Expert and Target Scoring: Their relation, corresponding test instructions, and their effects on the construct validity of the video-based social understanding test (VSU)

Kristin Conzelmann and Panja Goerke

Department of Aviation and Space Psychology, DLR-German Aerospace Center, Sportallee 54a, 22335 Hamburg, Germany. kristin.conzelmann@gmx.de

This study investigated the relation between expert and target scoring of a video-based social understanding test (VSU) under two different types of instructions (internal and observer). The effects of the scoring methods and instructions on the VSU's construct validity were also examined. A total of 529 pilot applicants completed the VSU (some with internal and some with observer instructions), cognitive ability and knowledge tests, and a personality questionnaire. A subsample (*n* = 132) completed the VSU again with the other instructions and participated in an assessment center (AC). The two scores were moderately correlated; correlations decreased when the instructions were considered. Neither expert nor target scores showed convergent validity with AC variables; none of the scoring-instruction combinations showed significant associations with the remaining measures.

1. Introduction

Social understanding (also known as social inference, social interpretation, or social judgment) is the ability to understand social stimuli against the background of the corresponding social situation (O'Sullivan, Guilford, & DeMille, 1965; Wedeck, 1947). This involves the recognition of the mental states that are behind people's words (Moss, Hunt, Omwake, & Woodward, 1955), the comprehension of observed behaviors in the social context in which they occur (O'Sullivan & Guilford, 1966), and the decoding of social cues (Barnes & Sternberg, 1989). We regard social understanding as a cognitive ability that makes up the central part of social intelligence (SI) and define it as the ability to understand social information in a given situation and to interpret it correctly (Weis, Seidel, & Süß, 2006). Social understanding is seen as a potential-based construct (Guilford, 1967; Moss et al., 1955; Wedeck, 1947) that is considered to be a precondition for behaving in a socially competent manner. As with other cognitive constructs (e.g., academic intelligence), social understanding is best measured with performance tests. Studies indicate that social understanding can be measured reliably and is sufficiently distinguishable from academic intelligence and personality (Conzelmann, Weis, & Süß, 2013; Seidel, 2008; Weis, 2008). Comparable findings have also been obtained for emotional understanding, which represents emotional intelligence – EI (e.g., Austin, 2010; Freudenthaler & Neubauer, 2005; MacCann, 2010; Sharma, Gangopadhyay, Austin, & Mandal, 2013).

In line with other researchers (i.e., Davies, Stankov, & Roberts, 1998; Kang, Day, & Meara, 2005; Mayer, Salovey, & Caruso, 2000; Weis & Süß, 2005), we see SI as overlapping with the construct of EI when viewed as an ability (see Salovey & Mayer, 1990). Studies on the relation between SI and EI performance tests (i.e., Davies et al., 1998; Weis & Süß, 2007) are both scarce and hard to compare because of different theoretical underpinnings. Since theoretical conceptions of EI and SI usually ignored each other, contradictions resulted. For example, Salovey and Mayer (1990) originally devised 'emotional intelligence as a subset of SI' (p. 189). In later publications (Mayer & Salovey, 1993, 1997), EI was viewed as an extension of SI, and thereafter, SI was ignored. Barchard (2003) sees SI as part of EI without elaboration. According to Weis (2008), in comparing tasks from the most prominent measure of EI (i.e., the

International Journal of Selection and Assessment

Mayer–Salovey–Caruso Emotional Intelligence Test – Mayer, Salovey, & Caruso, 2002) with an SI cognitive ability model (Weis & Süß, 2005), SI focuses on others (i.e., emotions in social interactions), whereas EI includes the self and others. However, we view SI as involving an understanding of cognitions and behavior in addition to emotions in social situations (see Conzelmann et al., 2013). Nonetheless, significant data-based analyses of the relation between the two constructs are needed to consider the cognitive requirements of both constructs as well as the content domains in which they are represented (for more details, see Weis et al., 2006).

A major issue in dealing with tests of SI and EI is the problem of defining the correct answer (Mayer, Caruso, & Salovey, 2000). An appropriate scoring key is much more complex for social and emotional matters than for academic intelligence and knowledge. Academic intelligence tests apply logical rules (typically mathematical problems), semantic rules (the dictionary), scientific rules (truthfulness as determined by scientific studies), or knowledge (Guttmann & Levy, 1991) to unequivocally determine the correct answers to reasoning, spatial, memory, numerical, and verbal tasks (Matthews, Zeidner, & Roberts, 2002; Roberts, Zeidner, & Matthews, 2001). A common approach to measuring social abilities with a performance-based concept is the application of situational judgment tests (SJT), which are used as a realistic way to simulate everyday problem situations. They do not allow the formulation of unambiguously correct solutions that correspond to reality since both real-world social situations and individuals are ambiguous, dynamic, and complex. Thus, for these kinds of social and emotional ability tasks, there is no absolute right or wrong response to a question or situation. In the past, three main approaches for determining the correct answer to a social or emotional question have been implemented: group consensus scoring, expert scoring, and target scoring. However, each approach has both advantages and shortcomings.

1.1. Group consensus scoring

Group consensus scoring reflects the majority opinion of laypersons (i.e., test takers). Proportion-based consensus scoring is the most prominent consensus scoring approach and is also typically used in EI testing (Mayer et al., 2002). A score is allocated to a response according to the proportion of people endorsing that response (e.g., in a sample, if 50% choose response A as correct, 30% response B, and 20% response C, response A receives a score of .50, response B a score of .30, and C of .20). Nevertheless, there are various problems encountered with consensually scored responses (MacCann, Roberts, Matthews, & Zeidner, 2004): distributions cannot be normal since the majority must score the highest mark by definition. If tests are internally consistent, the same majority of test takers will score correct answers for each question, and this will result in a lack of discriminability of test takers of high and average ability. If scores are distributed evenly, different people will be scored with the correct answer for different items. Consequently, inter-item correlations will be low or negative. If group consensus scoring is applied to a difficult dichotomous item (if an item is difficult, this means that it is solved correctly by only a minority of test takers), a test taker with high ability may receive a low score (Matthews, Zeidner, & Roberts, 2005; Schulze, Wilhelm, & Kyllonen, 2007). Hence, a less able person who gives the same answer as the majority of test takers will receive a higher score. This means that an analysis of item difficulties is not possible (Legree, Psotka, Tremble, & Bourne, 2005; Schulze et al., 2007) and that scoring will depend on the mean level of ability in the sample. Considering these issues, consensus scoring is not suitable for scoring a performance-based test.

1.2. Expert scoring

In expert scoring, a specified group of experts determines the correct response to an item. Legree (1995) considers expert scoring to be a special case of consensus scoring. Experts are expected to have a richer knowledge base and to provide more reliable and accurate judgments, particularly in areas such as clinical psychology and physics (see Mayer & Geher, 1996; Mayer, Salovey, Caruso, & Sitenarios, 2001).

A challenge to expert scoring is the need to determine the criteria for expertise and the context of expertise since there are no absolute criteria by which to determine who is an expert. According to Weis (2008), an expert in SI is a specialist who understands the internal states of an individual better than any outside observer and better than the individual him or herself. Expert scoring has been criticized for two main reasons. First, experts' answers tend to suffer from specific problems such as response bias, stereotypical thinking, sympathy and similarity with the target, and implicit personality theories (Cline, 1964). Second, the distinction between expert scoring and consensus scoring seems to be a function of group size (Legree, 1995): Expert ratings may converge with the group consensus when enough experts are consulted (e.g., Roberts et al., 2001; see also Mayer et al., 2002). This emphasizes the importance of selecting experts who all have excellent knowledge of the relevant topic area.

Expert scoring seems to function better than consensus scoring. The majority of people are probably not always right when answering a social understanding item if we assume a normal distribution of the ability. An expert sample is a selected group of people who probably deviate from the consensus of an unselected group of

people. Items for which not all experts chose the same category but their answers were varied can be taken into account by calculating the Euclidean distance of the test taker's to the expert's view.

1.3. Target scoring

Target scoring is based on the notion that the target (i.e., creator of the item stimuli) has more information about his or her mental state (i.e., emotions, motivations, cognitions, intentions) than any outside observer (Mayer & Geher, 1996). The target determines the correct answers (i.e., the target him or herself decides which emotion is expressed by his or her voice). In a study by O'Sullivan (2007), targets were more accurate at judging their own emotional experiences, whereas acquaintances were more accurate at rating the target's behavior. However, targets may not be accurate reporters of their own emotions because of response biases or poor self-understanding, or they may report only pleasant emotions when they are actually feeling something else (Geher, Warner, & Brown, 2001). Thus, the reliability of the target person's answer is a very important precondition of the successful use of target scoring.

1.4. Relations between different scoring procedures

Since all scoring methods have their problems and perils of unreliability, Colvin and Bundick (2001) required the application of multiple criteria for judging accuracy. Several studies exist concerning the relation between expert and consensus-based scoring as well as between target and consensus-based scores. In summary, there is often a lack of correspondence between results obtained with different scoring methods. For example, Geher et al. (2001) found a correlation between target and consensus scores on the Emotional Accuracy Research Scale (EARS) of only .02. They explained the orthogonality of these measures by the observation that humans are not adept at knowing the actual reasoning that underlies their cognitive processes. Therefore, it makes sense that consensus and target measures yield different results. This is in line with Nisbett and Wilson (1977), who argued that more often than we believe, people are not privy to the actual reasoning underlying their decisions. Knowing what an individual thinks he or she is feeling may be quite different from knowing what others think the same individual feels (Mayer & Geher, 1996).

Findings with respect to the Magdeburg Test of Social Intelligence (MTSI, Süß, Seidel, & Weis, 2008) showed some correspondence between target and consensus scoring. In two studies with different groups totaling 317 test takers (students and a heterogeneous group of adults), correlations of $r = .61$ to .82 between target and consensus scoring were found for different modality-specific social understanding scales, that is, pictorial, auditory, and video-based (Seidel, 2008; Weis, 2008). Possible reasons for the diverging findings may have been the availability of objective social cues (e.g., tone of voice, gesture, and posture) in the task material of the MTSI social understanding tasks that were also used by the test takers to judge the respective scenes. In addition, the targets were carefully selected according to their openness and were well known by one of the test developers, conditions that should have increased honesty and reliability.

Weis (2008) observed that no investigations in the literature had reported agreement between target and expert scoring. This absence was addressed by the current study as we applied a newly developed video-based social understanding test (VSU), which, like the MTSI, contains objective social cues in the task material. These cues (e.g., voice, facial expression, body language, and contextual features) were expected to be available for both types of scoring. This was expected to result in some overlap between expert and target scoring. On the other hand, differences in scores were expected since experts tend to take a more objective observer perspective, whereas targets are subjectively involved and thus may suffer from a lack of self-insight (Nisbett & Wilson, 1977; Reilly & Doherty, 1992). Furthermore, different meta-analyses have shown that the self-other agreement of performance is low to medium (Conway & Huffcutt, 1997; Harris & Schaubroek, 1988; Heidemeier & Moser, 2009). Consequently, we expected low to moderate correlations between expert and target scores.

1.5. Different instructions for different scoring procedures

In earlier studies comparing expert and consensus scoring (e.g., Mayer et al., 2002; Roberts et al., 2001) or target and consensus scoring (e.g., Geher et al., 2001; Mayer & Geher, 1996), different scoring keys were applied to the same data to investigate the effects of different ways of scoring on the scores that were awarded. However, different perspectives and instructions are also involved when different scoring approaches are applied. If the test is evaluated by consensus scoring, the test taker should adopt a group perspective (answer according to the majority of the group). A test taker evaluated by expert scoring should give ratings from the perspective of an observer. An applicant evaluated by a target-scored test should give ratings from an internal perspective (answer according to the target's point of view). The scoring-specific instructions also affect the item difficulty. There can be different item difficulty values for the same item depending on the instructions if

an easy item means that most of the subjects chose the correct response and a difficult item means that only a few test takers answered correctly. But the interpretation of 'correct' depends on whether subjects are instructed to identify the answer that the majority of the tested group would give, to select the answer that experts would choose, or to answer like the target person would.

So far, research has not yet been conducted to investigate how these different perspectives, which are the results of these different instructions, influence the relations among different methods of scoring and the effects of the scores on convergent and discriminant validity criteria. However, Freudenthaler, Neubauer, and Haller (2008) examined the effects of maximum versus typical performance instructions on the convergent and discriminant validity of an EI performance test. They found that test takers scored higher on the EI performance test under maximum performance instructions. There were different relations of the emotional performance test to personality and cognitive ability measures depending on the instructions for completing it: using typical performance instructions, they found substantial relations between personality, depression, and life satisfaction scales, whereas when using maximum performance instructions, relations to cognitive ability were significant and also higher. Studies from other streams of research have also yielded evidence concerning how different perspectives result in different reactions. Bateson, Early, and Salvarani (1997) distinguished between an 'image other' perspective (imagining how another person feels) and an 'image self' perspective (imagining how you would feel in the other person's shoes). In an experiment, the 'image other' perspective produced empathy and eventually resulted in altruistic motivation, whereas the 'image self' perspective created empathy and eventually turned into personal distress and egoistic motivation. Thus, projecting oneself into different perspectives resulted in different motives and reactions. Epley and Caruso (2009) also observed interindividual differences in this ability.

Results such as these indicate that not considering the nature of the instructions may lead to inaccurate or misleading results. It follows that the different perspectives operationalized by the instructions may represent different tasks, which could even address different underlying abilities. Therefore, we expected that scores that took the instructions into account would differ more than scores that excluded the effects of the instructions.

1.6. Hypotheses and questions

The aim of this paper was, first, to investigate the overlap and distinctiveness of target and expert scoring of the VSU while taking into account the effects of instructions. A second aim was to examine how the different scoring procedures and instructions would affect the convergent (measured as the relation to assessment center performance) and discriminant construct validity of the test (measured as relations to basic cognitive ability, knowledge, and personality).

H1: Expert and target scoring of an identical item pool will be correlated to a moderate degree.

H2: Expert and target scores that take into consideration the respective instructions will be correlated to a lesser degree than scores that disregard the instructions.

The investigation of the effects of the instructions and scoring method on the construct validity of the VSU is rather exploratory. Convergent measures (AC performance) and divergent measures (basic cognitive ability and personality) were applied. Since both types of VSU scores are assumed to represent aspects of social understanding (either different or the same) and both rely on a performance test, we did not expect large differences between expert and target scores. However, we did not know of any other comparable study that has taken into account the effects of both scoring and instructions on validity.

2. The VSU and its scoring keys

The VSU items were developed according to the principles of SJTs (e.g., Kyllonen & Lee, 2005; McDaniel & Nguyen, 2001) and integrated the postdiction paradigm (O'Sullivan, 1983) into the scenario approach. Like typical SJTs, the VSU involves the (social) context and works with realistic scenarios. However, the VSU has to be distinguished from the standard SJT in the following ways: First, the situations and items are unscripted and entirely natural, real-life video scenes. Second, items are constructed differently and thus require different answer patterns. On an SJT, test takers are asked to assess the best behavior in a hypothetical situation and to either choose the most appropriate response from several alternatives or rank the responses. For the VSU, test takers have to interpret or infer a target person's thoughts, feelings, and relationships and judge answer alternatives according to their degree of appropriateness. The test taker's answer is compared with either an expert view (expert scoring) or the view of the target who was in the situation (target scoring). The test principle was rooted in the social understanding tasks of the MTSI (Süß et al., 2008). The MTSI is a SI test that, including social understanding, assesses social memory and social perception with veridical video-based, auditory, pictorial, and written material and, unlike other SI tests, involves the social context. It is described in detail in Conzelmann et al. (2013). Unlike the MTSI, the VSU (1) uses complete video scenes (audio and video) in-

stead of separating the auditory from the visual channel, (2) exclusively focuses on occupational scenes (in an aviation context), and (3) integrates findings from peer ratings into test development.

The VSU contains two scenarios dealing with two male pilot trainees (targets) from a well-respected German airline. Targets were volunteers who were chosen according to their scores on a personality scale measuring openness (Freiburger Persönlichkeitsinventar [Freiburg Personality Inventory]; Fahrenberg, Hampel, & Selg, 2010). Since the reliability of target scoring depends on the reliability of the targets' answers, the preferred targets were high on openness and low on social desirability so that they would be likely to provide information honestly. The potential targets were deliberately selected to differ in their personality profiles (NEO- five factor inventory [NEO-FFI], Borkenau & Ostendorf, 1993; Interpersonal Circumplex, Horowitz, Strauß, & Kordy, 2000) to ensure that the VSU would assess a more general ability that would not depend on a certain personality type (except for openness). However, it is unclear how both the similarity and sympathy of the test taker with respect to the target can influence the test result, but this was yet another reason for choosing targets with different personality profiles. The video recordings took place during the targets' typical everyday lives in the flight training school. All participants in a social situation (i.e., target persons, flight trainers, peers, and friends) gave their written consent to use the collected material and data for test development and research.

2.1. Development of target scoring

The targets answered questions about their mental state (i.e., emotions, cognitions, and relationships to people interacting in the corresponding situation) on a visual analog rating scale. For example, in one VSU scenario, two flight trainees provide feedback to one another. An emotion item example is 'How difficult is it for the target person to provide feedback to his peer?' with a rating scale ranging from 'not at all difficult' to 'extremely difficult.' A cognition item example is 'Why does the target laugh as he provides feedback to his peer?' (Item 1), 'He feels insecure.' (Item 2), 'He wants to soften his criticism.' Both items have to be rated on a scale ranging from 'does not apply at all' to 'totally applies.'

A peer who was well known by the target person and present in most of the recorded situations completed the same questions as the target from the target's perspective. The target person's answer was considered the 'correct' answer to the question (i.e., test item) as long as the peer agreed. In addition, the reliability of the target's rating was examined using objective social cues that were present in the item material and that provided hints about the answer. A lack of correspondence between the target's answer and the available cues in the material resulted in the exclusion of that item from the test. The target's answer was then transformed from the visual analogue rating scale by assigning a rating category to a target's score according to seven equally distributed steps on the analogue scale (7-point rating scale). For example, for the emotion item mentioned above, the target's answer fell in category '2,' meaning that it was rather easy to provide feedback to the peer (ranging from 1 = *extremely easy* to 7 = *extremely difficult*).

Finally, the two targets had to individually complete the test to ensure that the selection of material and questions corresponded to their real answers. When there was any lack of correspondence, the item was excluded. To compute the test taker's score on an item, first, the absolute difference (in either direction) between the target's answer and the test taker's answer was computed. Then, the test taker's score was always given a negative sign when it did not match the target's answer perfectly. In addition, the score was weighted by the greatest possible deviation in order to give each item the same weight.

2.2. Development of expert scoring

Experts were 18 experienced psychologists in the field of personnel selection; 15 were aviation psychologists, and three were consultants from other branches. All experts were nominated by the test developers because of their experience in interpreting human behavior. The experts were between the ages of 29 and 62 with a mean age of 38 ($SD = 9.5$); 72% of the experts were female. The experts were asked to provide their personal answers to the questions after watching realistic video scenes. The test taker's score on an item was determined by computing Euclidean distances, resulting in small positive scores when differences from the mean expert score were small, and in large positive scores when differences from the mean expert score were large. Using the standard deviation of the group ensured that a larger amount of variability in the experts' answers was included in the evaluation of a test taker's answer. The final score was formed by taking the square root of the term, such that better performances were indicated for test takers with smaller deviations from the experts' mean value (i.e., smaller scores).

2.3. Scenarios and instructions

Each scenario in the VSU involved six scenes encompassing between 2 and 10 items. The scenario was introduced with a short self-presentation by the target to allow test takers to become familiar with the target's voice and physical appearance. Each scene began with a short text that presented basic background information, which contained situational cues that were necessary for

understanding the scene. Test takers were required to judge the mental states of the target (i.e., the target's emotions, cognitions, and relationships to others) on the basis of the information provided (Costanzo & Archer, 1993; Moss et al., 1955; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979).

Depending on the scoring procedure to be used, test takers were asked to take different perspectives. The instructions to answer the questions from the target's perspective (internal perspective) accompanied the target scoring. For the expert scoring condition, the test takers were instructed to interpret the target's behavior from their own point of view (observer perspective).

3. Method

3.1. Procedure

The study took place during the *ab initio* pilot selection procedure at the DLR. Applicants completed a multistage selection procedure with four stages in total: (1) basic cognitive ability and knowledge tests, and an aviation-specific personality questionnaire, (2) an assessment center, (3) a fixed-base simulator test, (4) and an interview. The VSU was administered during the first (Sample 1) and the second (Sample 2) selection stages. Thus, only applicants who were successful in the first selection stage were tested again (the selection rate from the first to second stage of testing was 35.7%). Therefore, the variance in Sample 2 was more restricted than in Sample 1, which had already involved a very homogeneous group of people.

3.2. Study design and participants

To investigate the relation between the scoring methods and the effect of the instructions, we utilized a crossed design. Sample 1 consisted of 529 applicants for a pilot training program between the ages of 17 and 29 ($M = 20.46$, $SD = 2.58$); 86% were male. Subjects were randomly assigned to one of the two test conditions (internal perspective with target scoring $n = 247$ or observer perspective with expert scoring $n = 282$).

A subsample of $n = 132$ participants (Sample 2) completed the VSU again using the other perspective (internal $n = 76$; observer $n = 56$) if they had been successful in the previous selection stage. The time interval between the two administrations was 21.57 weeks on average ($SD = 9.88$ weeks) with a minimum of 6.71 weeks and a maximum of 59.71 weeks. The time interval varied considerably since (1) there was variability in the Stage 1 test dates; (2) there was variability in the Stage 2 test dates; (3) there were sometimes personal restrictions (i.e., since the selection procedure takes a long time in total, some applicants had decided to attend a university or go abroad in the meantime). The different sample sizes for the internal and observer instructions occurred by chance and can be ascribed to applicants dropping out in selection Stage 1 (for basic cognitive ability and knowledge tests, see 3.1). The mean age of the subsample was 20.61 years ($SD = 2.46$ years), and 90.9% were male.

3.3. Measures

3.3.1. Social understanding

The VSU (see Section 2) was used as a measure of social understanding in Samples 1 and 2.

3.3.2. Basic cognitive ability tests, knowledge tests, and grade point average (GPA)

Basic cognitive ability tests were administered in Sample 1 for the domains of numerical operations, memory, speed of information processing, and spatial orientation. Knowledge tests consisted of knowledge of technical information and English. These tests are well established in the selection system and are continually validated. Cronbach's α ranged from $\alpha = .76$ (numerical operations) to $\alpha = .97$ (speed of information processing). Performance on the VSU was also related to GPA.

3.3.3. Personality

The Temperament Structure Scales (TSS, Maschke, 1986) were used as an aviation-specific personality questionnaire in Sample 1. The scales measure personal styles/tendencies such as achievement motivation, extraversion, compromise, assertion, and empathy. The TSS uses a forced-choice item format. The internal consistency of the scales ranged from $\alpha = .61$ (achievement motivation, compromise) to $\alpha = .82$ (assertion) (Goeters, Timmermann, & Maschke, 1993).

3.3.4. Assessment Center (AC) performance

The AC was comprised of three exercises: role-play, a problem-solving group exercise, and a conflict-oriented group exercise. Each candidate was independently assessed on the behavioral dimensions leadership, cooperation, and commitment by two trained observers (an aviation psychologist and an airline pilot) using an observer rotation plan. No major differences were found between the ratings given by airline pilots and aviation psychologists. In the current sample, inter-rater agreement (*ICC 1,2*) was .70 (leadership), .62 (cooperation), and .72 (commitment) regardless of the raters' professions.

3.4. VSU scale building and calculation of internal consistency

The total scales were built for both Samples 1 and 2. Given that the appropriate procedure for evaluating the

score is to examine participants' answers in an instruction-specific manner (i.e., participants who were administered the observer instructions were evaluated by expert scoring), we built two scales that took the instructions into consideration (expert scoring with observer instructions and target scoring with internal instructions). In addition, to examine the differences between the types of scoring independent of the type of instructions as was done in past studies, an instruction-unspecific scale was built for each type of scoring (= 'both'; i.e., data were target or expert scored, excluding the effect of the instructions). Some of the items yielded negative item-total correlations. These items were eliminated from the scale. Next, items scored by target and expert scoring were synchronized in order to compare the scores that were computed on the same items with one another. In order to estimate the reduction in reliability because of the synchronization, item selection was also computed separately for each scale, target, and expert, resulting in optimized scores (if only target or expert scoring had been applied). We refrained from building separate scales for the different types of VSU items (cognition, emotion, relationship) because of their limited and varying numbers (e.g., six relationship items, 13 emotion items, 20 cognition items, one about the typicality) and insufficient Cronbach's α values (.30 to .45 in Sample 1; .31 to .59 Sample 2) for the final synchronized scale. Instead, we decided to aggregate these items into one total score.

4. Results

4.1. Psychometric properties

Table 1 presents the psychometric properties of the total synchronized scales for both samples. Compared with the optimum item pools, Cronbach's α as well as the number of selected items decreased. The optimized scores were different for target ($k = 51$; $\alpha = .71$ when the internal instructions were considered, $\alpha = .73$ when they were not) and expert scoring ($k = 58$, $\alpha = .70$ when the observer instructions were considered, $\alpha = .68$ when they were not). Cronbach's α values remained consistent for the target and expert scoring in Sample 2; however, the item-total correlation showed more variation. For the expert-scored VSU scale, there was only one item with a negative item-total correlation (both types of instructions and observer instructions); and for the target-scored VSU scale, there were three items with negative item-total correlations. Three of the four critical items were part of scenario A, and the majority of the items dealt with emotions.

Descriptive statistics from the basic ability and knowledge tests, personality scales, and AC performance can be found in Table 2. Correlations between all study variables can be found in the Table A1 (see Appendix A).

4.2. Relation between target and expert scoring

The examination of the relation between target and expert scores was conducted in two stages. First, the expert and target scores were correlated, excluding the influence of the specific instructions (as was done in previous studies on scoring issues). Second, the subsample who took the test twice (Stages 1 and 2) was further investigated with regard to how the scores from the first testing stage (which were either target or expert scores) were related to the scores from the second testing stage (which were expert scores for the group that had target scores in the first stage and target scores for the group that had expert scores in the first stage). The synchronized scale was used in both analyses.

The correlation[1] between target and expert scores was moderate in both Samples 1 ($r = -.47$, $p < .01$, $n = 529$) and 2 ($r = -.34$, $p < .01$, $n = 132$) and higher in Sample 1 when the type of instructions (observer vs. internal) was not considered. However, correlation coefficients decreased drastically when the type of instructions was considered. If both the target version of the VSU was applied with internal instructions in Stage 1 and test takers completed the VSU with expert scoring

Table 1. Psychometric properties of the total VSU synchronized scales (Samples 1 and 2)

Synchronized scale	Scoring	Instructions	Item count	*M*[a]	*SD*[a]	r_{it}-Range	α
Sample 1	Target	Both	40	−2.09	.41	.05 to .38	.69
		Internal	40	−2.08	.38	.05 to .38	.65
	Expert	Both	40	2.08	.77	.07 to .27	.62
		Observer	40	2.12	.78	.06 to .28	.64
Sample 2	Target	Both	40	−1.87	.38	−.13 to .48	.70
		Internal	40	−1.87	.39	−.15 to .50	.72
	Expert	Both	40	1.74	.67	−.07 to .37	.68
		Observer	40	1.70	.65	−.17 to .47	.68

[a]Means of target and expert scores cannot be compared directly since target scores are means of weighted differences from the targets' answer and expert scores are means of Euclidean distances from the experts' opinion.

Table 2. Descriptive statistics of the variables used to assess convergent and discriminant validity

Variable	Mean	*SD*	Min	Max	Scale
AC performance[b] (Stage 2 sample)					
Leadership	2.93	6.70	1.50	4.67	1 to 6
Cooperation	3.13	5.90	1.67	4.33	1 to 6
Commitment	2.81	6.49	1.50	4.33	1 to 6
Basic cognitive abilities and knowledge[a] (Stage 1 sample)					
English knowledge	0	.99	−2.94	2.41	−3 to +3
Technical knowledge	0	.92	−2.05	2.63	−3 to +3
Numerical operations	0	.86	−2.54	1.86	−3 to +3
Memory	0	.88	−2.47	2.16	−3 to +3
Speed of information processing	0	.91	−2.91	1.74	−3 to +3
Spatial orientation	0	.91	−2.22	1.24	−3 to +3
GPA	2.5	.63	1.0	3.8	1 to 4
Personality[b] (Stage 1 sample)					
Achievement motivation	8.15	3.10	.00	15.00	0 to 15
Extraversion	11.37	3.02	.00	15.00	0 to 15
Compromise	5.71	3.03	.00	15.00	0 to 15
Assertion	9.33	3.84	.00	15.00	0 to 15
Empathy	10.49	3.45	.00	15.00	0 to 15

Note: GPA = grade point average (corresponding to the final grade of the university entrance diploma); AC = assessment center. [a]Regression factor scores: mean = 0. [b]Raw scores.

with observer instructions in Stage 2, the correlation dropped to a nonsignificant level ($r = -.18$, $p > .05$, $n = 56$). The same occurred when the roles were reversed ($r = -.21$, $p > .05$, $n = 76$). This decrease was significant for the difference between the correlation coefficient in Sample 1 (both instructions) compared with the instruction-specific correlations ($Z = 2.38$, $n = 76$ and $Z = 2.28$, $n = 56$, both $ps < .05$, two-tailed testing).

No applicant worked on the same version of the test twice. The instructions always differed from the first to the second administration. Test takers completed the VSU either with the internal instructions first and the observer instructions second or vice versa. Typical retest reliability could not be assessed. Also, the target and expert scores could not be compared directly because target scores are difference values, whereas expert scores are Euclidean distance scores. However, when the specific instructions were not considered, test takers improved from the first to the second administration. Applying expert scoring to all applicants at Stage 1 and Stage 2 resulted in improved VSU scores at Stage 2 ($M_{stage1} = 2.04$, $SD = .70$; $M_{stage2} = 1.74$, $SD = .67$). The findings for the application of target scoring to all applicants at Stage 1 and Stage 2 were comparable ($M_{stage1} = -2.05$, $SD = .42$; $M_{stage2} = -1.87$, $SD = .38$). Because all applicants improved, this did not affect the correlation.

4.3. Relation between VSU performance and AC performance

There were no significant correlations between VSU target scores and AC performance in the assessed dimensions (see Table 3). The correlations with the VSU expert scores were also low. There was only one significant positive relation observed with the commitment shown in the AC. The more committed the applicants were, as evaluated with the AC tasks (see Section 3.3.4), the worse they performed on the expert-scored version of the VSU with observer instructions. Out of the AC dimensions, commitment is surely the least social and the least similar to social understanding. Nevertheless, we do not have an interpretation for this finding and prefer to first investigate whether this result can be confirmed in a similar study.

Table 3. Correlations between VSU scores and performance in the assessment center (Stage 2 sample)

	Target scoring		Expert scoring	
AC performance	Both instr ($n = 132$)	Internal instr ($n = 76$)	Both instr ($n = 132$)	Observer instr ($n = 56$)
Leadership	−.05	−.11	.02	.21
Cooperation	−.01	−.09	.04	.14
Commitment	.05	.07	.06	.28*

Note: *$p < .05$. instr = instructions; AC = assessment center.

4.4. Relation of VSU performance to basic cognitive abilities, knowledge, and personality

In both samples, the VSU scores were correlated with knowledge (technical and English), basic cognitive ability tests (i.e., spatial orientation, numerical operations,

Table 4. Relations of VSU scores with basic cognitive ability and knowledge tests, GPA, and personality scales

	Stage 1 sample				Stage 2 sample			
	Target scoring		Expert scoring		Target scoring		Expert scoring	
	Both instr ($n = 529$)	Internal instr ($n = 247$)	Both instr ($n = 529$)	Observ instr ($n = 282$)	Both instr ($n = 132$)	Internal instr ($n = 76$)	Both instr ($n = 132$)	Observ instr ($n = 56$)
Basic cognitive abilities and knowledge								
English knowledge	.02	.06	−.06	−.08	.00	−.02	.01	−.09
Technical knowledge	.00	−.07	.03	.04	.04	.04	−.03	−.07
Numerical operations	−.04	−.14*	.04	.00	.04	.14	.09	.21
Memory	.09*	.02	−.01	−.05	−.05	−.04	.16	.02
Inf. processing speed	.02	−.03	.03	.01	−.08	−.06	.16	.21
Spatial orientation	.05	−.03	−.01	−.02	−.04	−.02	.12	.07
GPA[a]	.01	.12	−.03	.02	.11	.21	−.18	−.17
Personality								
Achievement motivation	−.02	−.08	.01	−.11	−.05	−.14	.19*	.24
Extraversion	.04	.09	−.04	−.01	.00	−.16	.09	.07
Compromise	−.04	−.13*	.05	.03	−.06	−.06	.14	.23
Assertion	−.02	−.01	−.05	−.13*	−.09	−.18	−.04	.04
Empathy	−.01	−.04	−.08	−.05	.06	−.06	−.01	.00

Note: *$p < .05$. **$p < .01$. GPA = grade point average (corresponding to the final grade of the university entrance diploma); inf = information; observ = observer; instr = instructions. [a]The sample size deviates from the maximum possible sample size since some of the applicants did not have the university entrance diploma when they applied.

memory, and information processing speed), and the personality scales (see Table 4).

With the exception of a significant correlation in the Stage 1 sample between the target score and numerical operations (internal instructions), there were no meaningful positive relations with basic cognitive ability tests. The nonsignificant relation between VSU performance and the final grade on the university entry diploma complemented this finding (see Table 4).

Similarly, there were only a few significant relations with personality variables. In Sample 1, 'compromise' was correlated with the target-scored VSU performance (internal instructions) and 'assertion' with the expert-scored VSU performance (observer instructions). In Sample 2, 'achievement motivation' was correlated with the expert-scored VSU performance (both instructions).

5. Discussion

In this study, we investigated the relation between target and expert scoring with respect to the influence of the scoring-specific instructions. Furthermore, we analyzed the differential effects of the scoring procedures and instructions of the VSU on convergent and discriminant validity.

5.1. Relation between expert and target scoring of the VSU depending on the instructions

As hypothesized, target and expert scores were moderately correlated (thus confirming *H1*). A possible reason for the difference between the scores may be the lack of the target's self-insight (see Nisbett & Wilson, 1977; Reilly & Doherty, 1992) compared with the more objective expert ratings. Different instructions considerably influenced the relation between the two scoring methods. The correlation between target- and expert-scored VSU performances showed an obvious decrease when the type of instructions was also considered (thus confirming *H2*). This may occur because the test taker who completes the VSU from an observer perspective is actually performing a different task than the test taker who completes the VSU from an internal perspective, and these different tasks may measure different underlying constructs. There is some confirmation of the influence of different perspectives in other areas of research; for example, different perspectives resulting in different underlying motives and reactions (Bateson et al., 1997). Also, neuropsychological findings indicate that different areas of the brain become involved according to whether a person's own perspective or another person's perspective is taken (Ruby & Decety, 2004).

5.2. Convergent validity depending on scoring and instructions

Independent of the instructions and the type of scoring, there were no meaningful relations of the VSU with AC performance. First, not many studies have investigated the relations between social understanding measured with a performance test and AC performance. Runde and Etzel (2004) found a correlation of $r = .35$ between AC performance and performance on their video-based

expert-scored social understanding test called VISION. However, this measure differs from the VSU in important aspects. VISION involves video scenes that are acted out, and therefore, the process of item generation was completely different from the VSU. There are several possibilities that may account for why the VSU did not show a meaningful relation with AC performance in this study. First, differences in measures may account for the lack of correlation. Whereas the VSU is a psychometric test, AC measures involve behavioral observation. Second, there are differences in the underlying construct as the VSU is intended to evaluate the cognitive precondition for socially competent behavior, whereas the AC assesses the behavior itself. Third, the VSU was developed on the basis of real-life situations, whereas the AC makes use of constructed situations. AC exercises make use of obvious social scenarios, such as extreme conflicts. By contrast, the VSU presents subtle social situations such as cooperative social everyday-life situations. Fourth, the findings may also be due to the specific preselected sample of pilot applicants. They all held a university entrance diploma and showed high technical interests. The selection effect was even greater in the AC subsample since the participants were among the 35.7% of applicants who passed the academic intelligence and knowledge test in selection Stage 1. Thus, the variance was restricted and may have played a role in reducing the strength of the correlations.

5.3. Discriminant validity depending on scoring and instructions

Significant relations of the VSU with cognitive ability, GPA, and personality were scarce, independent of the type of scoring method and instructions.

The lack of meaningful relations between the VSU scores and both basic cognitive ability and knowledge tests are somewhat similar to findings obtained with the MTSI (see Conzelmann et al., 2013). In a confirmatory factor analysis, social understanding measured with the MTSI was separate from academic intelligence measured with the Berlin Intelligence Structure Test (Jäger, Süß, & Beauducel, 1997). Mayer and Geher (1996) also did not report significant relations between EARS target scores and self-reported Scholastic Aptitude Test scores. The lack of significant correlations between the VSU scores and GPA, which is usually associated with performance in basic cognitive ability measures, supports this hypothesis. However, whether the VSU is indeed separate from academic intelligence tests, has to be investigated with factor analytic techniques and other measures of academic intelligence.

A lack of correlations between social understanding and personality is a common finding in the field of SI (see Brown & Anthony, 1990; Riggio, Messamer, & Throckmorton, 1991). It remains an open question whether social understanding and personality constructs are independent or whether this result has to be attributed to the application of different methods (performance test vs. self-report measure).

5.4. Conclusions and future research needs

Concerning general research on scoring issues in the field of social understanding, this study made clear that the type of instructions must be considered when conducting studies and interpreting their results. Taking the effects of scoring and instructions into account may help to explain the divergent findings reported across research studies. The criterion should be considered before making a final decision about which scoring method to use. Putting oneself in the position of a target and thinking, feeling, and behaving from the target's point of view may be required for assessing the performance of actors. Yet, a general ability to interpret other people's relationships, cognitions, and feelings may be useful for assessing the suitability of someone applying for a social profession such as a teacher, salesperson, lawyer, or social worker.

Concerning the application of the VSU in the practice of pilot selection, the use of expert scoring with observer instructions is recommended. Working successfully as a pilot requires that one is able to interpret the other pilot's and crew members' intentions, feelings, and behavior from an observer perspective instead of putting oneself in their shoes. Being able to interpret behavior the way the VSU requires should not only result in fewer human errors in aviation but also in positive social and emotional effects (Arriaga & Rusbult, 1998; Epley & Caruso, 2009) such as better and more satisfactory relationships among crew members.

The expert scoring procedure that was applied in this study is restricted since experts were not involved in the item development, but the items were developed on the basis of the targets' answers, and distractors were generated by the test developers. Thus, the appropriate answer from the experts' perspective might be missing. However, this does not affect the task of the VSU, which is to judge the target person's cognitions, emotions, and behavior according to specified aspects that are prevalent in the item alternatives. It is not necessarily important that the 'correct' answer is included.

The expert scoring method reported in this study should be validated by another preferentially gender-balanced expert sample including experts from different professions that deal with the interpretation of others' behavior. In future studies, similarity to and sympathy with the target should be controlled because Davis, Conklin, Smith, and Luce (1996) found indications that a person's self-representation, particularly with regard to positive traits, influences the evaluation of the target.

In the future, it would be worthwhile to collect information on convergent validity with appropriate measures. Therefore, the VSU should be related to the MTSI and other tests that measure the same or related constructs such as the Situational Test of Emotion Management and Emotion Understanding (MacCann, 2006) and the Face Recognition Test (Wilhelm et al., 2010).

To demonstrate the usefulness of the VSU for personnel selection, there is a need to relate it to appropriate criteria collected in job-relevant social situations.

The VSU itself needs to undergo further development. In Sample 2, very few item-total correlations were still negative. Why these items worked well only in Sample 1 has to be examined further. Target-scored items were more susceptible to negative item-total correlations (see also Conzelmann et al., 2013). Despite the advantages of target scoring, the quality of the target's answer might be impaired by two aspects: First, targets need a high level of self-reflection to identify their real mental states and then to adequately communicate these states. Second, a target might have unintentionally or intentionally been drawn to the tendency to respond in a socially acceptable manner (MacCann et al., 2004; Mayer & Geher, 1996).

This study yielded the first findings about the relation between expert and target scores. Moreover, it clarified the idea that scoring issues need to be examined comprehensively, taking into account the test instructions.

Note

1. The negative weighting of correlation coefficients was according to our expectations.

References

Arriaga, X. B., & Rusbult, C. E. (1998). Standing in my partner's shoes: Partner perspective taking and reactions to accommodative dilemmas. *Personality and Social Psychology Bulletin*, 24, 927–949.

Austin, E. J. (2010). Measurement of ability emotional intelligence: Results for two new tests. *British Journal of Psychology*, 101, 563–578.

Barchard, K. A. (2003). Does emotional intelligence assist in the prediction of academic success? *Educational and Psychological Measurement*, 63, 840–858.

Barnes, M. L., & Sternberg, R. J. (1989). Social intelligence and decoding of nonverbal cues. *Intelligence*, 13, 263–287.

Bateson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imagining how you would feel. *Personality and Social Psychology Bulletin*, 23, 751–759.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae [NEO-Five-Factor-Inventory (NEO-FFI) according to Costa and McCrae]*. Göttingen: Hogrefe.

Brown, L. T., & Anthony, R. G. (1990). Continuing the search for social intelligence. *Personality and Individual Differences*, 11, 463–470.

Cline, V. B. (1964). Interpersonal perception. In B. A. Maher (Ed.), *Progress in experimental personality research* (pp. 221–284). New York: Academic Press.

Colvin, C. R., & Bundick, M. J. (2001). In search of the good judge of personality: Some methodological and theoretical concerns. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity* (pp. 47–65). Mahwah, NJ: LEA.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analytic analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, 10, 331–360.

Conzelmann, K., Weis, S., & Süß, H.-M. (2013). New findings about social intelligence: Development and application of the Magdeburg Test of Social Intelligence (MTSI). *Journal of Individual Differences*, 34, 119–137.

Costanzo, M., & Archer, D. (1993). *The interpersonal perception task-15 (ipt-15) [videotape]*. Berkeley, CA: University of California Extension Media Center.

Davies, M., Stankov, L., & Roberts, R. D. (1998). Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology*, 75, 989–1015.

Davis, M. H., Conklin, L., Smith, H., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: A merging of self and other. *Journal of Personality and Social Psychology*, 70, 713–726.

Epley, N., & Caruso, E. M. (2009). Chapter 20: Perspective taking: Misstepping into others' shoes. In K. D. Markman, W. M. P. Klein, & J. A. Suhr (Eds.), *Handbook of imagination and mental simulation* (pp. 295–312). New York: Psychology Press.

Fahrenberg, J., Hampel, R., & Selg, H. (2010). *Das Freiburger Persönlichkeitsinventar FPI. Handanweisung* (8. Überarbeitete und neunormierte Aufl.) *[The Freiburg Personality Inventory FPI. Technical manual.]*. Göttingen: Hogrefe.

Freudenthaler, H. H., & Neubauer, A. C. (2005). Emotional intelligence: The convergent and discriminant validities of intra- and interpersonal emotional abilities. *Personality and Individual Differences*, 39, 569–579.

Freudenthaler, H. H., Neubauer, A. C., & Haller, U. (2008). Emotional intelligence: Instruction effects and sex differences in emotional management abilities. *Journal of Individual Differences*, 29, 105–115.

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

Goeters, K.-M., Timmermann, B., & Maschke, P. (1993). The construction of personality questionnaires for selection of aviation personnel. *Journal of Aviation Psychology*, 3, 123–141.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Guttmann, L., & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence*, 15, 79–103.

Harris, M. M., & Schaubroek, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43–62.

Heidemeier, H., & Moser, K. (2009). Self-other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology*, 94, 353–370.

Horowitz, L. M., Strauß, B., & Kordy, H. (2000). *Inventar zur Erfassung Interpersonaler Probleme (IIP-D) (2.,überarb. Und neunormierte Aufl.) [German version of the Inventory of Interpersonal Problems (IIP-D) (2nd edition)]*. Göttingen: Beltz.

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-test. Form 4 [Berlin intelligence structure-test. Version 4]*. Göttingen: Hogrefe.

Kang, S., Day, J. D., & Meara, N. M. (2005). Social and emotional intelligence: Starting a conversation about their similarities and differences. In R. Schulze & R. D. Roberts (Eds.), *International handbook of emotional intelligence* (pp. 91–105). Göttingen: Hogrefe.

Kyllonen, P. C., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 11–25). Thousand Oaks, CA: Sage.

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

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *An international handbook of emotional intelligence* (pp. 156–179). Göttingen: Hogrefe.

MacCann, C. (2006). *New Approaches to Measuring Emotional Intelligence: Exploring Methodological Issues with Two New Assessment Tools*. Unpublished doctoral thesis, University of Sydney, Australia.

MacCann, C. (2010). Further examination of emotional intelligence as a standard intelligence: A latent variable analysis on fluid intelligence, crystallized intelligence, and emotional intelligence. *Personality and Individual Differences*, 49, 490–496.

MacCann, C., Roberts, R. D., Matthews, G., & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance-based emotional intelligence (EI) tests. *Personality and Individual Differences*, 36, 645–662.

Maschke, P. (1986). *Temperament-Struktur-Skalen (TSS). Testmanual. [Temperament Structure Scales (TSS). Test manual]*. Forschungsbericht DFVLR-FB 86-58. Köln: Deutsche Forschungsanstalt für Luft- und Raumfahrt.

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

Matthews, G., Zeidner, M., & Roberts, R. D. (2005). Emotional intelligence: An elusive ability? In O. Wilhelm & R. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 79–100). Thousand Oaks, CA: Sage.

Mayer, J. D., Caruso, D., & Salovey, P. (2000). Selecting a measure of emotional intelligence: The case for ability scales. In R. Bar-On & J. D. A. Parker (Eds.), *The handbook of emotional intelligence* (pp. 320–342). New York: Jossey-Bass.

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

Mayer, J. D., & Salovey, P. (1993). The intelligence of emotional intelligence. *Intelligence*, 17, 433–442.

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 3–31). New York: Basic Books.

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

Mayer, J. D., Salovey, P., & Caruso, D. R. (2002). *The Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT): User's manual*. Toronto: Multi-Health Systems.

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

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 9, 103–113.

Moss, F. A., Hunt, T., Omwake, K. T., & Woodward, L. G. (1955). *Manual for the George Washington University series social intelligence test*. Washington, DC: The Center for Psychological Service.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.

O'Sullivan, M. (1983). Measuring individual differences. In J. M. Wiemann & R. P. Harrison (Eds.), *Nonverbal interaction* (pp. 243–270). Beverly Hills, CA: Sage.

O'Sullivan, M. (2007). Trolling for trout, trawling for tuna: The methodological morass in measuring emotional intelligence. In G. Matthews, M. Zeidner, & R. D. Roberts (Eds.), *The science of emotional intelligence: Knowns and unknowns* (pp. 258–287). Oxford: Oxford University Press.

O'Sullivan, M., & Guilford, J. P. (1966). *Six Factor Test of Social Intelligence: Manual of instructions and interpretations*. Beverly Hills, CA: Sheridan Psychological Services.

O'Sullivan, M., Guilford, J. P., & DeMille, R. (1965). *The measurement of social intelligence*. Los Angeles, CA: University of Southern California.

Reilly, B. A., & Doherty, M. E. (1992). The assessment of self-insight in judgment policies. *Organizational Behavior and Human Decision Process*, 53, 285–309.

Riggio, R. E., Messamer, J., & Throckmorton, B. (1991). Social and academic intelligence: Conceptually distinct but overlapping constructs. *Personality and Individual Differences*, 12, 695–702.

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

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore, MD: Johns Hopkins University Press.

Ruby, P., & Decety, J. (2004). How would you feel versus how do you think she would feel? A neuroimaging study of perspective-taking with social emotions. *Journal of Cognitive Neuroscience*, 16, 988–999.

Runde, B., & Etzel, S. (2004). VISION (Videobasierte Identifikation Sozialer Intelligenz-Online) [Video-Based Identification of Social Intelligence – Online]. In W. Sarges & H. Wottawa (Hrsg.) (Eds.), *Handbuch wirtschaftspsychologischer Testverfahren (2., überarbeitete und erweiterte Auflage) [Handbook of business psychological tests]* (pp. 823–830). Lengerich: Pabst.

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition and Personality*, 9, 185–211.

Schulze, R., Wilhelm, O., & Kyllonen, P. C. (2007). New approaches to assessing emotional intelligence. In G.

Matthews, M. Zeidner, & R. D. Roberts (Eds.), *Emotional intelligence: Knowns and unknowns* (pp. 199–229). Oxford: Oxford University Press.

Seidel, K. (2008). *Assessment of social and auditory intelligence: New perspectives and approaches*. Lengerich: Pabst.

Sharma, S., Gangopadhyay, M., Austin, E., & Mandal, M. K. (2013). Development and validation of a situational judgment test of emotional intelligence. *International Journal of Selection and Assessment*, 21, 57–73.

Süß, H.-M., Seidel, K., & Weis, S. (2008). Neue Wege zur leistungsbasierten Erfassung sozialer Intelligenz und erste Befunde [New ways of performance-based measurement of social intelligence and first results]. In W. Sarges & D. Scheffer (Hrsg.) (Eds.), *Innovative Ansätze für die Eignungsdiagnostik [Innovative approaches for personnel assessment]* (pp. 129–143). Göttingen: Hogrefe.

Wedeck, J. (1947). The relationship between personality and 'psychological ability'. *British Journal of Psychology*, 37, 133–151.

Weis, S. (2008). *Theory and Measurement of Social Intelligence as a Cognitive Performance Construct*. Doctoral dissertation, Otto von Guericke Universität Magdeburg, Germany.

Weis, S., Seidel, K., & Süß, H.-M. (2006). Messkonzepte sozialer Intelligenz – Literaturübersicht und Ausblick [Measurement concepts of social intelligence – Review and future perspectives]. In R. Schulze & R. D. Roberts (Hrsg.) (Eds.), *Emotionale Intelligenz. Ein internationales Handbuch [Emotional intelligence. An international handbook]* (pp. 213–234). Göttingen: Hogrefe.

Weis, S., & Süß, H.-M. (2005). Social intelligence – a review and critical discussion of measurement concepts. In R. Schulze & R. D. Roberts (Eds.), *An international handbook of emotional intelligence* (pp. 203–230). Göttingen: Hogrefe.

Weis, S., & Süß, H.-M. (2007). Reviving the search for social intelligence. *Personality and Individual Differences*, 42, 3–14.

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces – one element of social cognition. *Journal of Personality and Social Psychology*, 99, 530–548.

Appendix A

Table A1. Correlations between all study variables

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. S1_T_B	1 (529)										
2. S1_T_I	1 (247)	1 (247)									
3. S1_E_B	−.47** (529)	−.50** (247)	1 (529)								
4. S1_E_O	−.45** (282)	–	1 (282)	1 (282)							
5. S2_T_B	.74** (132)	.71** (56)	−.32** (132)	−.21 (76)	1 (132)						
6. S2_T_I	–	–	−.21 (76)	−.21 (76)	1 (76)	1 (76)					
7. S2_E_B	−.18* (132)	−.18* (56)	.42** (132)	.30** (76)	−.34** (132)	−.34** (76)	1 (132)				
8. S2_E_O	−.18 (56)	−.18 (56)	.58** (56)	–	−.35** (56)	–	1 (56)	1 (56)			
9. English knowledge	.02 (529)	.02 (247)	−.06 (529)	−.08 (282)	.00 (132)	−.02 (76)	.01 (132)	−.09 (56)	1 (529)		
10. Technical knowledge	.00 (529)	−.07 (247)	.03 (529)	.04 (282)	.04 (132)	.04 (76)	−.03 (132)	−.07 (56)	−.34** (529)	1 (529)	
11. Numerical operations	−.04 (529)	−.14* (247)	.04* (529)	.00 (282)	.04 (132)	.14 (76)	.04 (132)	.04 (56)	.33** (529)	.53** (529)	1 (529)
12. Memory	.09* (529)	.02 (247)	−.01 (529)	−.05 (282)	−.05 (132)	−.04 (76)	.16 (132)	.02 (56)	.35** (529)	.47** (529)	.54** (529)
13. Inf. processing speed	.02 (529)	−.03 (247)	.03 (529)	.01 (282)	−.08 (132)	−.06 (76)	.16 (132)	.21 (56)	.29** (529)	.39** (529)	.50** (529)
14. Spatial orientation	.05 (529)	−.03 (247)	−.01 (529)	−.02 (282)	−.04 (132)	−.02 (76)	.12 (132)	.07 (56)	.32** (529)	.49** (529)	.55** (529)
15. GPA[a]	.01 (421)	.12 (209)	−.03 (412)	.02 (212)	.11 (111)	.21 (61)	−.18 (111)	−.17 (50)	−.33** (421)	−.32** (421)	−.40** (421)
16. Achievement motivation	−.02 (529)	−.08 (247)	.01 (529)	−.11 (282)	−.05 (132)	−.14 (76)	.19* (132)	.24 (56)	−.06 (529)	−.05 (529)	.06 (529)
17. Extraversion	.04 (529)	.09 (247)	−.04 (529)	−.01 (282)	.00 (132)	−.16 (76)	.09 (132)	.07 (56)	−.10* (529)	−.13** (529)	−.02 (529)
18. Compromise	.04 (529)	−.13* (247)	.05 (529)	.03 (282)	−.06 (132)	−.06 (76)	.14 (132)	.23 (56)	−.01 (529)	−.13** (529)	−.09* (529)
19. Assertion	−.02 (529)	−.01 (247)	−.05 (529)	−.13* (282)	−.09 (132)	−.18 (76)	−.04 (132)	.04 (56)	−.03 (529)	−.01 (529)	−.01 (529)
20. Empathy	−.01 (529)	−.04 (247)	−.08 (529)	−.05 (282)	.06 (132)	−.06 (76)	−.01 (132)	−.00 (56)	−.02 (529)	−.09* (529)	−.02 (529)
21. Leadership	−.02 (132)	.15 (56)	−.01 (132)	−.06 (76)	−.05 (132)	−.11 (76)	.02 (132)	.21 (56)	−.08 (132)	.04 (132)	.21* (132)
22. Cooperation	.01 (132)	.21 (56)	.01 (132)	.07 (76)	−.01 (132)	−.09 (76)	.04 (132)	.14 (56)	−.12 (132)	.04 (132)	−.11 (132)
23. Commitment	.05 (132)	.08 (56)	−.02 (132)	−.09 (76)	.05 (132)	.07 (76)	.06 (132)	.28 (56)	−.09 (132)	−.01 (132)	.19* (132)

	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.
1. S1_T_B											
2. S1_T_I											
3. S1_E_B											
4. S1_E_O											
5. S2_T_B											
6. S2_T_I											
7. S2_E_B											
8. S2_E_O											
9. English knowledge											
10. Technical knowledge											
11. Numerical operations											
12. Memory	1 (529)										
13. Inf. processing speed	.82** (529)	1 (529)									
14. Spatial orientation	.80** (529)	.80** (529)	1 (529)								
15. GPA[a]	−.36** (421)	−.31** (421)	−.32** (421)	1 (421)							
16. Achievement motivation	.08 (529)	.10* (529)	.03 (529)	−.29** (421)	1 (529)						
17. Extraversion	−.08 (529)	−.05 (529)	−.12** (529)	−.03 (421)	.19** (529)	1 (529)					
18. Compromise	−.09* (529)	−.12 (529)	−.12** (529)	.14** (421)	−.05 (529)	.05 (529)	1 (529)				
19. Assertion	.05 (529)	.08 (529)	.04 (529)	−.02 (421)	.20** (529)	.28** (529)	.21** (529)	1 (529)			
20. Empathy	−.06 (529)	−.03 (529)	−.08 (529)	.00 (421)	.28** (529)	.35** (529)	.06 (529)	.12* (529)	1 (529)		
21. Leadership	.13 (132)	.16 (132)	.11 (132)	−.32** (132)	.10 (132)	.18* (132)	−.12 (132)	.05 (132)	.16* (132)	1 (132)	
22. Cooperation	.05 (132)	.12 (132)	.00 (132)	−.13 (132)	.16* (132)	.15 (132)	−.15 (132)	−.20* (132)	.26** (132)	.47** (132)	1 (132)
23. Commitment	.18 (132)	.12 (132)	.05 (132)	−.33** (111)	.08 (132)	.17* (132)	−.14 (132)	.00 (132)	.21** (132)	.84** (132)	.49** (132)

Note: *$p < .05$. **$p < .01$. S1 = Study 1; S2 = Study 2; T = target scoring; E = expert scoring; I = internal instructions; O = observer instructions; Inf. = information; GPA = grade point average (corresponding to the final grade on the university entrance diploma). [a]The sample size deviates from the maximum possible sample size because some of the applicants did not have the university entrance diploma when they applied.