TEIFA-IRT = forced-choice scored with Thurstonian IRT; TEIRA = Three-Branch Emotional Intelligence Rating Scale Assessment (self-report rating scales). *$p < .05$. **$p < .01$.

suggest that the conventionally scored forced-choice test may be a reasonable estimate of between-person differences in emotion management, but not emotion understanding nor emotion perception.

Correlations of the TEIFA, TEIFA-IRT, and TEIRA with other EI tests. Table 4 shows the correlations of the TEIRA, TEIFA, and the TEIFA-IRT scores with both typical-performance and ability-based EI.

Ability EI. None of the forced-choice scores (either conventionally or IRT-scored) were significantly related to either the STEU-B or the STEM-B. The TEIRA Understanding subscale was significantly related to both the STEU-B and STEM-B, with small effect size. TEIRA Perception and Management were not significantly related to ability-based EI.

Typical-performance EI. When the forced-choice assessment was conventionally scored, the Perception and Understanding subscales showed significant negative correlations with the SREIS in five of six cases, including some large effect sizes (e.g., SREIS Management and TEIFA Perception $r = -.55$). However, TEIFA Management showed a large positive and significant correlation with SREIS Management. In comparison, all correlations with the SREIS were positive for all subscales of the TEIFA-IRT and TEIRA, as would be expected theoretically given that these scales all assess typical-performance EI. All correlations with the SREIS subscales were significant for the TEIRA, and seven of the nine were significant for the TEIFA-IRT.

However, same-branch correlations were not always higher than different branch-correlations for either the TEIRA or TEIFA-IRT. For example, TEIRA Perception showed the highest correlation with SREIS Management (not SREIS Perception) and TEIRA Understanding was more strongly related to SREIS Management and Perception than SREIS Understanding. This pattern was also true for the TEIFA-IRT.

That is, the largest correlations of the new measures with the SREIS were for the SREIS Management subscale.

Correlations of the TEIFA, TEIFA-IRT, and TEIRA with criterion variables. Table 4 also shows the correlations of the three sets of scores on the new items with Big Five personality and the criterion variables. Partial correlations after controlling for Big Five personality are also given.

Table 4. Correlations of the Three-Branch EI Assessments With Other EI Assessments, Personality, and Criterion Variables.

	Perception			Understanding			Management		
	TEIFA	TEIFA-IRT	TEIRA	TEIFA	TEIFA-IRT	TEIRA	TEIFA	TEIFA-IRT	TEIRA
STEU-B	.05	−.03	−.01	.09	−.01	.21**	−.11	−.06	−.12
STEM-B	.01	−.04	.07	.06	.03	.21**	−.07	−.02	−.05
SREIS: Perception	−.14**	.12**	.41**	−.04	.09	.45**	.02	.14**	.14**
SREIS: Understanding	−.14**	.16**	.42**	−.09**	.05	.37**	.07	.17**	.18***
SREIS: Management	−.55**	.43**	.67**	−.25**	.32**	.41**	.68*	.67**	.62**
Openness	.02	−.01	.29**	.01	−.01	.35**	.01	.02	−.02
Conscientiousness	−.20**	.21**	.37**	.09	.23**	.33**	.21*	.17**	.11
Extraversion	−.15*	.15*	.44**	−.08	.12*	.19**	.28*	.17**	.18**
Agreeableness	−.10	.09	.35**	−.09	.08	.35**	.16*	.11	.16**
Neuroticism	.46**	−.39**	−.59**	.18**	−.31**	−.27**	−.62*	−.59**	−.53**
SWLS	−.20**	.22**	.47**	−.08*	.20**	.24**	.26*	.25**	.23**
(SWLS partial)	(−.03)	(.07)	(.27)	(−.04)	(.07)	(.09)	(.06)	(.06)	(.10)
Loneliness	.31**	−.34**	−.54**	.08	−.22**	−.28**	−.36*	−.33**	−.32**
(loneliness partial)	(.12)	(−.20)	(−.29)	(.01)	(−.08)	(−.10)	(−.12)	(−.12)	(−.13)
High school GPA	.07	.03	.09	−.04	.01	−.03	.08	.10*	−.01
(GPA partial)	(.06)	(−.08)	(−.01)	(.06)	(−.07)	(−.05)	−.10	(−.06)	(−.08)

Note. Partial correlations controlling for Big Five in parentheses. EI = emotional intelligence; TEIFA = Three-Branch Emotional Intelligence Forced-Choice Assessment (forced-choice scored conventionally); TEIFA-IRT = forced-choice scored with Thurstonian IRT; TEIRA = Three-Branch Emotional Intelligence Rating Scale Assessment (self-report rating scales). STEU-B = Situational Test of Emotional Understanding–Brief; STEM-B = Situational Test of Emotional Management–Brief; SREIS = Self-Report Emotional Intelligence Scale; SWLS = Satisfaction With Life Scale. *$p < .05$. **$p < .01$.

Personality and well-being outcomes. Again, results for the TEIFA follow the expected pattern of correlations for Management, but not for Understanding (e.g., the correlation with Neuroticism is significantly positive, whereas the correlation with life satisfaction is significantly negative) nor for Perception (e.g., the correlations with Neuroticism and loneliness are significantly positive, and the correlations with life satisfaction, Extraversion, and Agreeableness are significantly negative). In contrast, correlations of the TEIFA-IRT and TEIRA are in the expected direction for all three branches (Perception, Understanding, and Management), and all subscales significantly predict both loneliness and life satisfaction. Correlations are higher for the TEIRA than the TEIFA-IRT. However, these various scales are single-stimulus rating scales and thus share a method effect with the TEIRA.

Partial correlations controlling for personality are given in parentheses. None of the partial correlations are significant for any set of scores (i.e., there is no evidence of incremental validity for these 18 items, irrespective of how they were administered or scored).

GPA. Only the TEIFA-IRT Management subscale significantly predicted GPA, and this was of very small effect size and not significant after controlling for personality. The single-stimulus rating scale showed no significant correlations with GPA.

Correlations of other EI variables with criterion variables. For comparison purposes, Table 5 provides a breakdown of individual correlations of SREIS, STEM-B, and STEU-B with personality and outcome (e.g., GPA) variables. Correlations with personality and criterion data are of similar magnitude for the TEIRA as the SREIS, indicating similar levels of validity. Correlations for the TEIFA-IRT are slightly lower, as might be expected given that outcome variables are singlestimulus rating scales. The ability-based assessments of EI (the STEM-B and STEU-B) show very small relationships with personality and do not significantly predict any of the criteria in this study.

Table 5. Correlations Among SREIS, STEU-B, STEM-B, and Personality and Outcome Measures.

	SREIS				
	Perception	Understanding	Management	STEU-B	STEM-B
Openness	.23**	.36**	.18**	.16**	.19**
Conscientiousness	.26**	.24**	.33**	.14*	.18**
Extraversion	.14*	.28**	.32**	−.17**	−.02
Agreeableness	.29**	.17**	.32**	.09	.17**
Neuroticism	−.17**	−.22**	−.71**	.01	−.03
SWLS (SWLS partial)	.19** (.06)	.25** (.11)	.37** (.10)	.06 (.07)	.05 (−.01)
Loneliness (Loneliness partial)	−.23** (−.09)	−.24** (−.09)	−.44** (−.11)	.02 (.02)	−.01 (.06)
High school GPA (GPA partial)	.02 (.01)	.07 (−.02)	.11* (−.06)	.03 (.03)	.03 (.04)

Note. Partial correlations controlling for Big Five in parentheses. SREIS = Self-Report Emotional Intelligence Scale; STEU-B = Situational Test of Emotional Understanding–Brief; STEM-B = Situational Test of Emotional Management– Brief; SWLS = Satisfaction With Life Scale.
*$p < .05$. **$p < .01$.

Discussion

Results demonstrated that the use of IRT methods to score a forced-choice EI assessment provides a sounder basis for comparing EI differences across people than do conventional forcedchoice scoring methods. Correlations with personality, EI, and life outcome criteria were similar for the single-stimulus rating scale and the IRT-scored forced-choice tests but were clearly not in the expected direction when the forced-choice assessment was scored in the conventional fashion.

Validity of the TEIRA

The CFA of the TEIRA items provided mixed evidence for its structural validity. Although a three-factor structure is reasonable, and the three subscales produced reliable scores, there is clearly misfit in the data, indicating that the factor structure

should be confirmed with new samples. Alternatively, further test development involving additional items may be required to produce a reasonable factor structure, given that only a subset of PIES items were used in the current study.

TEIRA scores showed strong relationships with another rating scale EI measure (the SREIS) and also significantly predicted life satisfaction and loneliness, with a moderate to strong effect (most strongly for the Perception subscale). However, there was also a moderate to strong association with personality, particularly with low levels of Neuroticism. Neuroticism correlations were so high that they could be interpreted as problematic from the perspective of discriminant validity evidence. The low discriminant validity of rating scale measures of EI with respect to personality (particularly Neuroticism) is well documented across multiple different tests (including the SREIS in this study), and so the TEIRA is representative of a rating scale measure of EI in this respect (Joseph & Newman, 2010a; Saklofske, Austin, & Minski, 2003). The forcedchoice scores (both conventionally scored and IRT-scored) also showed strong relationships with Neuroticism, suggesting that this issue with discriminant validity cannot be overcome by using forced-choice measurement techniques.

Perhaps as a result of the moderate correlations with personality, scores on the new assessment showed no evidence of incremental prediction of outcomes. This is also reasonably consistent with this study's findings for the SREIS and with other research of rating scales assessing EI (e.g., Amelang & Steinmayr, 2006; MacCann & Burrows, 2013). Similarly, the small positive associations with ability-based EI are consistent with prior research on the relationship between rating scales and ability scales based on the same theoretical model (e.g., Brackett & Mayer, 2003; Brackett et al., 2006). The TEIRA showed no relationship with GPA in this study. However, GPA was retrospectively reported from a sample of varying age, such that grade inflation over the last several decades may have made this an unreliable index of academic achievement (Woodruff & Ziomek, 2004). Indeed, no other EI measure included in this study was significantly associated with GPA.

Comparison of Forced-Choice EI Tests Under IRT-Based and Conventional Scoring

Validity evidence for the forced-choice assessment of EI is much stronger for scores derived from IRT-based scoring than scores derived from conventional methods. In particular, conventional scoring resulted in Perception subscales that showed many significant correlations in the *opposite* direction to those predicted by theory. In the current study, the ipsativity of the scores seems to have primarily damaged the Perception and Understanding subscales, whereas the Management subscale was actually fairly similar to the single-stimulus rating scale in terms of its correlates. This might be expected from ipsative scores, as the correlations between subscales and external criteria must sum to zero across all the subscales (Brown & Maydeu-Olivares, 2012). In contrast, the IRT-based scores on the forced-choice EI test were similar to the scores obtained from single-stimulus administration in terms of correlations to external criteria.

One of the advantages of the IRT-based approach is that the EI assessment can be given as a short stand-alone assessment. Conventionally scored forced-choice assessments often require a very large number of constructs to be included so that the negative effects of ipsative scores can be diluted across multiple constructs (e.g.,

Bartram, 1996). The ability to use short stand-alone forced-choice assessments in high-stakes situations has clear practical utility. Assessments can be administered quickly and can be combined with other tools as needed in different testing situations, providing greater flexibility.

Limitations and Future Directions

The primary purpose of developing a forced-choice counterpart to the standard rating scale administration of EI is to provide an alternative method of assessment that is less susceptible to faking. Thus, one obvious direction for future research is to test whether criterion correlations of IRT-based forced-choice scores hold under testing conditions that motivate faking. Either an instructed faking paradigm or incentivizing high scores to simulate a high-stakes environment could be compared with scores obtained under standard instructions in a low-stakes environment (Mueller-Hanson, Heggestad, & Thornton, 2006). This study demonstrated that forced-choice administration of EI tests can show comparable validity (when scored using IRT methods) to conventional single-stimulus rating scale administration. What is yet to be demonstrated is whether forced-choice assessments are in fact superior under high-stakes conditions (i.e., whether forced-choice administration formats retain their validity evidence under fake-high conditions).

In the current study, we developed a new set of items with unknown psychometric properties to compare forced-choice assessment with single-stimulus rating scale assessment administration procedures. This had the disadvantage that we did not know, a priori, whether the newly developed items would clearly fit the intended theoretical structure (and in fact, results are mixed as to whether they do fit such a structure). Future research could fruitfully apply this same design to existing research instruments where the structure is well known, and items are a priori known to load on a single construct only.

Conclusion

Brown and Maydeu-Olivares's (2011, 2012, 2013) IRT methods of scoring forced-choice assessments have resulted in superior evidence of validity compared with conventional scoring mechanisms such as subtracting a point for "least like me" and adding a point for "most like me." Although follow-up studies examining this effect in high-stakes environments or under-instructed faking are needed, this method looks promising as an efficient way to administer rating scale items in a forced-choice format that may be less susceptible to faking but also immune from the validity issues that ipsativity may produce.

Acknowledgment

We thank Patrick Barnwell, Meghan Brenneman, Jeremy Burrus, Mary Lucas, and Heather Walters for supporting the preparation of this article.

Authors' Note

All statements expressed in this article are the authors' and do not reflect the official opinions or policies of any of the authors' host affiliations or the U.S. Army Research Institute.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was written with the support of an U.S. Army Research Institute Contract to the Educational Testing Service (ETS).

References

Allen, V. D., MacCann, C., Matthews, G., & Roberts, R. D. (2014). Emotional intelligence in education: From pop to emerging science. In R. Pekrun & L. L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 162-182). New York, NY: Taylor & Francis.

Allen, V. D., Rahman, N., Weissman, A., MacCann, C., & Roberts, R. D. (2014). Development and validation of the Situational Test of Emotional Management–Brief (STEM-B) using item response theory and latent class analysis. *Personality and Individual Differences*. Manuscript submitted for publication.

Allen, V. D., Weissman, A., Hellwig, S., MacCann, C., & Roberts, R. D. (2014). Development of the Situational Test of Emotional Understanding–Brief (STEU-B) using item response theory. *Personality and Individual Differences*, *65*, 3-7.

Amelang, M., & Steinmayr, R. (2006). Is there a validity increment for tests of emotional intelligence in explaining the variance of performance criteria? *Intelligence*, *34*, 459-468.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.

Bar-On, R. (2000). Emotional and social intelligence: Insights from the Emotional Quotient Inventory. In R. Bar-On & J. D. A. Parker (Eds.), *Handbook of emotional intelligence* (pp. 363-388). San Francisco, CA: Jossey-Bass.

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology*, *69*, 25-39.

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.

Brackett, M. A., Rivers, S. E., Shiffman, S., Lerner, N., & Salovey, P. (2006). Relating emotional abilities to social functioning: A comparison of self-report and performance measures of emotional intelligence.
Journal of Personality and Social Psychology, *91*, 780-795.

Brown, A. (2010). *How IRT can solve problems of ipsative data* (Doctoral dissertation). University of Barcelona, Spain.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*, 460-502.

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*, 1135-1147.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*, 36-52.

Chalmers, P. R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29.

Choi, S., Kluemper, D. H., & Sauley, K. S. (2011). What if we fake emotional intelligence? A test of criterion validity attenuation. *Journal of Personality Assessment*, *93*, 270-277.

Davis, S. K., & Humphrey, N. (2012). The influence of emotional intelligence (EI) on coping and mental health in adolescence: Divergent roles for trait and ability EI. *Journal of Adolescence*, *35*, 1369-1379.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*, 71-75.

Gardner, K. J., & Qualter, P. (2010). Concurrent and incremental validity of three trait emotional intelligence measures. *Australian Journal of Psychology*, *62*, 5-13.

Grubb, W. L., & McDaniel, M. A. (2007). The fakability of Bar-On's Emotional Quotient Inventory Short Form: Catch me if you can. *Human Performance*, *20*, 43-59.

Hays, R. D., & Dimatteo, M. R. (1987). A short-form measure of loneliness. *Journal of Personality Assessment*, *51*, 69-81.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102-138). New York, NY: Guilford Press.

Joseph, D. L., & Newman, D. A. (2010a). Discriminant validity of self-reported emotional intelligence: A multitrait-multisource study. *Educational and Psychological Measurement*, *70*, 672-694.

Joseph, D. L., & Newman, D. A. (2010b). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology*, *95*, 54-78.

MacCann, C., & Burrows, C. K. (2013). Does self-report emotional intelligence incrementally predict student affective outcomes and GPA beyond Five-Factor personality? *The Psychology of Education Review*, *37*, 33-39.

MacCann, C., Joseph, D. L., Newman, D. A., & Roberts, R. D. (2014). Emotional intelligence is a secondstratum factor of intelligence: Evidence from hierarchical and bifactor models. *Emotion*, *14*, 358-374.

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, *8*, 540-551.

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66*, 209-228.

Maydeu-Olivares, A., & Brown, A. (2010).Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *45*, 935-974.

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

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

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. *Psychology Science*, *48*, 288-312.

Muthén, L. K., & Muthén, B. (1998-2010). *Mplus 6*. Los Angeles, CA: Author.

Petrides, K. V. (2009). Psychometric properties of the trait emotional intelligence questionnaire (TEIQue). In C. Stough, D. H. Saklofske, & J. D. A. Parker (Eds.), *Assessing emotional intelligence: Theory, research, and applications* (pp. 85-101). New York, NY: Springer.

Roberts, R. D., Schulze, R., & Sattler, J. (2005). *Research report on the Perception and Introspection of Emotions Schedule (PIES)*. New York, NY: Educational Testing Service.

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

Saklofske, D. H., Austin, E. J., & Minski, P. S. (2003). Factor structure and validity of a trait emotional intelligence measure. *Personality and Individual Differences*, *34*, 707-721.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*, 5-17.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., & Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and Individual Differences*, *25*, 167-177.

Tenopyr, M. L. (1988). Artifactual reliability of forced-choice scales. *Journal of Applied Psychology*, *4*, 749-751.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286.

Wong, C., Law, C. S., & Wong, P. (2004). Development and validation of a forced choice emotional intelligence measure for Chinese respondents in Hong Kong. *Asia Pacific Journal of Management*, *21*, 535-559.

Woodruff, D. J., & Ziomek, R. L. (2004). *High school grade inflation from 1991 to 2003 ACT Research Report Series* (Vol. 4). Iowa City, IA: ACT.