Journal of Research in Personality 97 (2022) 104182

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Validation of a performance measure of broad interpersonal accuracy☆

Nele Dael [a,*,1], Katja Schlegel [b,1], Adele E. Weaver [c], Mollie A. Ruben [d], Marianne Schmid Mast [e]

[a] Department of Organizational Behavior, University of Lausanne, Switzerland
[b] Institute of Psychology, University of Bern, Switzerland
[c] Department of Psychology, University of Maine, United States
[d] Department of Psychology, University of Maine, United States
[e] Department of Organizational Behavior, University of Lausanne, Switzerland

ARTICLE INFO

Keywords:
Interpersonal accuracy
Social perception ability
Test
Assessment
Individual differences

ABSTRACT

Accurately reading others' emotions, personality, intentions etc. (interpersonal accuracy, IPA) is crucial to successful interpersonal interactions. However, most existing tests to measure IPA focus on people's ability to recognize emotions and do not specifically target the workplace. The newly developed WIPS (Workplace Interpersonal Perception Skill) test assesses multiple aspects of interpersonal accuracy using brief video segments for which test-takers are asked to assess personality, intentions, future social behavior, thoughts, situational affect and social attributes of the targets in the video. Different criteria such as actual behavior shown were used to establish the correct answers in multiple-choice questions. Seven studies that subsequently tested the psychometric properties of a large item pool in English, French, and German are presented. The WIPS is unidimensional, shows acceptable internal consistency, and correlates in expected ways with emotion recognition, personality judgment accuracy, and a variety of other measures. Higher WIPS scores also predicted membership as well as leadership in student groups (e.g., in volunteer and music-oriented groups). These results contribute to the integration of various research strands under the broader IPA construct. The WIPS also complements existing, more specific tests and represents a useful tool for research and practice in the organizational field and beyond.

1. Introduction

Wanting to understand others occupies us at all levels of our daily life. Measuring how accurately one can guess others' thoughts, feelings, personality, or intentions seems to be a particularly popular recreational activity, considering the abundance of tests, pop quizzes, games, and even television shows on the topic (such as "Lie to me", or the Belgian and Dutch game show "De Mol" [The Mole]. Being good at reading others is also a precondition for successful social interaction (Custrini & Feldman, 1989; Nowicki & Duke, 2001), for instance in the workplace (Elfenbein, Foo, et al., 2007; Elfenbein, Polzer, et al., 2007; Hall, Andrzejewski & Yopchick, 2009; Schmid Mast & Latu, 2016).

Research shows that individuals who are more socially perceptive create more benefits and successful social interactions for themselves as well as for their interaction partners (e.g., supervisor ratings, sales numbers, likeability ratings from peers) (see *meta*-analyses by Elfenbein, Foo, et al., 2007; Hall et al., 2009; Momm et al., 2015; Schmid Mast & Hall, 2018; Human et al., 2020). For example, salespeople who accurately recognize emotions from others' nonverbal behavior earn higher salaries and interpersonally accurate car dealers sell more cars per year (Byron, Terranova & Nowicki, 2007). In job contract negotiations, recruiters with higher emotion recognition ability were perceived as more cooperative by their negotiation counterpart and created higher joint and individual gains (Schlegel, Mehu, van Peer, & Scherer, 2018). Furthermore, accuracy in judging others is studied in many other fields including socio-emotional development across childhood (e.g., Castro et al., 2015), aging and emotion (e.g., Sullivan & Ruffman, 2004), and research on couples and relationship satisfaction (e.g., Simpson et al.,

Abbreviations: IPA, Interpersonal Accuracy.

☆ This research was supported in part by the Swiss Innovation Agency, Innosuisse project nr. 28376.1 PFES-ES. We wish to thank the project's industrial partner, Vima Link, Martigny, in particular its founder, Raphaël Héraïef, for supporting the empirical work. No conflicts of interest are declared. The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

* Corresponding author at: Institute for Management Development (IMD), Chemin de Bellerive 23, P.O. Box 915, CH-1001 Lausanne, Switzerland.
E-mail address: nele.dael@unil.ch (N. Dael).

[1] These authors equally contributed to the research and the manuscript.

https://doi.org/10.1016/j.jrp.2021.104182
Received 30 March 2021; Received in revised form 14 December 2021; Accepted 18 December 2021
Available online 24 December 2021

2003), generally showing advantages for more accurate perceivers.

Research into the accuracy of social perception or inferences about others has been summarized using the term interpersonal accuracy (IPA), which was coined in 2016 by Hall, Schmid Mast, and West (2016). Such inferences can be made about another person's emotion, intention, personality, truthfulness, thoughts, future social behavior, status, and many more traits and states. Importantly, although research focusing on single types of traits or states (e.g., judging emotion or personality) has had a long tradition and produced a rich literature, these different strands of accuracy research have been largely independent from each other and used different measurement approaches and terminology (Bernieri, 2001). For instance, nonverbal sensitivity has often been used by researchers focusing on accurate judgments of affect (e.g., Rosenthal, DiMatteo, Rogers, & Archer, 1979), empathic accuracy has been used in relation to a specific method to measure accurate inference of others' thoughts and feelings (e.g., Ickes, 2001), and judgmental accuracy has frequently been used when referring to accurate inferences of personality traits (e.g., Letzring, 2008).

Schlegel, Boone, and Hall (2017) found small positive *meta*-analytic correlations between accuracy measures from different IPA subdomains, suggesting that IPA may be a set of loosely connected but largely distinct skills. That is, people who are good at detecting others' emotions may not necessarily be good judges of others' personality. On the other hand, these researchers pointed out that measurement approaches differed widely between the domains, several domains lacked standard tests to measure accuracy, and the internal consistency of many of the standard tests was quite low with an average reported Cronbach's alpha of 0.48. Schlegel et al. (2017) argued that these aspects could have led to an underestimation of the existence of a common core IPA skill. The paper concludes with the recommendation to unify measurement approaches and to develop more psychometrically sound standard tests of broad IPA.

1.1. Existing standard measures of IPA

Standard tests of IPA differ widely in terms of assessment paradigm, cue modalities, content domains, and specificity. A first broad distinction is made between self-report questionnaires and performance-based tests of IPA. Self-report questionnaires such as the Social Skills Inventory (Riggio, 1986, 2005; Riggio & Carney, 2003) require the participants to assess their own interpersonal accuracy. However, self-assessed interpersonal skill is subject to self-presentation bias and tends to be overestimated (Ames & Kammrath 2004), and accuracy depends on an individual's self-awareness of their social skills. In addition, self-assessed interpersonal skill correlates only weakly with objective, performance-based IPA tests (Hall et al., 2009).

In the second paradigm, IPA is measured by presenting participants with pictures, videos, or audio recordings of target individuals and asking them to judge each target's trait or state of interest (e.g., to what extent the target is extraverted or which emotion was being expressed). Participants' responses are then scored based on predefined criteria, for instance targets' self-reported personality traits or the instructions that targets had received (e.g., the emotion they were asked to express). Although such performance-based tests do not suffer from the drawbacks of self-report measures, they have some other limitations, some of which we attempt to address in the new WIPS test.

First, many tests have relied on static photographs of faces (e.g., Nowicki & Duke, 1994) which arguably tap only a small fraction of everyday nonverbal communication (Hall, 1978). Recently developed multimodal tests indeed appear to capture IPA more comprehensively (Bänziger, Grandjean & Scherer, 2009; Schlegel, Grandjean & Scherer, 2014), and may therefore more accurately reflect how we naturally integrate diagnostic cues from face, voice and body when inferring states and traits from others. The WIPS also combines these different modalities.

Second, most existing tests focus on measuring emotion recognition (e.g., Diagnostic Analysis of Nonverbal Accuracy, DANVA; Nowicki & Duke, 1994; Multimodal Emotion Recognition Test, MERT; Bänziger et al., 2009; Geneva Emotion Recognition Test, GERT; Schlegel et al., 2014). However, IPA is a much broader construct than just emotion recognition, including content domains such as personality (judging others' traits), situational affect (inferring a person's current situation), deception (detecting falsehood), thoughts and feelings, and social attributes (inferring others' social group membership and social characteristics) (Hall et al., 2016; Schlegel et al., 2017). In everyday life, people are likely judging multiple traits and states in one situation and make inferences across domains and contexts, for instance with respect to their friends, partners, or work colleagues. However, to our knowledge only two existing tests measure these additional content domains. The IPT (Costanzo & Archer, 1989) measures judgments related to social attributes using short videos with verbal content (spoken in English). As an example, participants view a clip of a man and a woman talking and are asked to infer whether they have been in a romantic relationship for three years or for 10 years. Situational affect is measured by the Profile of Nonverbal Sensitivity (PONS, Rosenthal et al., 1979; and MiniPONS short version, Bänziger, Scherer, Hall & Rosenthal, 2011). In this test, participants view brief clips and pictures of one encoder (a young woman) and have to decide which of two states she is expressing (e.g., "scolding a child" or "ordering food at a restaurant"). Verbal content is masked by filtering the woman's speech. While the IPT and the PONS/MiniPONS shed light on the more ignored but not less relevant domains of IPA, their stimuli are now very dated. For the domain of personality judgment, there is no standard test available at all. This is an important gap considering the burgeoning research on personality judgment in general (for a review, see Back & Nestler, 2016). Although some researchers believed that the "good judge of personality" does not exist (for a review, see Rogers & Biesanz, 2019), various studies have now provided evidence for the existence of individual differences in accuracy (e.g., Christiansen et al., 2005; Letzring, 2008; Colman et al., 2017; Human & Mendes, 2018). The WIPS test therefore includes items measuring accurate judgments of others' personality traits as well as items targeting other IPA domains such as judgments of social characteristics (e.g., status; see Table 1 for a full list).

Third, no IPA test to date has focused on situations in the general professional domain (i.e., beyond medical encounters) as a context for judging others' intentions, behavioral outcomes, personality, status, affect etc., although much of the IPA research has been conducted to examine professional success (e.g., Elfenbein, Foo, et al., 2007; Elfenbein, Polzer, et al., 2007). Therefore, the new WIPS test assesses interactions that typically occur in the workplace. Finally, in order make the WIPS useful for researchers and practitioners in different countries, the test was designed to be language-independent; that is, it relies only on the interpretation of nonverbal facial, vocal, and bodily cues.

1.2. Goals of the present research

The present research attempted to create a standard test to measure broad IPA as a unidimensional skill in the professional domain, consisting of items in which participants are asked to judge various traits (e.g., personality) and states (e.g., behavioral intentions) of targets shown in short video clips. This test (Workplace Interpersonal Perception Skill (WIPS) test) intends to complement the battery of tests available to both IPA researchers as well as practitioners who wish to measure job applicants', clients', patients', students', and other participants' accuracy in judging others.

It is important to point out that, as a standard test with a fixed set of stimuli or items, the WIPS is limited to measuring IPA as a passive, observer-type skill, which may not be the same skill that people display in real-life face-to-face interactions. The alternative would be to use a (dyadic) interaction paradigm in which participants engage in a conversation with someone else and are asked to judge their interaction partner's personality, thoughts and feelings, or other traits or states

Table 1
Item Types in the WIPS.

Item type	Characteristics of video segments	Criteria for correct responses	Example item from final test
Behavioral intentions	Segments that preceded a decision made by one of the actors (e.g., segments prior to accepting or rejecting an offer)	Decision made after the segment as defined by verbal content or behavior (e.g., saying yes or no)	You will see a team member who is in charge of preparing the promotional material for an upcoming event. The intern (not in the picture) offers to do the layout. What will the man in the picture respond? A: He will accept the offer. /// B: He will decline the offer.
Personality traits (Big five, dominance)	Actors with maximally different self-reported traits (e.g., one actor with very high extraversion and one actor with very low extraversion) engaging in interactions without specific behavioral instructions (e.g., to be more assertive) were identified. From these interactions, brief sections containing nonverbal cues relevant to the target trait were selected	Value on self-reported personality trait in question based on the BFI and PRF (dominance)	You will see two negotiators. Which of the two has a more extraverted (outgoing) personality? A: The person on the left. /// B: The person on the right
Status	Sections in which different roles (e.g., team leader, intern, candidate etc.) became apparent, such as at the beginning or end of a team meeting or negotiation	Status/ role assigned to actors	You will see 6 people enter the room for a team meeting. Who is the team leader? – Response options are 6 photos; one of each person
Interpersonal attitudes	Interactions in which actors had received specific instructions to be more or less assertive, yielding/ helpful/ cooperative, or motivated in their respective role were reviewed for segments revealing nonverbal cues for these behaviors	Instructions given to the actor, post-interaction evaluation of self and other (only in helpdesk and negotiation interactions)	Who of the two persons in the following video is less motivated to take part in the preparations for the upcoming conference? — A: The man on the left. /// B: The woman on the right.
Behavioral outcomes	Sections after a decision was made or an outcome was reached; sections related to final negotiation results	The objective outcome as defined by points (in the negotiation) or verbal content	You will see the same person in two different negotiations, signing a contract. In which negotiation did the person negotiate the better deal for herself? — A: In the first one. /// B: In the second one.
Thoughts and feelings	Sections in which actors were (verbally and nonverbally) conveying their thoughts and/ or feelings on a given topic to another person. Sound was muted in these items and it was ensured that the response could not be read from the actor's lips	Verbal content right before, during, or after the segment	You will see brief sequence of a man reacting to something that the team leader just did. What happened? — A: The man in the video is overwhelmed with the number of tasks he just got from the team leader. /// B: The man in the video is disappointed because the team leader chose someone else to give a presentation to a client.

afterwards (e.g., Ickes & Hodges, 2013). Participants' IPA would then be calculated, for instance, as the correspondence of these judgments with the partner's self-rated traits or states. However, such measures of IPA are difficult to implement in applied contexts and are confounded with variables such as the interaction partner's expressiveness (e.g., Back & Nestler, 2016).

1.3. Overview of the present studies and hypotheses

The first three studies were conducted in three languages (English, French, German) to successively test, extend, and refine a pool of test items that would form an internally consistent and unidimensional instrument. These will be referred to as preliminary studies as they only contained subsamples of the final item set. In Studies 4 to 7, the nomological network of the final version of the WIPS was examined in order to assess its construct and external validity. Based on the available literature, we expected the WIPS to show the following associations with other variables and measures:

With respect to demographic variables, we expected a small advantage of women over men in terms of performance, in line with robust findings in emotion recognition (e.g., *meta*-analyses by Hall, 1978, and Thompson & Voyer, 2014). There is also accumulating evidence for the same gender difference in personality judgment accuracy (e.g., Letzring, 2008; Vogt & Colvin, 2003; Jaksic & Schlegel, 2020). In addition, we expected the WIPS to be uncorrelated with age. Although studies in emotion recognition often found a decline in performance with increasing age (e.g., Ruffman, Henry, Livingstone, & Phillips, 2008), such studies were usually based on tests in which emotional expressions were expressed without social context and in only one modality (usually the face). Several researchers have suggested that when rich scenarios with context and information from multiple modalities are presented (as is the case in the WIPS), the age difference may disappear (e.g., Isaacowitz & Stanley, 2011).

We also hypothesized that the WIPS should positively correlate with other performance-based tests in the IPA domain and the domain of emotional competence (specifically, emotion understanding, e.g., Scherer, 2007). If the WIPS indeed measures a broad IPA skill that taps various content domains, performance should be positively correlated with tests or tasks measuring several of these domains. Previous studies typically found medium to large positive effects for associations of standard emotion recognition tests with other standard tests from the same domain and medium to large positive correlations with emotion knowledge (e.g., Schlegel & Scherer, 2016; Schlegel et al., 2017; Schlegel et al., 2019; Schlegel & Scherer, 2018). With respect to general intelligence, we expected a small positive correlation based on a previous *meta*-analysis (Schlegel et al., 2020).

With respect to self-reported personality traits (e.g., Big Five personality traits, the perceived quality of one's social relationships, perceived emotional sensitivity, and emotional clarity), we expected very small associations with the WIPS, as in previous *meta*-analyses (Davis & Kraus, 1997; Hall et al., 2009), all of these constructs showed only small associations (generally, $r < 0.10$) with IPA. One reason might be that people are not very accurate when evaluating their own abilities (Davis & Kraus, 1997). Further, as an ability, high IPA might not in itself be strongly related to behavioral dispositions (Davis & Kraus, 1997).

Finally, we expected the WIPS to predict students' membership and leadership in university student groups such as volunteer groups, music groups, academic interest groups, or cultural groups. We expected students with higher WIPS scores to be more likely to be members in such groups because they may experience more pleasant and successful interactions with others as a result of more accurate judgments of other members' intentions, mood, needs, or traits (see further details in Study 5).

2. Studies 1–4: WIPS test development

4 studies were conducted to successively test and adapt a large pool of items; starting with 25 items in Study 1 and resulting in a final version with 41 items in Study 4. Items fulfilling the selection criteria (see below) were retained and administered again in the following study,

along with a set of newly created items and/or modified versions of previously excluded items. The goal of this stepwise item selection procedure was to select enough items to reach an acceptable internal consistency (McDonald's omega) of 0.70 while keeping the duration of the test short enough for administration in varied and applied contexts (about 20 min). Due to this restriction and in line with the idea of one general IPA skill, we did not create separate subtests for each of the 6 content domains (see Table 1), but aimed for a unidimensional, internally consistent test in which we did not distinguish between domains.

2.1. Method

2.1.1. Material recording

For the creation of the WIPS stimulus material, 30 francophone adults (15 females, aged between 21 and 63, $M = 30.15$, $SD = 13.16$) participated in several role-playing interactions under the direction of a professional acting coach and instructor. All actors but one were previous students in a drama class for business given by the instructor at the University of Lausanne, and had diverse educational and professional backgrounds. Actors were fully briefed about the goals of the research and signed informed consent in accordance with the Department's ethics committee and with the Swiss Federal Act on Data Protection. All agreed to make their audio–video recordings and questionnaire data available for the purpose of creating the test. Through the questionnaire we collected basic demographic information and assessed self-reported personality using the French version of the Big Five Inventory (Plaisant et al., 2010) and the Dominance subscale of the Personality Research Form (Jackson, 1984). The values on these measures later served as criteria for determining the correct answers in items regarding personality. Recording sessions lasted over three days. The raw video material comprised 55 role-playing interactions and over seven hours of footage. Each interaction was recorded from multiple camera angles (capturing the whole scene, each actor's full body and face), which actors were allowed to review during the three months following recording.

We recorded three types of scenarios representing common workplace situations (see supplemental material S1). Each original scenario was only broadly scripted so that actors were free to express themselves according to their natural preferences and tendencies (intended to allow expression of personality traits, thoughts and feelings). All actors played in several scenarios and in different roles in order to have a maximum number of possible dyads or group configurations. While in most scenarios, actors were behaving naturally within their broadly defined role (e.g., as a client), occasionally their behavior was manipulated through additional instructions as described below.

Actors separately familiarized themselves with the scripts during 10–20 min prior to recording and were additionally coached by the acting director and the two first authors (see supplemental material S1 for detailed instructions and scripts).

The first scenario describes a recruiter and an applicant negotiating the terms of a work contract consisting of four negotiation topics (salary, start date, working location and reimbursement of moving expenses). This scenario was based on negotiation exercises used in many empirical studies (e.g., Elfenbein et al., 2007; Schlegel et al., 2018). Recruiters and applicants each received a different schedule of priorities indicating 1) which negotiation topics were important for them and which negotiation topics were not, and 2) which of the five given options they preferred for each negotiation topic. The recruiter was not aware of the applicant's preferences, and vice versa. One negotiation topic was distributive, meaning that both parties had opposite preferences because the preferred option of one party was exactly opposite to the preferred option of the other party. One negotiation topic was compatible, meaning that both parties shared the same option preferences. The other two negotiation topics offered integrative potential through tradeoffs (logrolling) because one topic was more important to one party, and the other topic was more important to the other party. We instructed the actors to achieve the options that approximate their personal preference within 10–15 min of recorded discussion where they could come up with arguments in their personal favor. If they reached an agreement, the dyad signed a "contract" indicating the options for each negotiation topic they agreed upon. After the negotiation, the actors completed a brief questionnaire assessing to what extent they were satisfied with the outcome (agreement made), and to what extent they evaluated themselves and the other party as cooperative or as competitive.

The second scenario describes a typical helpdesk interaction where a client requests assistance from the helpdesk service representative about a technical problem (regarding printing or card access). After playing the "natural" version (no attitudinal instructions to the role of client or representative), some actors played a second, modified version where one of the two (or both) received an additional instruction (unknown to the dyadic partner) to be more or less motivated (for the role of the representative) or more or less friendly (for the role of the client). After each helpdesk interaction, the actors completed a brief questionnaire assessing to what extent they were satisfied with the outcome (how the problem was treated) and to what extent they evaluated the client as friendly or the service representative as motivated to help.

The third scenario describes a company team meeting of 4 to 6 members discussing the organization of an upcoming event (a public exhibition of a newly developed product). The scenario scripted different job roles (team leader, project coordinator, lab manager, designer, administrative assistant, intern) and tasks that were to be discussed during the 20-minute meeting. The actors were free to behave and construct their role around the basic description provided in the scenario. The description stipulated different levels of motivation, seniority, and member relationships (e.g., the administrative assistant is on good terms with the lab manager) in order to have a group dynamic common to existing teams. A modified version of the scenario was additionally enacted, instructing the team leader to be more or less assertive.

2.1.2. Item creation

Authors 1 and 2 independently reviewed the video material (including the different camera perspectives) for brief segments that appeared suitable for creating items (i.e., questions) pertaining to behavioral intentions, personality traits, status, interpersonal attitudes (competitiveness/cooperativeness and motivation), behavioral outcomes, and thoughts and feelings. The characteristics of these segments along with sample items are described in Table 1. We combined multiple sources of information for defining the criterion for each item, as detailed in Table 1. First, the actors' self-report information served as criteria for personality judgment in video segments depicting interactions without any attitudinal instructions ("natural" version). Second, behavioral manipulations through experimenter instructions served as criteria for the judgment of interpersonal attitude (for example, to act more or less motivated) and for the judgment of status. Third, the actors' actual decisions, verbalized thoughts and feelings, and interpersonal attitudes conveyed during or after the recorded interaction served as criteria for the judgment of behavioral intentions, interpersonal attitudes (corroborating the experimenter instructions), behavioral outcomes, and thoughts and feelings.

Identified segments varied in duration between approximately 3 s and 45 s. For each identified segment, an item was written consisting of a question (e.g., In the following video, you will see 6 people enter the room for a team meeting. Who is the team leader?) and between two and six response options (e.g., choose the team leader by clicking on one of the six pictures shown below). For some items, the video consisted of multiple short segments (e.g., You will see the same person in two different negotiations signing a contract. In which negotiation did the person negotiate the better deal for herself?). For each item, one response option was correct (based on the criteria given in Table 1) and was awarded one point when selected; all other response options were awarded zero points.

Table 2
Pretest Study Sample Characteristics, Administered and Retained Items, Descriptive Statistics, and Reliability Indices based on Retained Items.

	Study 1	Study 2	Study 3	Study 4
Sample (EN = English-speaking, FR = French-speaking, GE = German-speaking)	EN (United States) MTurkers: $N = 65$, age $M = 37.4$, $SD = 11.3$ (male = 59%) FR undergraduate business school students: $N = 76$ age $M = 21.3$, $SD = 3.3$ (male = 53%) GE undergraduate psychology students: $N = 108$, age $M = 22.3$, $SD = 2–9$ (male = 19%)	EN (United States) undergraduate students: $N = 83$ age $M = 20.7$, $SD = 4.6$ (male = 30%) FR undergraduate business school students: $N = 114$ age $M = 22.9$, $SD = 5.2$ (male 38%) GE Prolific and Psytests panel: $N = 85$ age $M = 29.4$, $SD = 9.3$ (male = 53%)	EN Prolific: $N = 104$ age $M = 34.3$, $SD = 12.5$ (male = 52%) FR Prolific: $N = 95$ age $M = 28.4$, $SD = 8.3$ (male = 52%) GE Prolific: $N = 99$ age $M = 30.7$, $SD = 9.4$ (male = 50%)	EN Prolific[a]: $N = 98$ age $M = 30.3$, $SD = 10.7$ (male = 28%) FR Prolific[a]: $N = 100$ age $M = 30.6$, $SD = 11.3$ (male = 57%) GE Prolific[a]: $N = 97$ age $M = 29.7$, $SD = 9.8$ (male = 50%)
Number of items administered	25	32 (8 from Study 1)	45 (14 from Study 2)	54 (27 from Study 3)
Number of items retained	8	14	27	41
Mean and Standard Deviation of WIPS score based on retained items	EN: $M = 0.83$, $SD = 0.19$ FR: $M = 0.76$, $SD = 0.20$ GE: $M = 0.80$, $SD = 0.16$	EN: $M = 0.80$, $SD = 0.13$ FR: $M = 0.82$, $SD = 0.12$ GE: $M = 0.79$, $SD = 0.16$	EN: $M = 0.79$, $SD = 0.12$ FR: $M = 0.81$, $SD = 0.11$ GE: $M = 0.78$, $SD = 0.12$	EN: $M = 0.79$, $SD = 0.11$ FR: $M = 0.81$, $SD = 0.09$ GE: $M = 0.79$, $SD = 0.11$
Cronbach's α	0.43	0.45	0.62	0.72
McDonald's ω_t	0.60	0.49	0.65	0.73
Guttman's λ6	0.42	0.46	0.66	0.76

Note. [a] Prolific workers that had participated in a prior WIPS study were not enrolled in the study.

Video segments were muted if the verbal content could have helped identifying the correct response to a question. The possibility that lip reading could help identifying the correct answer (at least for francophone speakers) was excluded through appropriate framing of the question (not containing literal speech) or a camera angle not showing the speaker. This way, we ensured that segments contained only nonverbal cues to the correct answers, i.e., facial expressions, gestures, posture, and prosody (when the video was not muted). We ensured that all items were unrelated, i.e., did not reveal information that could be used to infer the correct response in another item. For instance, items that probe behavioral intentions (e.g., accept a job offer or not) preceded items of the same target person that probe behavioral outcomes (e.g., evaluation of a negotiation outcome upon contract signature). Similarly, items that revealed a person as a team leader come after items that required the test-taker to assess the role of that same person. Furthermore, we explicitly and repeatedly instructed participants to judge each item independently because the same actor could appear in multiple roles and situations. When playing different roles, actors were seated differently and changed clothing items such as their blazer or shirt. Across the 41 items of the final test, the same actor appeared six times at most.

2.1.3. Participants

Native speakers of English, French, and/or German were recruited from various sources including undergraduate student subject pools from different universities in Switzerland and the US and online panels (Prolific, Amazon Mechanical Turk, Psytests – a German panel of volunteers who participate in psychological research). Final sample sizes for each language, recruitment methods, mean age, and gender distribution are reported in Table 2.

All participant data were collected in a manner consistent with ethical standards for the treatment of human subjects, as required by the respective university hosting the study. Mechanical Turk, Prolific, and business school student participants received payment; Psytests panel received personal feedback on their performance on the WIPS; all other student participants received course credit. Business school participants received a bonus payment if they were among the 10 best performers on the WIPS test or on another performance-based test (GERT-S).

2.1.4. Procedure

In all studies, WIPS items were presented in three blocks, one for each interaction context (negotiation, help desk, team meetings) using the survey software Qualtrics. For each item, participants first read the question and the response options, then watch the respective video, and finally see the question again and choose one response option. This was done to provide participants with the context of each video and to guide their attention to its relevant aspects. The order of the blocks and the order of items within each block was randomized. Participants were told that the same actor could appear in different roles and situations and that they were to judge each item independently. The additional instruments in each study (see below) were completed after the WIPS items. All studies were completed online, with the exception of the German-speaking sample in Study 1 which was tested in a laboratory in groups of up to four participants.

In all studies, participants who reported difficulties in playing the videos were excluded from the analyses. In the Mechanical Turk sample in Study 1, five attention check questions were embedded in the survey (e.g., What was the topic of the last item?); participants failing on more than two questions were excluded ($N = 6$). In the other samples of Studies 1 and 2, scores on another performance-based test, the GERT-S (see below), were checked for very low values around guessing level (<15% correct) in order to exclude potentially inattentive participants. No participants had to be excluded based on this criterion. Similarly, in Studies 3 and 4, WIPS scores were checked for outliers (<30% correct); all participants had scores above this value.

2.1.5. Measures

Additional instruments administered along with the WIPS items in each study are shown in Table 3. In Study 2, students of the English- and French-speaking samples also indicated their grade point average of the previous year.

2.1.6. Analysis

Item selection criteria. In all studies, suitable WIPS items were selected based on the following criteria (regardless of the item type): First, suitable items are items that are likely to be able to discriminate individuals but also contain enough information to be answered correctly. The first criterion was thus that items needed to be solved better than chance (e.g., for items with two response options: accuracy above 50%) but by less than 95% of the sample. Items that were answered correctly by more than 95% of participants were too easy and would fail to discriminate among individuals of different ability levels. Second, suitable items needed to show similar difficulty (i.e., mean percentage of correct responses) across the three languages (English, French, and German) to ensure that verbal content (for items with sound) or cultural differences did not affect performance. ANOVAs were conducted for each item with language as a factor; the resulting *p*-values were multiplied by the number of items tested to adjust for the large number of ANOVAs. Items with significant language differences ($p < .05$) were excluded. Third, suitable items needed to show an item-total correlation (i.e., item discrimination) of above 0.10 across the three languages. This was necessary to ensure that items were measuring a similar skill to the other items. All items not matching these criteria were excluded from the item pool. Some items were later modified (e.g., muted, response options changed, video shortened) and tested again in the following study.

Reliability analysis. The retained items in each study were used to compute reliability indices including McDonald's omega, Guttman's lambda 6, and Cronbach's alpha (using the *omega* function in the *psych* package in R; Revelle & Zinbarg, 2009). McDonald's omega total (ω_t) is the proportion of test variance explained by all factors retained in a factor analysis with oblique rotation. Guttman's λ6 is a measure of item communality based on the amount of variance in each item that can be explained by a linear regression of all other items. Both indices were proposed as alternatives to Cronbach's alpha which tends to underestimate reliability, especially with binary items (e.g., Revelle & Zinbarg, 2009). The test development process was considered finalized (i.e., no new items were added) once all reliability indices reached a value above 0.70. This cut-off was chosen as it has been commonly recommended in the literature (e.g., based on Nunally, 1978). At the same time, it is well above the average alpha of 0.48 found for IPA tests in the *meta*-analysis of Schlegel and colleagues (2017).

Power analysis. Associations with other instruments (see Table 3) were computed using Pearson correlations to evaluate construct validity. A power analysis conducted with G*Power (Faul et al., 2007) revealed a necessary *N* of 109 to detect medium effect sizes ($r = 0.30$) for a power of 0.90 and an alpha level of 0.05 for a two-tailed test. For small effects ($r = 0.20$), the same analysis yielded a sample size of 255. As described in the last section of the introduction ("Overview of the present studies and hypotheses"), based on previous studies we expected medium positive effects for associations with emotion-focused performance-based tests, and a small effect for the association with gender favoring women. For age and self-reported personality traits, no significant effects were expected, and no minimal required sample size was calculated for these associations. The total samples of Studies 1 to 4 were large enough to detect small effects for gender. Associations with emotion-focused tests (GERT-S and GEMOK Blends) were only assessed for some languages, but these subsamples were large enough to detect medium effects, with the smallest relevant *N* being 179. Self-reported grade point average was examined as a correlate in two subsamples (Study 2 English, $N = 76$; Study 2 French, $N = 105$), which were smaller than necessary to detect small or medium effects. These numbers were nevertheless deemed appropriate given that grade point average was not one of the central variables for construct validation and was included in an exploratory fashion.

2.2. Results

Table 2 shows how many items were administered and retained using the selection criteria in each study, as well as the reliability indices and descriptive statistics for the respective test version containing all *retained* items up until that point (for example, for Study 3, the reported statistics are based on 27 items, not 45). The set of 41 retained items in Study 4 fulfilled the a priori defined goal of all three reliability indices being above 0.70 and taking about 20 min to complete. This version was therefore considered the final WIPS version. The distribution of the final 41 items in terms of item content was as follows: Behavioral intention – 11 items, behavioral outcome – 7 items, interpersonal attitude – 5 items,

Table 3
Overview of Additional Measures.

Study	Measure	Description
Study 1–2	Geneva Emotion Recognition Test - short version (GERT-S, Schlegel & Scherer, 2016)	Video-based emotion recognition test
	Geneva Emotion Knowledge test (GEMOK, Schlegel & Scherer, 2018)	Emotion knowledge test based on fictional written scenarios of emotional situations
	Social Skills Inventory - Emotional sensitivity subscale (SSI-ES, Riggio, 1986)	Self-report questionnaire on social communication skills
Study 4	brief Profile of Nonverbal Sensitivity (MiniPONS, Bänziger et al., 2011)	Video-based test of situational affect
	Ten-Item Personality Inventory (TIPI, Gosling et al., 2003)	Big Five personality self-report questionnaire
Study 5 - GE	Fluid intelligence test (mini-q, Baudson & Preckel, 2016)	Quick judgment test solving simple logical problems
	Geneva Emotional Competence Test -Emotion Understanding subtest (GECo, Schlegel & Mortillaro, 2019)	Emotion knowledge subtest based on fictional written scenarios of emotional situations
	Trait Meta Mood Scale (TMMS, Otto et al., 2001)	Self-report questionnaire of attention to feelings, clarity of feelings, and mood repair
	Ryff Wellbeing - Positive Relations with Others subscale (Ryff & Keyes, 1995)	Self-report questionnaire of social relationship quality
Study 5 - EN	Emotional Sensitivity (SSI, Riggio, 1986)	Self-report subscale of the Social Skills Inventory
	Diagnostic Analysis of Nonverbal Accuracy – Adult Faces (DANVA-2AF, Nowicki & Duke, 1994)	Emotion recognition test based on still pictures of faces
Study 6	Multifactor Leadership Questionnaire - version 6S (MLQ, Bass & Avolio, 1992)	Self-report questionnaire assessing transformational, transactional, and laissez-faire leadership style
	Personality judgment accuracy test (Jaksic & Schlegel, 2020)	Ad-hoc rating task using video material of negotiation interactions

Note: EN = English-speaking sample, GE = German-speaking sample.

personality – 7 items, status – 3 items, thoughts and feelings – 8 items. Twenty of the final items originated from the team meeting scenario, 6 from the helpdesk scenario, and 15 from the negotiation scenario. The correlations of the preliminary WIPS versions with the GERT-S, the GEMOK Blends, the SSI, and age and gender in Studies 1–3 are presented in the supplemental material (S2). Overall, women scored significantly higher on the WIPS than men. GERT-S, GEMOK and age showed either no significant correlation or a significant positive correlation with the WIPS. WIPS scores were not significantly correlated with the SSI and self-reported grades. Correlations of the final WIPS (Study 4) with other measures and demographics are presented in Table 4. Again, a gender difference favoring women, but no age effect was found. The WIPS was largely unrelated to the Big Five but was strongly positively correlated with the MiniPONS, providing evidence for construct validity.

3. Study 5: unidimensionality and work-related outcomes

The first goal of Study 5 was to evaluate the internal consistency and factor structure of the final WIPS test, as well as its conformity with the Rasch model. The Rasch model is the simplest model in Item Response Theory (IRT). It assumes that the probability of solving an item only depends on a person's location on a latent ability dimension θ and on the difficulty of an item, both of which are displayed on the same latent dimension. Item difficulty determines the location on θ where this item has the highest measurement precision; that is, an easy item will measure most precisely at low ability levels. Because both item parameters and participants' ability estimates can be displayed on the same θ dimension, their distributions can be compared, allowing to evaluate the difficulty of the overall test as well its measurement precision in relation to the particular sample that was studied. If the Rasch model fits the data, the sum or average score of all test items is considered a sufficient estimate of a person's ability, because the model assumes that all items discriminate equally well and can therefore be equally weighted. In a second step, we assessed construct validity as well as predictive validity.

The second goal of Study 5 was to examine construct and predictive validity. In terms of predictive validity, we expected the WIPS to predict students' membership and leadership in university student groups such as volunteer groups, music groups, academic interest groups, or cultural groups. Perceptive students may have a stronger and more rewarding experience in terms of group cohesion and self-other overlap (Parkinson et al., 2005), and may therefore be more likely to join such groups. They might also be more likely to be accepted for membership in these groups because they are perceived as more likeable, empathic etc. by others (Hall et al., 2009). In addition, for some groups, more specific mechanisms might explain higher IPA among members as compared to non-members. For instance, musical training may provide the ability to hear and decode prosody and expressive meaning in voices, thus fostering IPA (Thompson et al., 2004; Hall et al., 2009). Members in music-oriented groups might thus have higher IPA than non-members. Higher IPA has also been linked to more prosocial behavior (e.g., Kaltwasser et al., 2017), suggesting that such individuals may be more likely members of volunteer groups. Finally, through the positive link between IPA and intelligence (Schlegel et al., 2020), high IPA students may be more likely to join academic interest groups or honors societies.

As for leadership, we expected leaders in student groups to outperform non-leader members on the WIPS, because successful interactions with others (team members, other leaders, sponsors etc.) make up major part of their tasks and higher IPA students may be more likely to achieve these (Schmid Mast & Latu, 2016). Students who make it into leadership positions within their group are likely to be those who are very well attuned to other people's needs, desires, and feelings, which is also a component of transformational leadership style (Bass & Riggio, 2006). In line with this hypothesis, two studies, one in a factory setting and one using a laboratory task, found that managers or leaders were more interpersonally accurate than subordinates (Zhong et al., 2013; Schmid Mast & Darioly, 2014).

Table 4
Correlations of WIPS Scores with Other Measures in Study 4 (N = 295).

Measure	*N*	*Cronbach's α*	*M (SD)*	*r (95% CI)*
age	295			0.06 (−0.05; 0.18)
gender	295			0.14* (0.03; 0.25)
MiniPONS	295	0.66	0.75 (0.08)	0.52*** (0.43; 0.60)
TIPI Extraversion	295	0.81	3.58 (1.47)	−0.04 (−0.15; 0.07)
TIPI Agreeableness	295	0.24	4.76 (1.12)	0.18** (0.07; 0.29)
TIPI Conscientiousness	295	0.67	5.13 (1.25)	0.03 (−0.08; 0.14)
TIPI Openness	295	0.44	5.05 (1.17)	0.07 (−0.04; 0.18)
TIPI Emotional Stability	295	0.65	4.37 (1.39)	−0.03 (−0.14; 0.08)

Note: MiniPONS = brief Profile of Nonverbal Sensitivity (scale: 0 to 1), TIPI = Ten-Item Personality Inventory (scale: 1 to 7). * $p < .05$, ** $p < .01$, *** $p < .001$.

3.1. Method

3.1.1. Participants

Sample characteristics are described in Table 5. The participant recruitment approach was the same as described above for Studies 1–4. Seven participants from the US undergraduate sample were excluded because they failed to correctly respond on either of two attention check questions which asked participants to select a specific option in a multiple-choice question.

3.1.2. Measures

Participants completed the WIPS and additional instruments (see Table 4). The US undergraduate sample (610 out of the 646 participants)

Table 5
Study Sample Characteristics, Descriptive Statistics, and Reliability Indices of the Studies Conducted with the Final 41 Item WIPS Test.

	Study 5	Study 6	Study 7
Sample (EN = English-speaking, FR = French-speaking, GE = German-speaking)	EN: undergraduate students (United States), $N = 646$, and Prolific[a], $N = 53$ age $M = 22.6$, $SD = 8.8$ (male = 38%) GE Prolific[a]: $N = 143$ age $M = 30.6$, $SD = 11.1$ (male = 63%)	EN (United States) managers recruited through Prolific[a]: $N = 83$, age $M = 41.2$, $SD = 9.5$ (male = 60%)	EN (United States) undergraduate students: Time 1 $N = 116$ age $M = 19.20$, $SD = 1.93$ (male = 38%) Time 2 $N = 76$ age $M = 19.37$, $SD = 2.23$ (male = 46%)
Mean and Standard Deviation of WIPS score	EN: $M = 0.75$, $SD = 0.12$ GE: $M = 0.80$, $SD = 0.08$	$M = 0.78$, $SD = 0.10$	Time 1: $M = 0.79$, $SD = 0.10$ Time 2: $M = 0.77$, $SD = 0.12$
Cronbach's α	0.74	0.67	T1: 0.68, T2: 0.74
McDonald's ω_t	0.75	0.71	T1: 0.71, T2: 0.80
Guttman's λ6	0.76	0.84	T1: 0.81, T2: 0.91

Note. [a] Prolific workers that had participated in a prior WIPS study were not enrolled in the study. Distributions of the total WIPS scores are displayed in supplemental material S3.

was also asked if they were or had ever been a member of the following groups: sports teams, Greek life (i.e., sororities and fraternities), student government, religious, cultural, academic interest, volunteer, or music-oriented clubs, and honor societies, which was then labeled as a dichotomous variable (member vs. non-member). In addition, participants indicated the highest position they held in each group on a multiple-choice question (group member, executive board, president, or other positions with a fill-in option), which was also dichotomized to leader (president and executive board) vs. non-leader (all other positions).

In contrast to the first 3 studies (random item order within the three blocks negotiation, team meeting, and helpdesk), the items in the final WIPS (used in Studies 5, 6, and 7) were presented in a fixed order within each block. The order was created in a way as to avoid possible carry-over effects between items. For instance, an item asking who the team leader out of six people in a video clip is was presented before an item containing video material from the same interaction that might contain additional clues about who the team leader is.

3.1.3. Power analysis

As described in Studies 1–4, the necessary sample size was 109 to detect medium effects and 255 to detect small effects, which was well exceeded by the *N* available for gender ($N = 800$, see Table 6). The subsamples that examined associations with emotion-focused tests were large enough to detect the expected medium effects (see Table 6). For intelligence, only a small correlation was expected (see last section of the introduction), and the subsample examining this association was smaller than necessary to detect this effect with an *N* of 143. For the expected differences in student group membership and leadership, t-tests were calculated. Given that we are unaware of previous studies making similar comparisons, we calculated the sample size required for small effects (Cohen's $d = 0.20$) and small-to-medium effects ($d = 0.30$) with a power of 0.90 at an alpha level of 0.05, which were $N = 858$ and $N = 382$, respectively. With an of $N = 610$, the present sample was not big enough to detect small effects, but well exceeded the number needed for small-to-medium effects.

3.2. Results

All three reliability indices were above 0.70 (see Table 5), indicating good reliability of the WIPS in a large sample. In order to test whether the final WIPS is a unidimensional test, we ran a one-factor confirmatory factor analysis (CFA) as well as a set of exploratory factor analyses (EFAs) with two to six factors over the final set of 41 items in Mplus (Muthén & Muthén, 2011) on the whole sample ($N = 842$). Model fit was evaluated using the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). The CFA with all 41 items loading on one common factor showed good model fit ($\chi 2 = 881.670$, $df = 779$, $p = .006$, $CFI = 0.955$, $TLI = 0.952$, $RMSEA = 0.013$). The EFAs for two to six factors did not produce interpretable patterns based on content domains, with many items loading similarly on multiple factors. The excel sheet in the supplemental materials (S4) provides the factor loadings for the CFA and the EFAs.

After testing the unidimensionality assumption, we fitted the Rasch model for binary data using the *eRm* package in R (Mair et al., 2019). Model fit was examined for each item using standardized Infit and Outfit t statistics, with values between −2 and 2 indicating good fit. Inspection of Infit and Outfit indices revealed good fit for 38 out of 41 items; we therefore concluded that overall, the Rasch model appropriately describes the WIPS and the mean score across all items is a sufficient indicator of a person's IPA. The comparison of the person parameter (i.e., ability score) and item parameter (i.e., item difficulty) distributions on the latent dimension (see Person-Item map, supplemental material S5) revealed that overall, most items have the highest measurement precision in the lower ability range and only few items have a high measurement precision in the medium ability range. In other words, the WIPS is an easy test when compared to the overall ability level of the studied sample. This is an issue common to most IPA or emotion recognition tests (see Boone & Schlegel, 2016; Kenny, 2013).

Correlations between the WIPS and the other instruments are shown in Table 6. A gender difference favoring women and a positive correlation with age were found. In addition, the WIPS correlated positively with the GECo Emotion Understanding test and the DANVA-2 Adult Faces test. Correlations with self-report measures were very small and mostly not significant.

Table 6
Correlations of WIPS Scores with Other Measures in Studies 5, 6, and 7.

Study	Subsample	Measure	*N*	*Cron-bach's α*	*M (SD)*	*r (95% CI)*
Study 5	all	age[a]	801			0.12* (0.05; 0.19)
	all	gender[a]	800			0.13* (0.06; 0.20)
	GE	mini-q	143	0.91	0.64 (0.19)	0.11 (−0.06; 0.27)
	GE	GECo Emotional Understanding subtest	143	0.56	0.73 (0.14)	0.25** (0.09; 0.40)
	GE	TMMS clarity of own feelings	143	0.86	3.54 (0.67)	0.01 (−0.15; 0.17)
	GE	TMMS mood repair	143	0.80	3.33 (0.83)	−0.05 (−0.21; 0.12)
	GE	TMMS attention to own feelings	143	0.89	3.54 (0.71)	0.00 (−0.16; 0.16)
	GE	Ryff quality of social relationships	143	0.89	4.60 (1.09)	0.08 (−0.09; 0.24)
	EN	SSI Emotional Sensitivity	610	0.81	5.78 (1.15)	0.08* (0.00; 0.16)
	EN	DANVA-2AF	601	0.60	0.75 (0.12)	0.34* (0.27; 0.41)
Study 6	EN	age	83			0.16 (−0.06; 0.36)
	EN	gender	83			0.17 (−0.05; 0.37)
	EN	MLQ transformational leadership	83	0.91	3.93 (0.60)	−0.11 (−0.32; 0.11)
	EN	MLQ transactional leadership	83	0.69	3.89 (0.57)	−0.08 (−0.29; 0.14)
	EN	MLQ Laissez-faire	83	0.68	3.03 (0.75)	−0.08 (−0.29; 14)
	EN	trait accuracy (personality judgments)	83	0.18	0.14 (0.10)	0.30** (0.09; 0.48)
	EN	profile accuracy (personality judgments)	83	0.33	0.25 (0.09)	0.25* (0.04; 0.44)
	EN	distinctive profile accuracy (personality judgments)	83	0.26	0.08 (0.10)	0.33** (0.12; 0.51)
Study 7	EN	age	116			0.04 (−0.14; 0.22)
(T1)	EN	gender	116			0.26* (0.08; 0.42)

Note: EN = English-speaking, GE = German-speaking. mini-q = fluid intelligence test (scale: 0–1), GECo = Geneva Emotional Competence Test (scale: 0–1), TMMS = Trait Meta Mood Scale (scale: 1–5), SSI = Social Skills Inventory Emotional sensitivity subscale (scale: 1–9), DANVA-2AF = Diagnostic Analysis of Nonverbal Accuracy-Adult Faces (scale: 0–1), MLQ = Multifactor Leadership Questionnaire (scale: 1–5). Values on the Ryff scale could range from 1 to 7. Gender was coded 0 = male, 1 = female. Descriptive statistics for age and gender can be found in Table 5. For personality judgments, accuracy scores are correlation coefficients (possible range −1 to 1). [a] for 41 participants, age and gender were not recorded, 1 person reported nonbinary gender orientation. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 7
WIPS Scores as a Function of Student Group Membership and Role.

Group Type	*N*	*M (SD)*	*t*	*p*	*d* (95% CI)
Music-Oriented					
Yes	200	.78 (.10)	5.88	< .001	.45 (.29; .63)
No	410	.73 (.13)			
Volunteer					
Yes	167	.77 (.12)	2.50	.013	.23 (.05; .41)
No	443	.74 (.12)			
Honor Society					
Yes	193	.78 (.11)	4.17	< .001	.37 (.17; .52)
No	417	.73 (.13)			
Academic Interest					
Yes	83	.79 (.11)	3.19	.001	.37 (.14; .61)
No	527	.74 (.12)			
Religious					
Yes	189	.76 (.10)	2.60	.010	.21 (.04;.38)
No	421	.74 (.13)			
Cultural					
Yes	23	.79 (.08)	2.76	.010	.40 (.02; .81)
No	587	.74 (.12)			
Sports					
Yes	543	.75 (.12)	.72	.471	.11 (.15; .36)
No	67	.73 (.14)			
Student Government					
Yes	58	.76 (.11)	1.08	.283	.15 (.12; .42)
No	552	.74 (.12)			
Greek Life					
Yes	126	.71 (.14)	3.09	.002	.35 (.15; .54)
No	484	.76 (.12)			
Outperforming groups[a]					
Leaders	108	.78 (.11)	1.98	.050	.21 (.00; .43)
Members	376	.75 (.12)			
Other groups[b]					
Leaders	305	.75 (.12)	0.838	.402	.07 (.09; .23)
Members	294	.74 (.11)			

Notes. Total $N = 610$. In leader vs. non-leader comparisons, total *N*'s are lower as only participants who indicated membership in at least one of the respective groups were included.

[a] Groups in which members scored higher on the WIPS than non-members are: music, volunteer, honor society, academic, religious, and cultural groups.

[b] Groups in which members did not score higher on the WIPS than non-members are: sports, student government, Greek life.

In order to assess group differences of members vs. non-members and leaders vs. non-leaders of student groups, *t*-test were conducted. In line with our assumptions, US undergraduate members of music groups, academic interest groups, cultural groups, volunteer groups, religious groups, and honors societies scored higher on the WIPS than non-members in each of these groups (see Table 7). Membership in student government and sports clubs was unrelated to WIPS scores; it may be that membership in these clubs is more strongly driven by interests that are unrelated to interpersonal interactions (e.g., achieving high athletic performance). Somewhat unexpectedly, members in sororities and fraternities scored significantly lower on the WIPS than non-members. It might be that interpersonally accurate students prefer the first set of groups (those with a specific focus), and due to the time spent in these groups, less often join Greek life organizations. Overall, the total number of groups in which students were members positively correlated with WIPS scores; $r = 0.21$, $p < .001$. With respect to leadership in student groups, we found that leaders in groups that outperformed on the WIPS (i.e., people that indicated being a leader in one or more of the following groups: music, academic interest, cultural, volunteer, religious, and honors society) scored even higher on the WIPS than members of those groups without a leading role, with a small effect size. Leaders in the other three groups (student government, sports, and Greek life) did not have higher WIPS scores than non-leader members (see Table 7).

4. Study 6: WIPS & personality judgment accuracy

4.1. Method

4.1.1. Participants

83 participants with a manager role were recruited via Prolific (see Table 5 for sample details and supplementary material S6 for additional characteristics). Native English speaking US residents who fulfilled the following additional criteria (as provided as screening variables by the panel) were invited to participate: Employed in an organization (not self-employed), having supervisory responsibilities, having experience in a management position, and currently working full- or part-time. These selection criteria narrowed the attainable sample size below the required *N* of 109 to detect medium effects (see power analysis in Studies 1–4). Among native speakers of French or German, at the time of data collection only few managers fulfilling our selection criteria were available on Prolific. Therefore, this study was conducted only with English speakers.

4.1.2. Measures

All participants completed the final 41 item WIPS test, provided information on their leadership experience and organization, and completed the 21-item *Multifactor Leadership Questionnaire* (MLQ version 6S; Bass & Avolio, 1992) as well as a *personality judgment task* developed for this specific study given that no standard measures to assess personality judgment accuracy exist to date. Descriptive statistics on leadership experience and organization size, number of subordinates etc. are provided in the supplemental material (S6). As intended by the initial screening through Prolific, all participants had supervisory responsibility. The MLQ consists of seven subscales with three items each that are rated on a 5-point Likert scale (Idealized Influence, Inspirational motivation, Intellectual stimulation, Individual consideration, Contingent reward, Management-by-exception, and Laissez-faire). The first four scales were summarized to form an index of transformational leadership which is a leadership style enabling organizational change by understanding the organization's culture and showing sensitivity to the needs or others (Bass & Avolio, 1992). Contingent reward and Management-by-exception were summarized to form an index of transactional leadership, which is characterized by working within existing organizational rules and norms. The Laissez-faire subscale was left as such, referring to a style that avoids responsibilities and making decisions (Bass & Avolio, 1992).

In the personality judgment accuracy task, participants were asked to rate personality traits for 16 target individuals, which were 7 male and 9 female undergraduate students from a university in the US (age $M =$ 18.1 years). These targets were shown in a muted 25 s video clip (sitting on a chair; filmed at about a 45-degree angle from their front; whole body except lower legs were visible). The video clip showed the beginning of an employee-recruiter negotiation with an unacquainted other participant of the same gender (also an undergraduate student; not visible in the video), including the greeting and the start of their discussion. Details on the negotiation task can be found in Schlegel et al. (2018). Immediately after the negotiation, targets had rated their competitiveness on four items (5-point Likert scale), "It was important to me to win the negotiation", and "During the exercise, I tried to … (1) be competitive, (2) be persistent/ dogged, (3) be firm" which averaged into one competitiveness score. Targets had also completed the Ten Item Personality Inventory (TIPI; Gosling et al., 2003) and the emotional sensitivity subscale of the Social Skills Inventory (SSI; Riggio, 1986) in a different session of the study. The 16 targets were selected from a larger pool of participants such that for each trait (the Big Five, competitiveness, and emotional sensitivity) high, low, and medium values were represented.

Participants (managers) in the current study were asked to rate each target on the ten items of the TIPI, competitiveness (to what extent is the person competitive, striving to win), and emotional sensitivity (to what

extent is the person empathic, understanding what makes people tick) on a 5-point Likert scale (1 = this trait does not apply to the person at all/ does not describe the person well at all, to 5 = this trait fully applies to the person/ describes the person very well). Participants were instructed to consider how they believed each person acts in real life (i.e., judge a person's personality rather than their current state). From the 12 ratings provided by each participant for each of the 16 targets, three personality judgment accuracy scores were calculated following procedures described in the literature (e.g., Back & Nestler 2016; Hall et al. 2017; Letzring & Funder 2018; but see Biesanz, 2021, for a different approach to estimating accuracy).

Trait accuracy refers to the ability to discriminate *among different targets* on one given trait (e.g., to discern whether one target is more or less competitive than another target). Trait accuracy was calculated as follows: First, for each of the Big Five, the two respective TIPI items were combined into one score in both the targets' self-rating and participants' (managers') ratings. Second, for each of the seven traits (Big Five, competitiveness, and emotional sensitivity), participants' ratings across the 16 targets were correlated with the targets' self-ratings on the trait, resulting in seven trait accuracy scores. These seven scores were averaged to yield one overall trait accuracy score per participant.

Profile accuracy refers to the ability to accurately judge relative levels of different traits *within one target* (e.g., to judge whether a person is more agreeable than competitive). Profile accuracy was calculated by correlating the 12 ratings provided by each participant for one target with this targets' self-ratings on the same 12 items. The correlations across the 16 targets were averaged for each participant to form an overall profile accuracy score.

Finally, *distinctive profile accuracy* was calculated, which statistically removes the average target's profile and the average profile as rated by participants (judges) from profile accuracy scores. This is done to account for the finding that high profile accuracy can be achieved simply by attributing a typical personality profile to all targets without knowing whether or how a target's actual profile differs from the typical profile. Distinctive profile accuracy thus yields a measure of a judge's ability to evaluate targets' unique, distinctive personality profiles. It was calculated as described in Furr (2008) and Jaksic and Schlegel (2020).

It should be noted that competitiveness was rated in terms of a personality trait by participants (see instructions above), but targets' self-ratings of competitiveness referred only to the negotiation, thus rather describing a state. This mismatch between self- and other-ratings may have resulted in lower accuracy scores.

4.2. Results

Reliability of the WIPS was satisfactory with all three indices close to or exceeding 0.70 (see Table 5). However, the reliability of the three personality judgment scores was low (see Table 6) which is in line with the values obtained in other studies as discussed by Rogers, Furr, and Wood (2018). Results showed that managers with higher WIPS scores were more accurate judges of unacquainted others' personality (see Table 6). As predicted, significant positive correlations (medium effect sizes) were found for all three personality judgment accuracy scores including the ability to rank target individuals on specific traits (trait accuracy), the ability to assess targets' personality profiles (profile accuracy), and the ability to discern targets' unique trait constellations as compared to the average person's profile (distinctive profile accuracy). Given the low reliability of the personality judgment scores, these correlations may underestimate the magnitude of the true associations. Correcting for attenuation (i.e., for measurement error due to imperfect reliability of the measures) according to the procedure outlined by Schmidt and Hunter (2014) substantially increased the correlations between the WIPS and trait accuracy ($r = 0.70$), profile accuracy ($r = 0.49$), and distinctive profile accuracy ($r = 0.89$). There were no significant correlations of the WIPS with age, gender, or the self-rated MLQ leadership variables. While the results for the MLQ are not unexpected due to its self-report format, the absence of a significant gender difference may partly be due to the relatively low sample size ($N = 83$), which is lower than the required N of 255 to detect small effects (see power analysis in Studies 1–4). However, due to the narrow selection criteria for study participation, we considered 83 managers a satisfactory sample.

5. Study 7: test-retest reliability

5.1. Method

Sample characteristics are described in Table 5. Participants were invited to complete the WIPS twice. There were approximately 2–4 weeks between the two time-point measures ($M = 19$ days, $SD = 4$). Five participants had to be excluded because their identifier codes on time 1 and time 2 could not be matched. Previous studies reported test–retest correlations of $r = 0.56$ (TAPPA; Hall et al., 2014), $r = 0.64$ (MiniPONS; Bänziger et al., 2011), and $r = 0.78$ (MERT; Bänziger et al., 2009). A power analysis revealed a necessary N of 35 to detect an effect of $r = 0.56$ for a power of 0.90 and an alpha level of 0.05 for a two-tailed test. Although the final sample of $N = 71$ in the present study well exceeded this number, it should be noted that correlations obtained with relatively small samples such as the present can be unstable (Schönbrodt & Perugini, 2013). Participants received course credit for taking part in the study. Age and gender information were collected; no other measures were administered.

5.2. Results

Reliability for each of the two individual samples was satisfactory with all three indices close to or exceeding 0.70 (see Table 5). Regarding test–retest reliability, the correlation between the time 1 and the time 2 sample was $r = 0.68$, 95% CI [0.54, 0.79]. We further estimated the single score intraclass correlation coefficient (ICC) using the R statistical package "irr" (Gamer et al., 2012), based on a Two-Way Mixed-effects Model using absolute agreement (Koo & Li, 2016; Shrout & Fleiss, 1979). This ICC = 0.65, 95% CI [0.49, 0.77]. These results indicate moderate reliability (Koo & Li, 2016). There was no significant group difference between the two time-points (Table 5), $t(70) = -1.98$, $p = .05$, 95% CI [0.04, 0.00]. There was a gender difference favoring women, and no correlation with age (see Table 6).

6. Cross-study results and mini-meta-analysis

Taken together, the WIPS can be considered a unidimensional test that conforms to the Rasch model. Reliability was satisfactory in the four studies with the final item set, with all three indices (McDonald's ωt, Guttman's λ6, and Cronbach's α) generally exceeding 0.70 (see Tables 2 and 5).

When the three languages were assessed separately, reliability indices showed some variation between studies. In French, the final WIPS was only tested in one study (Study 4), with an ωt of 0.60 which was somewhat lower than omegas for German (0.74) and English (0.77) in the same study. Omegas for the German version were 0.74 (Study 4) and 0.60 (Study 5). There was no apparent reason for this discrepancy among the German samples, as both samples were recruited through Prolific and reached a similar mean WIPS score. However, variability in the German Study 5 sample was the lowest ($SD = 0.08$) of all seven separate language samples tested with the full WIPS version, with lower variability leading to lower item-total-correlations and hence, to lower reliability. In English, omegas were consistent across four studies with 0.77 (Study 4), 0.77 (Study 5), 0.71 (Study 6), and 0.71 and 0.80 (Study 7, T1 and T2). Taken together, reliability appears to be acceptable for all three languages, but this should be confirmed with more studies especially for the French and German versions.

In terms of construct validity, with respect to performance-based

measures in the affective domain, as expected, WIPS scores significantly positively correlated with the MiniPONS test and with the DANVA Face subtest in Studies 4 and 5, with medium to large effects. The WIPS also positively correlated with the GECo Emotion Understanding subtest (Study 5) with a small-to-medium effect. These results are supported by similar correlations obtained with other tests from the emotion domain (GERT-S and GEMOK) in preliminary Study 2 (see Table S1). For fluid intelligence, the association found with the WIPS was only small (r = 0.10) and not statistically significant (Study 5), but it was in the expected direction and not too dissimilar from the *meta*-analytic correlation found for the emotion recognition domain of interpersonal accuracy (r = 0.19; Schlegel et al., 2020).

As expected, correlations between the WIPS and self-reported personality traits and other variables were mostly very small and not statistically significant. Only the correlation with agreeableness in Study 4 (r = 0.18, p < .05) and the correlation with emotional sensitivity (SSI; r = 0.08, p < .05) in Study 5 reached significance, but were still small. WIPS scores were unrelated to other affective traits (TMMS), quality of social relations, and leadership style among managers.

Given that participant age and gender were available in all studies, we conducted mini *meta*-analyses according to the procedures outlined by Goh, Hall, and Rosenthal (2016) with fixed and random effects to assess the overall association between WIPS score, age, and gender for Studies 4–7 in which the final WIPS was used. For age, the fixed effects analysis (weighted by sample size) yielded an effect of r = 0.10, and the random effects analysis (unweighted by sample size) yielded an effect of r = 0.10. Stouffer's Z was 1.530 (p = .063). In line with our expectations, no decline in WIPS performance was found for older participants. For gender, the fixed effects analysis yielded an effect of r = 0.15, and the random effects analysis yielded an effect of r = 0.18. Stouffer's Z was 2.53 (p = .006). This effect size is similar to the previously reported *meta*-analytic effect size of r = 0.19 (Thompson & Voyer, 2014) showing a small advantage of females.

In order to assess whether correlations of the WIPS with other measures and variables generalized across the three languages, we recalculated all correlations with age and gender by language for Studies 4–7. The correlations in the seven samples ranged from 0.11 to 0.26 for gender and from −0.06 to 0.16 for age, which can be considered relatively consistent and in line with our overall interpretation as well as the literature (small advantage for women; no association with age). For the TIPI (Study 4), the correlations for the five traits with WIPS in the three language samples ranged from −0.16 to 0.27, with only agreeableness reaching significant values in German and in English consistent with our expectation of low associations between WIPS and the Big Five. The MiniPons was significantly positively correlated with the WIPS in all three languages (Study 4), and samples from all three languages contributed other evidence for construct validity with different tests (e.g., German: correlation with emotion understanding using GECo in Study 5; French: correlation with emotion recognition using GERT-S and emotion understanding using GEMOK in Study 2; English: correlation with emotion recognition using DANVA in Study 5).

7. Discussion

Interpersonal accuracy (IPA) was recently proposed by Hall et al. (2016) as a term to integrate research from largely independent fields that all study accurate inferences of other people's traits and states from (mostly nonverbal) behavior. The purpose of the present research was to contribute to the broad IPA field by developing a new performance-based test that combines various content domains such as accurate judgments of personality, intentions, behavioral outcomes, and group membership within the context of workplace interactions.

7.1. Summary and discussion of main findings

The seven studies presented here showed that it was possible to develop a test that is both internally consistent and taps into different IPA domains, suggesting that IPA could be measured more broadly than what other, specific tests have done. The WIPS is a 20-min test that includes items related to the evaluation of others' behavioral intentions and outcomes, attitudes, thoughts and feelings, status, and personality traits, and covers different nonverbal channels (face, body, voice) while excluding the linguistic channel. It combines different criteria to establish correct responses (Bernieri, 2001), including actual decisions made and behaviors shown, verbalized thoughts and feelings, interpersonal attitudes conveyed by the target person (actor) during or after the recorded interaction, and actors' self-reported personality traits. As such, the WIPS covers more heterogeneous content and criteria than most existing tests in the IPA domain.

Despite this heterogeneity, the test shows acceptable internal consistency (Cronbach's alphas around 0.70) that exceeds the average across IPA tests according to a recent *meta*-analysis (0.48, see Schlegel et al., 2017). Even though the values obtained here are modest when compared with generally recommended cut-offs (e.g., 0.70, Nunally, 1978), the internal consistency obtained here can be seen as satisfactory when considering that the WIPS includes a much more diverse and naturalistic content than standard IPA tests, which are typically restricted to posed emotional expressions of one individual. In addition, Study 7 showed a satisfactory test–retest reliability of 0.68, which is at least as high or higher than for related IPA tests such as the TAPPA (Hall et al., 2014) and the MiniPONS (Bänziger, 2011). However, considering the relatively small sample of Study 7, more studies are needed to replicate this finding.

With respect to the nomological network of the test, our hypotheses were generally confirmed. The WIPS showed medium-to-large associations with two of the central IPA domains as measured by standard emotion recognition tests and a task measuring personality judgment accuracy. A somewhat smaller positive correlation was found with emotion understanding, reflecting that both constructs draw on knowledge of causes, characteristics, and consequences of affect. This in line with research on nonverbal cue knowledge (Rosip & Hall, 2004) and with a recent model of personality trait perception that emphasizes the role of temporary affective cues in the inference of more stable personality traits (State Trait Accuracy Model, STAM; Hall et al., 2017). The observed correlation of the WIPS with intelligence was somewhat lower than the *meta*-analytic correlation found for emotion recognition (r = 0.11 vs. r = 0.19; Schlegel et al., 2020). This could be because the intelligence test used here measured only a narrow facet of intelligence (making quick judgments on simple logical problems), and other facets such as crystallized verbal intelligence may contribute to IPA as well (Matthews et al., 2004; Roberts et al., 2010).

As predicted, the WIPS was largely unrelated to self-reported traits as well as self-rated leadership style, confirming that the WIPS measures an ability rather than a personality trait or interpersonal style (Davis & Kraus, 1997; Hall et al., 2009). Finally, in terms of gender and age, our hypotheses were generally confirmed as well. Women showed a small significant advantage in WIPS performance, with the effect size being similar to previous *meta*-analyses in the emotion domain (Thompson & Voyer, 2014). No significant correlation was found with age as predicted, which suggests that when contextually rich and multimodal scenarios are being evaluated, older adults may be able to compensate for potential deficits in single modalities and narrow domains (Isaacowitz & Stanley, 2011). This is also in line with studies suggesting that social perception may be more stable across adulthood than emotion perception (Castro & Isaacowitz, 2019). However, especially Study 5 mainly assessed younger adults with little variation in age. Future studies should therefore confirm the relationship of the WIPS with age.

Importantly, the WIPS predicted undergraduate students' membership in six out of nine student groups. Additionally, leaders in these six groups outperformed non-leader members in the same groups. Overall, the number of groups in which a student was a member was positively correlated with WIPS performance. These results underscore the

importance of IPA in social interactions within an organized setting (i.e., clubs or groups), although we did not assess the underlying mechanisms. Individuals with higher IPA might be more likely to be accepted as members in such organizations, but they might also be more likely to experience rewarding interactions and establish closer relationships within these groups. Members with particularly high IPA might be the ones to eventually get promoted into leadership positions in which understanding others may be even more important (Bass & Riggio, 2006; Zhong et al., 2013).

While these results provide first support for the predictive validity of the WIPS within organizational settings, more studies are needed that include practical outcomes and that examine the contribution of the measured "passive" perception ability to successful real-life interactions while controlling for potential third causal variables (Nowicki & Duke, 2001). For instance, future studies should examine how leaders' or managers' WIPS scores correlate with subordinates' perceptions of leadership effectiveness or job satisfaction, or with objective task performance of the team while controlling for intelligence (Elfenbein, Polzer, et al., 2007). These studies may also attempt to shed more light on the discrepancy often found between IPA correlations with peer- and self-ratings (e.g., Hall et al., 2009). That is, why exactly do IPA tests not correlate with self-reported leadership style (Study 6), social skill, or wellbeing (e.g., Hall et al., 2009; Schlegel, 2020) but show a variety of associations with other-rated interpersonal outcomes? One reason might be that high IPA is not always adaptive, but can have detrimental effects as well, for instance in terms of a hypersensitivity to emotional information (e.g., Schlegel, 2020).

7.2. Limitations and strength

A potential shortcoming of the WIPS is the low difficulty as compared to the ability distribution in our samples. Thus, in an Item Response Theory framework, the test measures IPA most precisely at low ability levels. This may have been caused by two opposing goals in the item selection process, similarly as for existing ERA tests (Schlegel et al., 2014). On the one hand, test items need to discriminate between people and the answers should not be obvious to the entire sample (i.e, too easy). On the other hand, difficult items (i.e., items that are solved by only few test-takers) may often be difficult solely because they contain ambiguous or too few valid cues that signal the correct answer, and thus lack internal validity (e.g., due to idiosyncratic actor performance). Such items are likely to be excluded during the test development process due to low item-total correlations, leaving mostly relatively easy items. As shown by Kenny (2013), in the IPA domain, easier items in fact yield more internally consistent tests than more difficult items.

The fact that the WIPS has higher measurement precision in the lower ability range does not mean that it is not useful for detecting individuals with higher ability, but it means that the test is less accurate when measuring more subtle differences among high-ability individuals than subtle differences among low-ability individuals. As the results in the studies presented here show, the test still yields practically useful correlates in domains characterized by normal or high interpersonal functioning, such as membership and leadership in student groups (Study 5) and personality judgment accuracy (Study 6). As such, the WIPS may be useful to improve recruitment or screening practices to help HR managers quantify so-called soft skills in a more objective way than by using self-report measures or by relying on recruiter intuition, which is susceptible to impression management tactics (Barrick et al., 2009) and bias such as perceived similarity (e.g., Graves & Powell, 1995).

The relatively low difficulty level could also represent an advantage for more clinically oriented studies with populations that have shown impairments in emotion recognition, e.g., when examining autistic traits, alexithymia, or disorders like depression (e.g., Kohler et al., 2009; Demenescu et al., 2010; Preis et al., 2020). Most studies in this domain use tests measuring only one modality (the face), and the WIPS could extend this research to more naturalistic scenarios presented in multiple modalities and to IPA beyond pure emotion recognition.

Although the WIPS items show situations in the workplace, we believe that many of the measured aspects, such as deciding who out of two individuals is more agreeable, how satisfied a person is with an interaction, whether a person is going to accept a proposition or not etc., could translate to other life domains as well. In addition, the WIPS requires only general knowledge of professional contexts that most adults likely have. Therefore, another applied setting could be personnel development and vocational training; the test could be used in addition to more established tests to help practitioners reliably identify low-scoring employees or students for training programs in fields in which interpersonal interactions are important (e.g., healthcare, consulting, customer support, therapy etc.). That said, such applied settings warrant additional research integrating the WIPS with other IPA measures before the WIPS can be used to make consequential decisions about individuals. Also, it remains an open empirical question whether IPA as measured in the WIPS generalizes to other contexts and life domains outside the organizational field.

The total administration time of 20 min should allow researchers and practitioners to include the WIPS in most existing test batteries or in selection and assessment protocols. Notably, given that the test was designed to rely on nonverbal content, it can potentially be used in languages other than English, French, and German examined here (i.e., only the items and instructions would need to be translated). However, further research is needed to assess the psychometric quality of such translations and the comparability of WIPS scores across different cultures. Previous studies suggest that there is only partial consensus between Western and Eastern or other non-Western respondents judging personality and affect from each respective culture (Albright et al., 1997; Gendron et al., 2014; Matsumoto & Kudoh, 1993). The interaction context may also activate different behavioral scripts across cultures, as has been shown for negotiation (Adair et al., 2001). Finally, Iizuka, Patterson, and Matchen (2002) found that Japanese and American participants scored similarly on the IPT in the visual only condition but not when audio-cues were available. These potential cultural or language differences might also be relevant for the WIPS.

Another limitation of the WIPS is that it is not suited to assess participants' relative strengths and weaknesses in the different IPA domains, given that each domain is represented with only few items and all items were specifically selected to measure a more global skill. In addition, as we limited the maximum duration of the test to about 20 min, it was not feasible within the current project to develop longer subtests for the different domains. Although in many professional settings, employees likely need several or all IPA domains, there are certainly contexts where one domain is particularly important. For instance, Hurley et al. (2014) showed that the accurate recognition of micro-expressions of emotions is beneficial for security officers, and deception detection more generally may be important in the fields of security and law enforcement. The WIPS does not measure these types of IPA.

Thus, whenever a researcher or practitioner is interested in a specific IPA domain, for instance for establishing training needs, more specific tests than the WIPS should be used if available. For instance, for measuring affect recognition in healthcare providers, the Patient Emotion Cue Test (PECT; Blanch-Hartigan, 2011) could be used. For several domains, however, no standard tests are available to date (e.g., deception detection or personality judgments), and we hope that the present research will inspire other researchers to start developing such tests. The implication of creating a subtest for personality, for example, would be to include a larger and more diverse set of items showing target persons who cover the entire range of possible values on the traits assessed, rather than choosing only targets with extreme values as was done in the WIPS. Other accuracy criteria should also be considered in addition to self-rated personality, such as personality ratings from knowledgeable informants (Funder et al., 1995).

7.3. Conclusion and outlook

The WIPS is a unidimensional and reasonably internally consistent test with heterogeneous content, suggesting that IPA possibly contains a broad common core skill across domains (Schlegel et al., 2017). Importantly, this does not preclude the simultaneous existence of more specific types of skills necessary in each domain, such as making inferences about trait *profiles* of targets, knowledge on specific traits or states, or the ability to detect very brief and subtle cues in single domains (e.g., facial micro-expressions; Hurley et al., 2014).

Further, the present research does not inform us about whether the IPA that people show in face-to-face interactions is the same IPA as measured in a standard test of passive and nonverbal "receiving ability" such as the WIPS. For instance, people can influence others' behavior to elicit cues that may facilitate accurate judgments, and they can also use linguistic information to assess others' states and traits. It is thus crucial for future research to examine how accuracy measured with standard tests correlates with accuracy measuring in live interactions. Another task for future research is to examine the real-life outcomes and correlates of IPA in more detail. While most research conceptualizes IPA as an adaptive skill, some studies also suggest that high IPA may have downsides (e.g., Elfenbein & Ambady, 2002; Bechtoldt et al., 2011; Schlegel, 2020).

In sum, the current results establish good psychometric properties of the WIPS test. This research offers the field a new and promising performance-based measure of broad IPA skill which hopefully will foster further theoretical development as well as the application of IPA testing in professional, educational, and other contexts.

The WIPS test is freely accessible for academic research purposes only, due to protected intellectual property rights. Interested academically affiliated researchers can apply for access via the webpage XXXX. They will receive a download link to the test in their mailbox.

CRediT authorship contribution statement

Nele Dael: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Katja Schlegel:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Adele E. Weaver:** Writing – original draft. **Mollie A. Ruben:** Writing – original draft. **Marianne Schmid Mast:** Conceptualization, Funding acquisition, Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2021.104182.

References

Adair, W. L., Okumura, T., & Brett, J. M. (2001). Negotiation behavior when cultures collide: The United States and Japan. *Journal of Applied Psychology, 86*(3), 371–385. https://doi.org/10.1037/0021-9010.86.3.371

Albright, L., Dong, Q., Fang, X., Malloy, T. E., Kenny, D. A., Winquist, L., & Yu, D. (1997). Cross-cultural consensus in personality judgments. *Journal of Personality and Social Psychology, 72*(3), 558–569. https://doi.org/10.1037/0022-3514.72.3.558

Ames, D. R., & Kammrath, L. K. (2004). Mind-reading and metacognition: Narcissism, not actual competence, predicts self-estimated ability. *Journal of Nonverbal Behavior, 28*(3), 187–209. https://doi.org/10.1023/B:JONB.0000039649.20015.0e

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 98–124). Cambridge University Press. https://doi.org/10.1017/cbo9781316181959.005.

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion, 9*(5), 691–704. https://doi.org/10.1037/a0017088

Bänziger, T., Scherer, K. R., Hall, J. A., & Rosenthal, R. (2011). Introducing the MiniPONS: A short multichannel version of the Profile of Nonverbal Sensitivity (PONS). *Journal of Nonverbal Behavior, 35*(3), 189–204. https://doi.org/10.1007/s10919-011-0108-3

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology, 94*(6), 1394–1411. https://doi.org/10.1037/a0016532

Bass, B. M., & Avolio, B. J. (1992). Developing transformational leadership: 1992 and beyond. *Journal of European Industrial Training, 14*(5), 21–27.

Bass, B. M., & Riggio, R. E. (2006). *Transformational leadership*. Psychology Press.

Baudson, T. G., & Preckel, F. (2016). mini-q: Intelligenzscreening in drei Minuten. *Diagnostica, 62*(3), 182–197. https://doi.org/10.1026/0012-1924/a000150

Bechtoldt, M. N., Beersma, B., Rohrmann, S., & Sanchez-Burks, J. (2011). A gift that takes its toll: Emotion recognition and conflict appraisal. *European Journal of Work and Organizational Psychology, 22*(1), 56–66. https://doi.org/10.1080/1359432X.2011.614726

Bernieri, F. J. (2001). Toward a taxonomy of interpersonal sensitivity. In J. A. Hall, & F. J. Bernierri (Eds.), *The LEA series in personality and clinical psychology. Interpersonal sensitivity: Theory and measurement* (pp. 3–20). Lawrence Erlbaum Associates, Inc.. https://doi.org/10.4324/9781410600424-8

Biesanz, J. C. (2021). The social accuracy model. In T. D. Letzring, & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment* (pp. 61–82). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.5.

Blanch-Hartigan, D. (2011). Measuring providers' verbal and nonverbal emotion recognition ability: Reliability and validity of the Patient Emotion Cue Test (PECT). *Patient Education and Counseling, 82*(3), 370–376. https://doi.org/10.1016/j.pec.2010.11.017

Boone, R. T., & Schlegel, K. (2016). Is there a general skill in perceiving others accurately? In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 379–403). Cambridge University Press.

Byron, K., Terranova, S., & Nowicki, S. (2007). Nonverbal emotion recognition and salespersons: Linking ability to perceived and actual success. *Journal of Applied Social Psychology, 37*(11), 2600–2619. https://doi.org/10.1111/j.1559-1816.2007.00272.x

Castro, V. L., Halberstadt, A. G., Lozada, F. T., & Craig, A. B. (2015). Parents' emotion-related beliefs, behaviours, and skills predict children's recognition of emotion. *Infant and Child Development, 24*(1), 1–22.

Castro, V. L., & Isaacowitz, D. M. (2019). The same with age: Evidence for age-related similarities in interpersonal accuracy. *Journal of Experimental Psychology: General, 148*(9), 1517–1537. https://doi.org/10.1037/xge0000540

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance, 18*(2), 123–149. https://doi.org/10.1207/s15327043hup1802_2

Colman, D. E., Letzring, T. D., & Biesanz, J. C. (2017). Seeing and feeling your way to accurate personality judgments: The moderating role of perceiver empathic tendencies. *Social Psychological and Personality Science, 8*(7), 806–815.

Costanzo, M., & Archer, D. (1989). Interpreting the expressive behavior of others: The Interpersonal Perception Task. *Journal of Nonverbal Behavior, 13*(4), 225–245. https://doi.org/10.1007/BF00990295

Custrini, R. J., & Feldman, R. S. (1989). Children's social competence and nonverbal encoding and decoding of emotions. *Journal of Clinical Child Psychology, 18*(4), 336–342. https://doi.org/10.1207/s15374424jccp1804_7

Davis, M. H., & Kraus, L. A. (1997). Personality and empathic accuracy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 144–168). Guilford Press.

Demenescu, L. R., Kortekaas, R., den Boer, J. A., & Aleman, A. (2010). Impaired attribution of emotion to facial expressions in anxiety and major depression. *PLoS ONE, 5*(12), Article e15058. https://doi.org/10.1371/journal.pone.0015058

Elfenbein, H. A., & Ambady, N. (2002). Predicting workplace outcomes from the ability to eavesdrop on feelings. *Journal of Applied Psychology, 87*(5), 963–971.

Elfenbein, H. A., Foo, M. D., White, J., Tan, H. H., & Aik, V. C. (2007). Reading your counterpart: The benefit of emotion recognition accuracy for effectiveness in negotiation. *Journal of Nonverbal Behavior, 31*(4), 205–223. https://doi.org/10.1007/s10919-007-0033-7

Elfenbein, H. A., Polzer, J. T., & Ambady, N. (2007). Team emotion recognition accuracy and team performance. In C. E. J. Härtel, N. M. Ashkanasy, & W. J. Zerbe (Eds.), *Research on emotion in organizations* (Vol. 3, pp. 87–119). Emerald Group Publishing Limited. https://doi.org/10.1016/S1746-9791(07)03004-0.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology, 69*(4), 656–672. https://doi.org/10.1037/0022-3514.69.4.656

Furr, R. M. (2008). A Framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality, 76*(5), 1267–1316. https://doi.org/10.1111/j.1467-6494.2008.00521.x

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *Various coefficients of interrater reliability and agreement (R package 'irr' v. 0.84)*.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion, 14*(2), 251–262. https://doi.org/10.1037/a0036052

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Graves, L. M., & Powell, G. N. (1995). The effect of sex similarity on recruiters' evaluations of actual applicants: A test of the similarity-attraction paradigm. *Personnel Psychology, 48*(1), 85–98. https://doi.org/10.1111/j.1744-6570.1995.tb01747.x

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*(4), 845–857. Doi: 0033-2909/78/8504-0845$00.75.

Hall, J. A., Andrzejewski, S., & Yopchick, J. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior, 33*(3), 149–180. https://doi.org/10.1007/s10919-009-0070-5

Hall, J. A., Gunnery, S. D., Letzring, T. D., Carney, D. R., & Colvin, C. R. (2017). Accuracy of judging affect and accuracy of judging personality: How and when are they related? *Journal of Personality, 85*(5), 583–592.

Hall, J. A., Schmid Mast, M., & West, T. V. (2016). Accurate interpersonal perception: Many traditions, one topic. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 3–22). Cambridge University Press. https://doi.org/10.1017/cbo9781316181959.001.

Hall, J. A., Ship, A. N., Ruben, M. A., Curtin, E. M., Roter, D. L., Clever, S. L., … Pounds, K. (2014). The Test of Accurate Perception of Patients' Affect (TAPPA): An ecologically valid tool for assessing interpersonal perception accuracy in clinicians. *Patient Education and Counseling, 94*(2), 218–223. https://doi.org/10.1016/J.PEC.2013.10.004

Human, L. J., Carlson, E. N., Geukes, K., Nestler, S., & Back, M. D. (2020). Do accurate personality impressions benefit early relationship development? The bidirectional associations between accuracy and liking. *Journal of Personality and Social Psychology, 118*(1), 199–212. https://doi.org/10.1037/pspp0000214

Human, L. J., & Mendes, W. B. (2018). Cardiac vagal flexibility and accurate personality impressions: Examining a physiological correlate of the good judge. *Journal of Personality, 86*(6), 1065–1077.

Hurley, C. M., Anker, A. E., Frank, M. G., Matsumoto, D., & Hwang, H. C. (2014). Background factors predicting accuracy and improvement in micro expression recognition. *Motivation and Emotion, 38*(5), 700–714. https://doi.org/10.1007/s11031-014-9410-9

Ickes, W. (2001). Measuring empathic accuracy. In J. A. Hall, & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 219–241). Lawrence Erlbaum Associates Publishers.

Ickes, W., & Hodges, S. D. (2013). Empathic accuracy in close relationships. In J. A. Simpson, & L. Campbell (Eds.), *The Oxford handbook of close relationships* (pp. 348–373). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195398694.013.0016.

Isaacowitz, D. M., & Stanley, J. T. (2011). Bringing an ecological perspective to the study of aging and recognition of emotional facial expressions: Past, current, and future methods. *Journal of Nonverbal Behavior, 35*(4), 261–278.

Iizuka, Y., Patterson, M. L., & Matchen, J. C. (2002). Accuracy and confidence on the interpersonal perception task: A Japanese-American comparison. *Journal of Nonverbal Behavior, 26*(3), 159–174. https://doi.org/10.1023/A:1020761332372

Jackson, D. N. (1984). *Personality Research Form manual* (3rd ed.). Research Psychologists Press.

Jaksic, C., & Schlegel, K. (2020). Accuracy in judging others' personalities: The role of emotion recognition, emotion understanding, and trait emotional intelligence. *Journal of Intelligence, 8*(3), 34. https://doi.org/10.3390/jintelligence8030034

Kaltwasser, L., Hildebrandt, A., Wilhelm, O., & Sommer, W. (2017). On the relationship of emotional abilities and prosocial behavior. *Evolution and Human Behavior, 38*(3), 298–308.

Kenny, D. A. (2013). Issues in the measurement of judgmental accuracy. In S. Baron-Cohen, H. Tager-Flusberg, & M. V. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 104–116). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199692972.003.0007.

Kohler, C. G., Walker, J. B., Martin, E. A., Healey, K. M., & Moberg, P. J. (2009). Facial emotion perception in schizophrenia: A meta-analytic review. *Schizophrenia Bulletin, 36*(5), 1009–1019.

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality, 42*(2), 914–932. https://doi.org/10.1016/j.jrp.2007.12.003

Letzring, T. D., & Funder, D. C. (2018). Interpersonal accuracy in trait judgments. In *The SAGE handbook of personality and individual differences: Applications of personality and individual differences* (pp. 253–282). SAGE Publications Ltd.. https://doi.org/10.4135/9781526451248.n11

Mair, P., Hatzinger, R., & Maier, M. J. (2019). *eRm: Extended Rasch Modeling. 1.0-0.* http://erm.r-forge.r-project.org/.

Matthews, G., Zeidner, M., & Roberts, R. (2004). *Emotional intelligence: Science and myth.* MIT Press.

Matsumoto, D., & Kudoh, T. (1993). American-Japanese cultural differences in attributions of personality based on smiles. *Journal of Nonverbal Behavior, 17*(4), 231–243. https://doi.org/10.1007/BF00987239

Momm, T., Blickle, G., Liu, Y., Wihler, A., Kholin, M., & Menges, J. I. (2015). It pays to have an eye for emotions: Emotion recognition ability indirectly predicts annual income. *Journal of Organizational Behavior, 36*(1), 147–163. https://doi.org/10.1002/job.1975

Muthén, L. K., & Muthén, B. O. (2011). *Mplus User's Guide* (6th ed.). Muthén & Muthén.

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior, 18*(1), 9–35. https://doi.org/10.1007/BF02169077

Nowicki, S., & Duke, M. P. (2001). Nonverbal receptivity: The Diagnostic Analysis of Nonverbal Accuracy (DANVA). In J. A. Hall, & F. J. Bernieri (Eds.), *The LEA series in personality and clinical psychology. Interpersonal sensitivity: Theory and measurement* (pp. 183–198). Lawrence Erlbaum Associates Publishers.

Nunally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Otto, J. H., Döring-Seipel, E., Grebe, M., & Lantermann, E.-D. (2001). Entwicklung eines Fragebogens zur Erfassung der wahrgenommenen emotionalen Intelligenz. Aufmerksamkeit auf, Klarheit und Beeinflussbarkeit von Emotionen. [Development of a questionnaire for measuring perceived emotional intelligence: Attention to, clarity. *Diagnostica, 47*(4), 178–187. https://doi.org/10.1026//0012-1924.47.4.178

Parkinson, B., Fischer, A., & Manstead, A. S. R. (2005). *Emotion in social relations: Cultural, group, and interpersonal processes.* Psychology Press.

Plaisant, O., Courtois, R., Réveillère, C., Mendelsohn, G. A., & John, O. P. (2010). Validation par analyse factorielle du Big Five Inventory français (BFI-Fr). Analyse convergente avec le NEO-PI-R. *Annales Médico-Psychologiques, Revue Psychiatrique, 168*(2), 97–106. https://doi.org/10.1016/J.AMP.2009.09.003

Preis, M. A., Schlegel, K., Stoll, L., Blomberg, M., Schmidt, H., Wünsch-Leiteritz, W., … Brockmeyer, T. (2020). Improving emotion recognition in anorexia nervosa: An experimental proof-of-concept study. *International Journal of Eating Disorders, 53*(6), 945–953. https://doi.org/10.1002/eat.23276

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika, 74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Riggio, R. E. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology, 51*(3), 649–660. https://doi.org/10.1037/0022-3514.51.3.649

Riggio, R. E. (2005). The Social Skills Inventory (SSI): Measuring nonverbal and social skills. In V. Manusov (Ed.), *The sourcebook of nonverbal measures: Going beyond words* (pp. 25–33). Lawrence Erlbaum Associates Inc.

Riggio, R. E., & Carney, D. R. (2003). *Social Skills Inventory manual* (2nd ed.). Mind Garden, Inc. (P, D).

Roberts, R. D., MacCann, C., Matthews, G., & Zeidner, M. (2010). Emotional intelligence: Toward a consensus of models and measures. *Social and Personality Psychology Compass, 4*(10), 821–840. https://doi.org/10.1111/j.1751-9004.2010.00277.x

Rogers, K. H., & Biesanz, J. C. (2019). Reassessing the good judge of personality. *Journal of Personality and Social Psychology, 117*(1), 186–200. https://doi.org/10.1037/pspp0000197

Rogers, K. H., Wood, D., & Furr, R. M. (2018). Assessment of similarity and self-other agreement in dyadic relationships: A guide to best practices. *Journal of Social and Personal Relationships, 35*(1), 112–134. https://doi.org/10.1177/0265407517712615

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

Rosip, J. C., & Hall, J. A. (2004). Knowledge of nonverbal cues, gender, and nonverbal decoding accuracy. *Journal of Nonverbal Behavior, 28*(4), 267–286. https://doi.org/10.1007/s10919-004-4159-6

Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews, 32*(4), 863–881. https://doi.org/10.1016/j.neubiorev.2008.01.001

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology, 69*(4), 719–727. https://doi.org/10.1037/0022-3514.69.4.719

Scherer, K. R. (2007). Component models of emotion can inform the quest for emotional competence. In G. Matthews, M. Zeidner, & R. D. Roberts (Eds.), *The science of emotional intelligence: Knowns and unknowns* (pp. 101–126). Oxford University Press.

Schlegel, K. (2020). Inter- and intrapersonal downsides of accurately perceiving others' emotions. In R. J. Sternberg, & A. Kostić (Eds.), *Social intelligence and nonverbal communication* (pp. 359–395). Springer International Publishing. https://doi.org/10.1007/978-3-030-34964-6_13.

Schlegel, K., Boone, T. R., & Hall, J. A. (2017). Individual differences in interpersonal accuracy: A multi-level meta-analysis to assess whether judging other people is one skill or many. *Journal of Nonverbal Behavior, 41*(2), 103–137. https://doi.org/10.1007/s10919-017-0249-0

Schlegel, K., Fontaine, J. R. J., & Scherer, K. R. (2019). The nomological network of emotion recognition ability: Evidence from the Geneva emotion recognition test. *European Journal of Psychological Assessment, 35*(3), 352–363. https://doi.org/10.1027/1015-5759/a000396

Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the Geneva Emotion Recognition Test: An example of Rasch-based test development. *Psychological Assessment, 26*(2), 666–672. https://doi.org/10.1037/a0035246

Schlegel, K., Mehu, M., van Peer, J. M., & Scherer, K. R. (2018). Sense and sensibility: The role of cognitive and emotional intelligence in negotiation. *Journal of Research in Personality, 74*, 6–15. https://doi.org/10.1016/j.jrp.2017.12.003

Schlegel, K., & Mortillaro, M. (2019). The Geneva Emotional Competence Test (GECo): An ability measure of workplace emotional intelligence. *Journal of Applied Psychology, 104*(4), 559–580. https://doi.org/10.1037/apl0000365

Schlegel, K., Palese, T., Schmid Mast, M., Rammsayer, T. H., Hall, J. A., & Murphy, N. A. (2020). A meta-analysis of the relationship between emotion recognition ability and

intelligence. *Cognition and Emotion, 34*(2), 329–351. https://doi.org/10.1080/02699931.2019.1632801

Schlegel, K., & Scherer, K. R. (2016). Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation. *Behavior Research Methods, 48*(4), 1383–1392. https://doi.org/10.3758/s13428-015-0646-4

Schlegel, K., & Scherer, K. R. (2018). The nomological network of emotion knowledge and emotion understanding in adults: Evidence from two new performance-based tests. *Cognition and Emotion, 32*(8), 1514–1530. https://doi.org/10.1080/02699931.2017.1414687

Schmidt, F. L., & Hunter, J. E. (2014). Meta-analysis of correlations corrected individually for artifacts. In F. L. Schmidt, & J. E. Hunter (Eds.), *Methods of meta-analysis: Correcting error and bias in research findings* (pp. 75–135). Newbury Park: Sage.

Schmid Mast, M., & Darioly, A. (2014). Emotion recognition accuracy in hierarchical relationships. *Swiss Journal of Psychology, 73*(2), 69.

Schmid Mast, M., & Hall, J. A. (2018). The impact of interpersonal accuracy on behavioral outcomes. *Current Directions in Psychological Science, 27*(5), 309–314. https://doi.org/10.1177/0963721418758437

Schmid Mast, M., & Latu, I. (2016). Interpersonal accuracy in relation to the workplace, leadership, and hierarchy. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 270–286). Cambridge University Press.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Simpson, J. A., Orina, M. M., & Ickes, W. (2003). When accuracy hurts, and when it helps: A test of the empathic accuracy model in marital interactions. *Journal of Personality and Social Psychology, 85*(5), 881.

Sullivan, S., & Ruffman, T. (2004). Social understanding: How does it fare with advancing years? *British Journal of Psychology, 95*(1), 1–18. https://doi.org/10.1348/000712604322779424

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion, 4*(1), 46.

Thompson, A. E., & Voyer, D. (2014). Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis. *Cognition and Emotion, 28*(7), 1164–1195. https://doi.org/10.1080/02699931.2013.875889

Vogt, D. S., & Colvin, C. R. (2003). Interpersonal orientation and the accuracy of personality judgments. *Journal of Personality, 71*(2), 267–295.

Zhong, Y., Zhang, S., & Chen, Y. (2013). Differences of interpersonal sensitivity between high-power people and low-power people. *Chinese Journal of Clinical Psychology*.