The typical-performance emotional management test:

Test-retest reliability and further validity evidence

Aljoscha C. Neubauer & Gabriela Hofer

University of Graz, Institute of Psychology, Graz, Austria

[Word count: 2,497]

Authors' Note

Aljoscha C. Neubauer <https://orcid.org/0000-0002-1326-3104>

Gabriela Hofer <https://orcid.org/0000-0003-4407-1487>

We have no conflicts of interest to disclose. Open data for this manuscript are available on the Open Science Framework ([Authors, 2021](#)). We want to thank our former master student Carina Limberger for her help in data collection.

Correspondence concerning this article should be addressed to Aljoscha C. Neubauer, Institute of Psychology, Section for Differential Psychology, University of Graz, Universitätsplatz 2, 8010 Graz, Austria. Email: aljoscha.neubauer@uni-graz.at

This manuscript is an earlier version of one that is now published in *Personality and Individual Differences*. Please refer to:

Neubauer, A. C., & Hofer, G. (2022). (Retest-)Reliable and valid despite low alphas? An example from a typical performance situational judgment test of emotional management. *Personality and Individual Differences*, 189, 111511. <https://doi.org/10.1016/j.paid.2022.111511>

Abstract

Situational judgments tests (SJTs) offer many advantages over traditional trait questionnaires but often show low internal consistency, presumably due to heterogeneity of the assessed constructs (Lievens et al. 2008; Whetzel & McDaniel, 2009). Authors have suggested that test-retest-reliability (TRT) might be a better-suited measure for SJTs. Here, we analyzed TRT of the typical-performance emotional management test (TEMT, Freudenthaler & Neubauer, 2005) – and SJT for emotional management. We additionally report further and hitherto not analyzed validity evidence with respect to a maximum-performance emotional management SJT (STEM, MacCann & Roberts, 2008) and the widely used Trait Emotional Intelligence Questionnaire (TEIQue, Petrides & Furnham, 2003). For a 2 week-interval, we found satisfying test-retest reliabilities for both scales of the TEMT (inter- and intrapersonal emotional management; .76 and .83, respectively). Intrapersonal ability correlated moderately to highly with all four TEIQue main factors and the total score, whereas interpersonal ability showed low to medium relations. Relations with the STEM were smaller. Our results support that SJTs can show good reliabilities – when assessed via TRT – and emphasize the importance to distinguish between typical- and maximum-performance tests of emotional abilities.

Keywords: Emotional Intelligence, Emotional Management, Typical Performance, Situational Judgment Test, Test-Retest-Reliability, Validity

The typical-performance emotional management test:

Test-retest reliability and further validity evidence

1. Introduction

In the field of emotional intelligence (EI), two main kinds of measurement approaches can be distinguished: While the ability approach focuses on the maximal performance of individuals and employs performance tests like the MSCEIT (Mayer et al., 2003), the trait approach focuses on people's typical behavior and uses self-report questionnaires like the TEIQue (Petrides and Furnham, 2003). A possible third way that has been suggested is a mixed approach, that means, employing a performance-like test but not asking for maximum performance but for typical behavior. An example of the mixed approach is the Typical Emotional Management Test (TEMT; Freudenthaler & Neubauer, 2005). In this situational judgment test (SJT), participants read performance items (emotional scenarios) and are asked to respond in accordance to how they would typically react. Past research showed several advantages over a maximum performance approach. Examples include higher reliabilities and validities, but also different correlational patterns with intelligence vs. big five traits (see Freudenthaler & Neubauer, 2005, 2007).

Despite displaying satisfying validities, SJTs are often plagued by low internal consistency (coefficient alpha; see Lievens et al., 2008). This has been the case for both the adult and the adolescent forms of the TEMT (Neubauer et al., 2018; Neubauer & Hofer, 2021) as well as for the Situational Test of Emotion Management (STEM, MacCann & Roberts, 2008) – a similar test of emotional management. According to Whetzel and McDaniel (2009), this is not surprising: SJTs are inherently heterogeneous and this makes alpha an unsuitable measure of their reliability. These authors have recommended parallel form or test-retest reliability as more adequate reliability coefficients for SJTs. In line with this, a two-week test-retest reliability of over .8 was reported for the STEM, despite its low internal consistency

(Libbrecht & Lievens, 2012). We know of no study that investigated the test-retest reliability of the TEMT – a gap in the literature that we aim to close.

Apart from assessing the TEMT's test-retest reliability, the present investigation had a second objective: To investigate the TEMT's relations to two other measures of emotional competence: The first one – the Trait Emotional Intelligence Questionnaire (TEIQue; Petrides and Furnham, 2003) – is an extensive self-report measure of emotional intelligence encompassing four factors – well-being, self-control, emotionality, and sociability – and 15 facets. In their German translation of the TEIQue, Freudenthaler and colleagues (2008) have found medium to high associations between the four TEIQue main factors and TEMT intrapersonal management (r between .36 and .55). TEMT interpersonal management showed small ($r = .2$ each for well-being and self-control) to medium associations ($r = .35$ and $r = .4$ for sociability and emotionality, respectively). We aimed to replicate these results but also to investigate the TEMT's associations with the 15 TEIQue facets. Our second validation criterion was the Situational Test of Emotion Management (STEM; MacCann & Roberts, 2008) – a maximum-performance emotional management test. The STEM is an SJT, in which participants rate the effectiveness of a person's reaction to an emotional situation. It shares some characteristics with the TEMT intrapersonal scale – both relate to a person's management of their own emotions – but differs from the TEMT in two important aspects: In the STEM, the subject in the scenarios is not the participant but a third person. Moreover, the instruction asks for the effectiveness of reactions (maximum performance) and not – as in the TEMT – how typical they are of the participant.

2. Methods

2.1. Participants

Participants were recruited via different online platforms. 169 participants between 18 and 60 years ($M = 29.58$; $SD = 12.03$) completed the first part of the study. 135 were female and 2 reported their gender to be diverse. 110 were students, 50 were employed, 2 were

school pupils, and one was unemployed. 89 had a high school degree, 64 a university or college diploma, and the remaining had completed compulsory schooling or an apprenticeship. 136 completed the retest (age: $M = 29.46$; $SD = 12.06$; 106 female, 2 diverse). Psychology students received course credits and all participants were offered feedback regarding their emotional competence. All participants provided their informed consent and the study protocol was approved by the local ethics committee.

2.2. Measures

The TEMT (Freudenthaler & Neubauer, 2005) is an SJT measuring emotional management abilities (EMA) with two subscales: intrapersonal EMA (the ability to manage one's own emotions; 18 items) and interpersonal EMA (the ability to manage the emotions of others; 24 items). Each item consists of a short scenario and four possible responses (for a sample item see OSF;

https://osf.io/pw3sr/?view_only=fb6b65fe24b646659dae4559c80f393d). Participants are explicitly instructed to choose the alternative that would best describe their actual behavior in the given situation. The adequacy/effectiveness of each response had been rated by a panel of 10 experts in the field of emotions (psychologists, psychotherapists) at scale construction. Based on these ratings, participants' responses were scored from 1 (least adequate reaction) to 4 (most adequate reaction). Like expected, internal consistencies were rather low (α between .58 and .68)¹. In Freudenthaler and Neubauer (2005), the test showed satisfying convergent and discriminant validities.

The TEIQue was used in the German translation by Freudenthaler et al.(2008) and consists of 153 7-point Likert scale items measuring 13 primary facets that are then further grouped into the 4 second-order factors of emotionality, sociability, self-control, and well-being (α between .74 and .88). Two additional primary scales are used in conjunction with the

¹ Internal consistencies, descriptive statistics, and intercorrelations of all study variables can be found in Tables 1 and 2 in the Appendix on the OSF.

13 other facets to form a general factor of Emotional Intelligence ($\alpha = .90$). Most primary scales displayed satisfying reliabilities ($\alpha > .70$), the only exception being “relationships” ($\alpha = .65$).

The third test was the STEM (German version by Hilger et al.; 2012). In this SJT, participants judge the effectiveness of four different reactions to each of 44 scenarios on a Likert scale from 1 (not effective at all) to 6 (very effective). Each individual response is then compared to the average rating of the sample (i.e., the score of a person reflects how much they deviate from the average response). Low scores indicate a small deviation from the consensus and, therefore, high performance. In our sample, the test showed good internal consistency.

2.3. Procedure

All tests were administered online. In the first session, participants responded to some demographic questions and then completed the TEMT, the TEIQue, and STEM (total duration 45-60 min). Two weeks later, participants were contacted via email and asked to complete a second survey within the following days. The actual test-retest interval varied between 13 days and 3 weeks ($Md = 14$ days). The second test session only contained the TEMT and took 15-20 minutes to complete.

3. Results and Discussion

We provide data and code on the OSF. The majority of results is based on $N = 169$; only the test-retest reliability is based on $N = 136$. Due to non-normal distribution of some variables and some violated multiple regression assumptions (see OSF), we based our interpretations on 95% percentile confidence intervals based on 2,000 bootstrap samples instead of conventional significant thresholds.

The correlations between the first and second administration of the TEMT were .83 (95% CI [.77, .88]) and .76 (95% CI [.68, .93]), for intra- and interpersonal EMA, respectively. In comparison with the rather low alphas, this confirms the assumption (Whetzel

& McDaniel, 2009; Libbrecht & Lievens, 2012) that the low alphas found for SJTs do not necessarily mean that these tests are unreliable. Based on these test-retest correlations, the TEMT can be considered a reliable measurement instrument for (intra- and interpersonal) emotional management.

Table 1 shows the correlations of the TEMT intra- and interpersonal subscales with the TEIQue and the STEM. For the TEIQue factors, correlations are comparable to those reported by Freudenthaler et al. (2008): The intrapersonal TEMT score showed medium to high correlations with all TEIQue second-order factors and a high correlation with the TEIQue general factor. The interpersonal TEMT score, on the other hand, showed low to medium correlations with the TEIQue second-order factors and the general TEIQue-EI score. Results for the 15 TEIQue facets corresponded to this pattern (see Table 2 in the appendix):

Correlations with the TEMT intrapersonal scale were mostly medium to high (only 3 out of 15 $r < .3$), whereas most correlations with the TEMT interpersonal scale were low (only 5 out of 15 $r \geq .3$). Taken together, these results lead us to conclude that the TEIQue is likely rather a measure of intrapersonal than of interpersonal management tendencies.

We had also expected the STEM to be related to the TEMT, particularly to intrapersonal EMA. Contrary to this, the STEM only correlated significantly with the TEMT interpersonal – and not the intrapersonal – subscale (see Table 1). One potential reason for this could be that the STEM-items are phrased in the third person and, thus, require test-takers to put themselves into someone else's shoes. In line with this, the STEM showed some of its highest – albeit still low – correlations with the TEIQue facets empathy and relationships (see Table 2 in the Appendix). Overall, the low correlations between both TEMT subscales and the STEM suggest that maximum vs. typical performance tests are two very different things and show – if any – only weak relations. This is also corroborated by the low correlations the STEM showed with the TEIQue.

Table 1*Correlations between the TEMT, the STEM, and the TEIQue*

Variable	TEMT intrapersonal EMA			TEMT interpersonal EMA		
	<i>r</i>	95% CI	<i>p</i>	<i>r</i>	95% CI	<i>p</i>
STEM EMA (low)	-.143	[-.290, .008]	.064	-.265	[-.418, -.105]	.002
TEIQue global trait EI	.583	[.474, .673]	< .001	.371	[.220, .506]	< .001
TEIQue factor well-being	.472	[.341, .585]	< .001	.180	[.030, .337]	.018
TEIQue factor self-control	.446	[.310, .564]	< .001	.283	[.126, .432]	< .001
TEIQue factor emotionality	.409	[.281, .530]	< .001	.418	[.264, .552]	< .001
TEIQue factor sociability	.506	[.365, .630]	< .001	.255	[.105, .398]	.002

Note. EMA = Emotion Management Abilities. EI = Emotional Intelligence. TEMT = Typical Performance Emotional Management Test. STEM = Situational Test of Emotion Management. TEIQue = Trait Emotional Intelligence Questionnaire. Low values in the STEM indicate high EMA. Results are for the whole sample ($n = 169$). Values in brackets represent 95% percentile bootstrap confidence intervals based on 2000 samples.

As a final test of the TEMT's validity, we conducted two hierarchical regressions – one per TEMT subscale – entering the STEM in the first step and the four TEIQue factors in the second step (for detailed results see Table 3 in the Appendix). For the intrapersonal subscale, only the second step (36.4%; TEIQue) contributed significantly to the model (1st step - STEM: 2%). Scores on the interpersonal subscale were significantly explained by both the first (7%; STEM) and second (17.5%; TEIQue) step. These findings again underline the TEMTs stronger relations to typical emotion-related behavior than to maximum performance.

While our study comes with some limitations (e.g., the predominantly female sample), it added to the literature demonstrating that SJTs can prove reliable if tested through repeated measurement and not internal consistency. Future work constructing or applying SJTs would,

therefore, benefit from reporting test-retest correlations as additional – and likely more suitable – reliability evidence. We also provided new validity evidence for the TEMT and further support that typical- and maximum-performance tests of EI are distinct measures.

What remains unanswered at present, is the question what test format – typical- or maximum-performance – allows for better predictions of external criteria like professional or personal success.

References

- Authors (2021). *The typical-performance emotional management test: Test-retest reliability and further validity evidence [Data set]*.
https://osf.io/pw3sr/?view_only=fb6b65fe24b646659dae4559c80f393d
- Freudenthaler, H. H. & Neubauer, A. C. (2007). Measuring emotional management abilities: Further evidence of the importance to distinguish between typical and maximum performance. *Personality and Individual Differences*, 42, 1561-1572.
- Freudenthaler, H. H., Neubauer, A. C., Gabler, P., Scherl, W. G., & Rindermann, H. (2008). Testing and validating the trait emotional intelligence questionnaire (TEIQue) in a German-speaking sample. *Personality and Individual Differences*, 45(7), 673–678.
<https://doi.org/10.1016/j.paid.2008.07.014>
- Freudenthaler, H.H. & Neubauer, A.C. (2005). The convergent and discriminant validities of intra- and interpersonal emotional abilities. *Personality and Individual Differences*, 39, 569-579.
- Hilger, L., Hellwig, S., & Schulze, R. (2012). *Deutschsprachige Adaptation des STEU sowie des STEM und erste Validitätsevidenz [German adaptation of the STEU and the STEM and first validity evidence]*. 48th Congress of the German Psychological Society, Bielefeld.
- Libbrecht, N., & Lievens, F. (2012). Validity evidence for the situational judgment test paradigm in emotional intelligence measurement. *International Journal of Psychology*, 47(6), 438–447. <https://doi.org/10.1080/00207594.2012.682063>
- Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, 37(4), 426–441.
<https://doi.org/10.1108/00483480810877598>
- MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. *Emotion*, 8(4), 540. <https://doi.org/10.1037/a0012746>.

- Mayer, J., Salovey, P., Caruso, D., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT v2.0. *Emotion*, 3(1), 97–105. <https://doi.org/10.1037/1528-3542.3.1.97>
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60(1), 63–91. <https://doi.org/10.1111/j.1744-6570.2007.00065.x>
- Neubauer, A. C., & Hofer, G. (2020). Self-estimates of abilities are a better reflection of individuals' personality traits than of their abilities and are also strong predictors of professional interests. *Personality and Individual Differences*, 109850. <https://doi.org/10.1016/j.paid.2020.109850>
- Neubauer, A. C., Pribil, A., Wallner, A., & Hofer, G. (2018). The self–other knowledge asymmetry in cognitive intelligence, emotional intelligence, and creativity. *Heliyon*, 4(12), e01061. <https://doi.org/10.1016/j.heliyon.2018.e01061>
- Petrides, K. V., & Furnham, A. (2003). Trait emotional intelligence. Behavioural validation in two studies of emotion recognition and reactivity to mood induction. *European Journal of Personality*, 17, 39–57.
- Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, 19(3), 188–202. <https://doi.org/10.1016/j.hrmr.2009.03.007>