Journal of Personality 84:4, August 2016
© 2015 Wiley Periodicals, Inc.
DOI: 10.1111/jopy.12170

Individual Differences in Accurately Judging Personality From Text

Judith A. Hall,[1] Jin X. Goh,[1] Marianne Schmid Mast,[2] and Christian Hagedorn[1]
[1]Northeastern University
[2]University of Lausanne

Abstract

This research examines correlates of accuracy in judging Big Five traits from first-person text excerpts. Participants in six studies were recruited from psychology courses or online. In each study, participants performed a task of judging personality from text and performed other ability tasks and/or filled out questionnaires. Participants who were more accurate in judging personality from text were more likely to be female; had personalities that were more agreeable, conscientious, and feminine, and less neurotic and dominant (all controlling for participant gender); scored higher on empathic concern; self-reported more interest in, and attentiveness to, people's personalities in their daily lives; and reported reading more for pleasure, especially fiction. Accuracy was not associated with SAT scores but had a significant relation to vocabulary knowledge. Accuracy did not correlate with tests of judging personality and emotion based on audiovisual cues. This research is the first to address individual differences in accurate judgment of personality from text, thus adding to the literature on correlates of the good judge of personality.

As far as we are aware, all research on individual differences in accurate interpersonal perception is based on judgment of nonverbal cues only (i.e., face, gesture, posture, vocal quality) or nonverbal cues combined with linguistic content (Hall & Bernieri, 2001; Nowicki & Duke, 2013; Vogt & Colvin, 2003). However, in this voluminous literature on measuring individual differences in accuracy of judging other people's states and traits, there have not been any using purely linguistic stimuli—what people say, isolated from other cues conveyed via voice, face, or body.

Yet, in life, people make use of others' words constantly when drawing inferences about them, whether in a face-to-face interaction where they have access to words along with other cues, or from reading what others have said in text messages, emails, blogs, Web pages, personal and professional correspondence, tweets, and online social media. In fact, the frequency with which people are exposed exclusively to nonverbal cues is very small compared to how often they have access to a target person's words. The present research begins to fill this gap by measuring individuals' accuracy of inferences based on purely linguistic stimuli in one content domain—judging Big Five personality traits—and investigating correlates of this skill.

Though the present research is about judging personality, one could also measure accuracy of judging other characteristics of people from their words—their emotions, background, or hierarchical status, for example. In fact, research on accuracy in "mind reading"—inferring the specific contents of another person's thoughts and feelings that are spontaneously revealed in a conversation—finds that accuracy when seeing and hearing real people interacting depends far more on their words than on the accompanying visible or audible nonverbal cues (Gesn & Ickes, 1999; Hall & Schmid Mast, 2007; Zaki, Bolger, & Ochsner, 2009). This is not to diminish the important role of nonverbal communication for drawing many kinds of inferences in daily life; we are only saying that typically the words are present and they are often important, perhaps even more important than nonverbal cues. Research should therefore focus more on accurate judgment of people through words.

The more general question of how personality is revealed and judged in words has received considerable research attention, however, and we begin with a brief overview of approaches that have been taken toward that end.

The Study of Personality in Text

Tskhay and Rule (2014), in a meta-analysis, summarized research that reported raters' agreement with each other when inferring the Big Five traits from the words contained in social networking sites and other online text-based media (e.g.,

The authors thank Mitja Back and Nicholas Rule for their advice on the article, and Aria Rad and Julia Zuroff for research assistance.

Correspondence concerning this article should be addressed to Judith A. Hall, Department of Psychology, Northeastern University, Boston, MA 02115. Email: j.hall@neu.edu.

blogs). Agreement (i.e., interjudge reliability) was highest for judging Extraversion from such material and lowest for judging Neuroticism, though still above zero. The specific cues raters utilize—rightly or wrongly—when drawing personality inferences from text have also been studied (cue utilization in lens model terminology; Brunswik, 1956). Küfner, Back, Nestler, and Egloff (2010) found, for example, that writers of brief creative writing samples were judged by readers to be high on Openness if the writing was sophisticated and was more creative, and they were judged to be more Agreeable if the writing had more positive and social orientation words.

To find out whether people with different personalities actually use words differently, researchers measure word usage or linguistic communication style and correlate that with the known personality of the writers or speakers (cue validities in lens model terminology). Often, measurement has been done with the Linguistic Inquiry and Word Count (LIWC) software (e.g., Beukeboom, Tanis, & Vermeulen, 2013; Fast & Funder, 2008; Hirsh & Peterson, 2009; Holtgraves, 2011; Pennebaker & King, 1999; Tausczik & Pennebaker, 2010). Linguistic correlates of personality have been identified this way. For instance, writers of Facebook messages who were more Neurotic used words indicating depression and the phrase “sick of” at a higher frequency than those who were less Neurotic (Schwartz et al., 2013; for a review, see Pennebaker, Mehl, & Niederhoffer, 2003).

Yet another question relating to personality and word use is whether perceivers are accurate. This is examined by correlating perceivers’ personality judgments based on reading the target persons’ words with criterion personality information. Holleran and Mehl (2008) found substantial accuracy for all of the Big Five traits based on reading 20-min stream-of-consciousness essays, and Wall, Taylor, Dixon, Conchie, and Ellis (2013) obtained accuracy for judging Openness from raters’ judgments of email conversations. The meta-analysis of studies based on online media samples done by Tskhay and Rule (2014) showed clearly that readers of such material are accurate above chance at judging all of the Big Five traits except for Neuroticism, but even there, accuracy was positive in direction.

And, finally, researchers have put judgments, cues, and criteria together to learn which cues are correctly and incorrectly used when perceivers make their personality inferences (e.g., Back, Schmukle, & Egloff, 2008; Gifford & Hine, 1994; Stopfer, Egloff, Nestler, & Back, 2014). As an example, Küfner et al. (2010) found that sophisticated and creative writing were correctly used in judging Openness, and positive and social orientation words were correctly used in judging Agreeableness.

The present research complements previous studies on personality in words by taking the individual differences approach not previously taken—by measuring accuracy and its correlates.

Accuracy of Interpersonal Inference

In setting out to understand individual differences in ability to infer personality from text, the present research can be compared to the large literature on correlates of individual differences in accuracy based on audiovisual cues. Although there is an abundance of such research (for meta-analyses, see Elfenbein & Ambady, 2002; Hall, 1984; Hall, Andrzejewski, & Yopchick, 2009; Hall, Schmid Mast, & Latu, 2014; Kirkland, Peterson, Baker, Miller, & Pulos, 2013; Marsh & Blair, 2008; Murphy & Hall, 2011; Savla, Vella, Armstrong, Penn, & Twamley, 2013; Thompson & Voyer, 2014), the great majority of studies of interpersonal accuracy are on judging affective states rather than personality, and none of them used text as stimuli to be judged. In fact, relatively little research has looked for correlates of accuracy in judging personality (Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, 2005; Funder, 2001; Letzring, 2008; Lippa & Dietz, 2000). The present six studies were designed to expand understanding of the so-called good judge of personality, with respect to judgments based on text. Some of the research questions were similar to those that have been asked with regard to accurate interpersonal perception based on audiovisual cues (e.g., correlations with gender, cognitive ability, personality), and some questions were novel to the present research (e.g., interest in personality, tendency to write about personality).

Across the six studies, some of the correlates were examined more than once and the results are presented meta-analytically. For this reason, the methodology of all of the studies is described first, followed by the results.

METHOD

Measuring Accuracy of Judging Personality From Text

Because an instrument did not exist, we created one for the present research, which is hereafter referred to as the APT, for Assessing Personality from Text. The goals for this test were that it would (a) have enough items to reflect a reasonable array of targets and excerpts while representing all Big Five traits, (b) be short enough for efficient administration, and (c) be easy to score in case future researchers wished to use it. The APT’s operational definition of accuracy is based on two criteria, both of which are prevalent in the interpersonal accuracy measurement literature (Hall & Bernieri, 2001; Hall, Bernieri, & Carney, 2005): self-other agreement, that is, agreement between the original target’s (i.e., excerpt writer’s) self-described personality and the test taker’s guess of that person’s personality; and an empirically derived criterion based on published research documenting how personality is reflected in linguistic behavior. Both of these are described below.

The APT items are 36 text excerpts, each a few lines in length, that were originally written by an actual person whose Big Five traits were measured (i.e., Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience). In all the present studies, the test-taking participants were given a brief description of context (i.e., what the original writer was writing about), and then they read the excerpt and indicated

Table 1 Sample Items for Assessing Personality From Text

Sample Item 1:		
Context: Think of a sport you like and explain why you like it.		
I love cross country. I ran all throughout high school doing cross country and track and field but cross country is where my passion is. I love getting a runner's high and running large distances on tracks generally located in natural settings. Finishing a long run is amazing. I am empty but full at the same time.		
Gender:	Male	Female
Anxiety:	*Low*	*High*
Activity level:	Low	High
Correct answer:	Low anxiety (i.e., low Neuroticism)	
Sample Item 2:		
Context: Write a message to your mother.		
I'm so thankful to have you in my life, you are the reason why I try my hardest at school, work, and any situation I come across. If there's one person I know I can count on, anywhere, anytime, it is you. Even when I am away from home. Despite all the mistakes I've made in my life so far, you still love me for who I am and I would be nowhere without you. Thank you Mama.		
Gender:	Male	Female
Anger:	*Low*	*High*
Cooperation:	Low	High
Correct answer:	Low anger (i.e., low Neuroticism)	

Note. Scored answer is shown in italics.

whether the target person was high or low on two named traits and whether the target person was male or female. One of the named traits and the gender question were distractors intended to keep the task from being too easy, whereas the other trait was scored for accuracy. As an example, if the target person was selected for being low on Neuroticism, the participant would be asked about Neuroticism along with one other personality trait and the target's gender, and only his or her answer regarding Neuroticism would be scored. Table 1 shows two sample APT items. Items were scored as 0 or 1 (incorrect or correct) to reflect whether the participant correctly identified the level of the trait in question (high or low), and then items were averaged so that the APT total score was the proportion of items correctly answered. The individual Big Five traits are not separately scored due to the relatively few items representing each. The test takes approximately 15 to 20 min to complete.

Collection of Writing Excerpts. To create a large pool of writing excerpts as potential test items, writing excerpts were collected from nine participants who were recruited from the Northeastern University Psychology Department Participant Pool and given partial course credit for participation, or from flyers offering $10 for participation. This and all studies described in the present article were approved by the Northeastern University Committee on Human Subject Protection. It was explained to participants that they were generating brief writing excerpts (not more than 10 lines long) that would be revised before being shown to later participants. Participants were told not to include any personally identifying information about themselves or other people and not to write anything they would not want others to read. They were also told that they did not have to be truthful in what they wrote and that they should not worry about grammatical errors, typos, or writing well. Participants were told they did not have to be truthful in order to ensure the confidentiality of their data and because some of the writing prompts were not necessarily events they would have encountered themselves.

The same 26 writing assignments (contexts) were given to all participants. Participants were told they should do as many as they could, but they were not told which ones to do. On average, participants were able to complete 17 of the 26 prompts in their experimental hour. For illustration, 10 of these contexts were as follows: (a) Write a message to your mother. Talk about anything you want. (b) Write about your family pet, or the pet of a friend or neighbor. Write anything you want on this topic. (c) Write about your sister or brother, or a relative who is in your generation (such as a cousin). Describe that person. (d) Write about something scary that happened to you. (e) Write a message of complaint to a company about their product or service, such as you might in an email to their complaint department. (f) Write about a tough semester or a tough exam you had to prepare for. (g) Think of a sport you like and explain why you like it. (h) Write a thank-you note to a professor for writing letters of recommendation for you. (i) Write a short telephone dialogue between two friends who are chatting about what to do this weekend. (j) Write about your study habits and whether you think they are good or bad. With this method, well over 100 excerpts were generated from which to select a set of test stimuli based on the criteria described next.

Each participant then filled in the Ten-Item Personality Inventory (TIPI), a brief measure of the Big Five (Gosling, Rentfrow, & Swann, 2003). Each participant's Big Five scores were scrutinized and, for each trait separately, each participant was classified as clearly high or low on the trait, or in the middle range. Thus, a given participant might (for example) be classified as high on Extraversion and low on Neuroticism, in which case that participant's excerpts were eligible for inclusion as Extraversion and Neuroticism items, but that same participant might have been classified as in the middle range on the remaining traits, in which case none of that participant's excerpts would be included to represent those traits. This procedure ensured maximum variance between the "high" and "low" trait excerpts.

Enhancement of Writing Excerpts. Because there are psychometric gains from having more items on a test, we deliberately kept the excerpts short. However, short excerpts are likely to contain fewer diagnostic cues to personality than longer excerpts. To increase the possibility of accurate judgment based on short excerpts, the excerpts were enhanced subtly. This was done by altering the excerpts in minor ways, by adding or subtracting word usages, so that the excerpt would more clearly reflect the original target person's actual personality (i.e., high or low on one of the Big Five traits). These slight alterations were based on valid personality cues that had been reported in published studies documenting how word usage differed for the Big Five traits (e.g., Hirsh & Peterson, 2009; Pennebaker & King, 1999; Pennebaker et al., 2003; Tauszik & Pennebaker, 2010; Yarkoni, 2010). Introducing this element of artificiality into the stimuli did not undermine the potential validity of the test, as the changes were made based on previously validated information and the point was not to develop a corpus of naturalistic or representative writing samples but rather to produce a set of test stimuli suitable for detecting individual variation among participants in their ability to pick up on personality-relevant cues.

Selection of Final APT Items. In Study 1 (N = 45; 39% male; M_{age} = 18.73), Northeastern University undergraduates from the Psychology Department Participant Pool were given 48 potential APT items. These 48 items were selected to represent all five traits approximately equally, with a rough balance of excerpts whose original writers were high versus low on each trait. On average, the excerpts were 83 words long. After the data were scored, items were dropped based on accuracy levels (extremely low or extremely high) and item-total correlations, leaving the 36 items composing the final test. Seven or eight items represented each of the Big Five traits, approximately balanced for a high or low level of the trait.

Overview of Studies and Variables

In all six studies, the APT's psychometric characteristics were assessed and the APT was correlated with other variables thought to be relevant to the construct. Hereafter, the term *participants* refers to people who took the APT, not the participants whose writing excerpts served as the stimuli.

Gender. The correlation of the APT with gender was examined in all studies. Gender is one of the most consistent correlates of accuracy of interpersonal perception in the broader literature based on audiovisual stimuli (Hall & Gunnery, 2013). Women typically score higher on tests involving affect judgment, and often on tests of personality judgment (e.g., Vogt & Colvin, 2003).

Personality Traits. Participants' own Big Five traits were examined in relation to the APT in all six studies. The more general literature based on audiovisual stimuli finds accuracy of interpersonal perception to be positively correlated with Extraversion, Conscientiousness, and Openness to Experience, and negatively correlated with Neuroticism (Hall et al., 2009).

Personality Dominance as a separate trait was measured in all six studies, but in an exploratory manner because the broader audiovisual literature is unclear on how accuracy of interpersonal perception is related to personality Dominance (Hall et al., 2014).

The personality traits of Agency and Communion were examined because the broader audiovisual literature finds that Femininity, but not Masculinity, is associated with accurate interpersonal perception (Hall et al., 2009), and Vogt and Colvin (2003) found that Communion as a personality trait predicted accuracy of judging personality.

The concept of empathy was measured in terms of several personality dimensions (e.g., Empathic Concern, Perspective Taking) because the broader audiovisual literature finds a positive association with accurate interpersonal perception (Hall et al., 2009).

Cognitive Ability. Self-reported SAT scores were gathered in one study, and vocabulary knowledge was measured in three studies to establish discriminant validity. Although the SAT scores were self-reported, research shows that self-reported SAT scores are very highly correlated with actual SAT scores (Cole & Gonyea, 2010). The broader literature on accuracy of interpersonal perception using audiovisual tests finds a positive, though not very strong, correlation between general cognitive ability and accuracy (Murphy & Hall, 2011). In the case of the APT, it was considered especially important to establish discriminant validity, especially with respect to verbal ability, because the test consists of making decisions about written material. Early research on accurate social perception or "social intelligence," based on written stimuli (scenarios, etc.), could not establish discriminant validity due to high correlations with general intelligence (Shanley, Walker, & Foley, 1971; Walker & Foley, 1973). Therefore, a relatively small correlation of the APT with SAT scores and vocabulary knowledge would demonstrate that the test was not importantly confounded with general cognitive ability.

Interest in Personality. Interest in personality has not been measured in any research on personality judgment accuracy to our knowledge, yet it would seem to be a plausible correlate of this skill. The causal relation could be as antecedent, consequence, or both. Interest in personality was measured in two ways. First, a set of questionnaire items was written and used in three of the studies to measure how participants thought about personality and their self-reported attentiveness to others' personalities. Second, participants in one study were asked to write descriptions of people and then their writings were analyzed for allusions to personality traits. The hypothesis for both of these measures was that higher scores on them would be associated with higher APT scores because this would indicate that people having more engagement with the personality construct—for example, saying they like to describe people's personalities when talking about them to friends or family, or spontaneously mentioning personality traits when writing about someone—would also possess more skill in judging personality.

Reading Habits. With a test of accuracy in processing written material, an obvious domain of relevance is a person's experience and habits relating to reading. Additional questionnaire items were written and used in three studies to measure reading habits, in particular regarding the reading of fiction literature, with the hypothesis that because fiction literature contains a great deal of personality description, more fiction reading or enjoyment of reading fiction would correlate positively with the APT.

Other Measures of Accurate Interpersonal Perception. In the literature on interpersonal accuracy based on audiovisual tests, authors have repeatedly noted the very weak, often negligible, correlations between different tests (e.g., Hall, 2001; Hall & Boone, 2014). Therefore, other tests could not be used as a basis for establishing convergent validity of the APT. Nevertheless, for exploratory purposes, in one of the present studies, five other measures of accuracy in interpersonal perception were administered (three on affect judgment and two on personality judgment), all of which were based on inferring the meanings of audiovisual cues.

Study 1 Method

The participants in this study were described above (N = 45). Only the 36 items selected for the final APT were analyzed for Study 1. Participants came singly or in pairs to the laboratory and completed the APT and all other measures individually on computers. Other measures were the following:

1. Self-reported Verbal and Quantitative SAT scores.
2. The TIPI (Gosling et al., 2003). The 10 items (each on a 1–7 scale) were scored after appropriate reversals so that high scores indicated Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience.[1]
3. Self-reported Dominant personality, captured in two items that were written using the same format as the TIPI: "dominant, controlling" and "a follower more than a leader, nonassertive" (reversed). These two items were averaged after appropriate reversal so that high scores indicated more Dominance (M = 4.31, SD = 1.55, range = 1–7; Cronbach's α = .74).

Study 2 Method

This study was conducted using Amazon Mechanical Turk (MTurk), selecting only Master Workers (i.e., very experienced workers) from the United States (N = 27; 26% male; M_{age} = 40.48). Participants were paid $1.50. After completing the APT, participants completed the TIPI and the Dominance items (M = 3.60, SD = 1.36, range = 1–5.50; Cronbach's α = .34), as well as 15 antonyms from the Verbal GRE that were taken from a GRE review book. An example is "Quixotic: slow, abstemious, pragmatic, benevolent, or grave" (M_{correct} = 10.37, SD = 3.27, range = 5–15; Cronbach's α = .78).

Study 3 Method

This was an MTurk study, but not based on Master Workers (N = 119, reduced to 118 due to deletion of a low outlier score of .29 on the APT; 49% male; M_{age} = 31.95). Participants were paid $1.50. Other variables were as follows:

1. TIPI and Dominance items as in Study 2 (Dominance scale M = 3.80, SD = 1.48, range = 1–7; Cronbach's α = .59).
2. The 24-item Personal Attributes Questionnaire (PAQ), consisting of unipolar scales for Masculinity (Agency) and Femininity (Communion) and a bipolar scale for Femininity-Masculinity (Spence & Helmreich, 1978). Sample items for the three scales, respectively, are "not at all independent . . . very independent," "not at all emotional . . . very emotional," and "never cry . . . cry very easily" (reversed).
3. The Interpersonal Reactivity Index (IRI), consisting of four scales measuring different facets of response to emotional stimuli (Davis, 1983): Fantasy, Perspective Taking, Empathic Concern, and Personal Distress. Sample items for the four scales are "I daydream and fantasize, with some regularity, about things that might happen to me" (Fantasy); "I try to look at everybody's side of a disagreement before I make a decision" (Perspective Taking); "I often have tender, concerned feelings for people less fortunate than me" (Empathic Concern); and "In emergency situations, I feel apprehensive and ill-at-ease" (Personal Distress).
4. The same GRE antonyms as in Study 2 (M = 9.86, SD = 3.23, range = 3–15; Cronbach's α = .75).
5. The nine "interest in personality" self-report items mentioned above. Items were answered on a 1–7 scale ranging from *Strongly Disagree* to *Strongly Agree* and were as follows: "I think a lot about how people differ in their personalities"; "I don't spend much time comparing people I know in terms of their personalities" (reversed); "Another person's personality matters a lot in terms of how I treat him or her"; "When I meet someone new, I immediately have an impression of their personality"; "There's no such thing as 'personality types'" (reversed); "I am generally right on target when I assess someone's personality"; "If you asked me to describe the personalities of my friends, I probably couldn't do a good job of it" (reversed); "I like to describe people's personalities when I talk about them to friends or

family"; and "A person's personality is reflected in how they act and in what they say." Cronbach's alpha for these nine items was .67, and they were averaged into an interest in personality scale (M = 5.08, SD = .72, range = 3.22–6.78).

6. Reading habits were assessed with three questions: "How much do you like to read for pleasure?" (1 = *Not much, I don't like to read*; 5 = *Extremely much, I'm an avid reader; M* = 3.92, SD = 1.03, range = 1–5); "How many hours in a typical week do you spend reading for pleasure? Please round to the nearest hour. If you read more than 30 hours per week, just put 30" (M = 8.74, SD = 6.44, range = 0–30); and "What kind of books do you tend to read?" (1 = *almost all non-fiction*, 5 = *almost all fiction*; M = 3.22, SD = 1.35, range = 1–5).

Study 4 Method

This was an MTurk study, for which participants were recruited as in Study 3 (N = 123; 33% male; M_{age} = 36.12). Participants were paid $1.50. Study 4 was the same as Study 3, with one new item added: "How many hours in a typical week do you spend reading for work? Please round to the nearest hour. If you read more than 30 hours per week, just put 30" (M = 6.75, SD = 8.07, range = 0–30). Descriptive statistics for the other three reading questions were liking to read for pleasure (M = 3.92, SD = 1.06, range = 1–5), hours spent reading for pleasure (M = 8.86, SD = 6.83, range = 0–30), and kind of books read (M = 3.33, SD = 1.29, range = 1–5). Other descriptive statistics were interest in personality scale (M = 5.16, SD = .79, range = 3.11–7.00; Cronbach's α = .80), GRE antonyms (M = 10.09, SD = 3.16, range = 2–15; Cronbach's α = .75), and Dominance scale (M = 3.84, SD = 1.40, range = 1–7; Cronbach's α = .49).

Study 5 Method

These participants were recruited from the participant pool as described above, and they participated in dyads (30 dyads plus three students who participated singly; N = 63; 49% male; M_{age} = 18.50). Participants also returned to the lab individually after 2 weeks in order to complete the APT again for test-retest purposes.

In the first session, mixed-gender dyads first completed the APT and then interacted with each other for 3 min on any topic they wanted. They then rated their own personality (TIPI; Dominance scale M = 4.32, SD = 1.28, range = 1.5–7; Cronbach's α = .56) and felt emotions (Positive and Negative Affect Schedule [PANAS]; Watson, Clark, & Tellegen, 1988), as well as judgments of their partners' personality and felt emotions using the same questionnaires. Interpersonal judgment accuracy scores were calculated for each dyad as self-other agreement profile correlations, by correlating the participants' ratings of the partner with the partner's self-ratings, across the personality items and across the emotion items (personality judgment accuracy M = .35, SD = .31, range = −.39–.71; emotion judgment accuracy M = .51, SD = .22, range = −.03–.74). For three participants who completed this first session alone, we obtained their APT score. All participants were then scheduled to return in 2 weeks. Two participants failed to return.

In the second session, participants completed the APT again. They then watched brief video clips of six job applicants whose Big Five personality traits were known (collected for a different study; Frauendorfer, Schmid Mast, Nguyen, Gatica-Perez, & Odobez, 2014) and made judgments of the applicants' personality traits using the TIPI. Accuracy scores for judging the applicants' personalities were calculated as profile correlations across rated items for each participant judging each applicant, and these were then averaged across applicants for each participant to yield one accuracy score reflecting the participant's ability to judge the personalities of the six applicants (M = .10, SD = .25, range = −.40–.79).

After this, participants completed the Adult Faces test from the Diagnostic Analysis of Nonverbal Behavior (DANVA2; Nowicki & Duke, 1994), followed by the shortened version of the Profile of Nonverbal Sensitivity (MiniPONS; Bänziger, Scherer, Hall, & Rosenthal, 2011). All of these tasks were scored to yield, respectively, accuracy in judging the Big Five from full-speech videos, accuracy in judging four basic emotions from photographs of facial expressions, and accuracy in judging situational affect from videos showing face, body, and/or content-masked speech.

Study 6 Method

This was an MTurk study, with participants recruited in the same manner as in Study 3. Four low-scoring outliers on the APT (accuracy below .40) were removed, leaving N = 404 (35% male, M_{age} = 36.21). Participants were paid $1.50. The TIPI, Dominance, interest in personality, and reading habits items were the same as in Study 4. Descriptive statistics were Dominance scale (M = 3.82, SD = 1.32, range = 1–6.5; Cronbach's α = .35), interest in personality (M = 5.23, SD = .75, range = 2.67–6.89; Cronbach's α = .73), liking reading for pleasure (M = 3.92, SD = 1.08, range = 1–5), hours reading for pleasure (M = 9.75, SD = 7.44, range = 0–30), hours reading for work (M = 8.56, SD = 8.92, range = 0–30), and kind of books read (M = 3.18, SD = 1.24, range = 1–5).

In this study, an additional task was added to measure participants' interest in personality—their spontaneous use of personality descriptions when describing people. Two person descriptions were requested, which participants typed into a box on the computer screen: One was a self-description, and the other was a description of "Mark, a 35 year old married heterosexual man, who works in middle management in a small company." Instructions made no specific mention of describing personality traits, so participants would be free to include them or not. Each description was rated by a trained coder for the extent of personality description, with "personality description" defined as the degree to which participants described their own attributes in the self-description, or the attributes of "Mark," by using specific personality trait terms such as *introverted*, *agreeable*, or *outgoing*, or by describing behavioral styles that strongly imply personality traits (e.g., I love hanging out with friends, Mark worries a lot about his social adequacy). Personality was defined to exclude physical descriptions (e.g., height) and sociodemographic characteristics (e.g., place of birth/ nationality, ethnicity, occupation, social class, religion, political affiliation, education, or marital status). An example of a low-rated personality description is "Mark likes to go to the store on his lunch break. He buys a tuna sandwich and a Coke and sits in the park on the bench. He feeds the birds his leftover pieces of bread." An example of a high-rated description is as follows: "Mark is just an ordinary man, with ordinary goals and conventional wants. He sticks to what he knows, and what he thinks is the "straight and narrow." He is not a risk-taker, and instead prefers the mundane routine of his everyday life."

Also, vividness of both the self-description and of "Mark" was rated by the same coder and was defined as the extent to which people made their writing unique and individualized using specific details.

In addition, participants were asked to write briefly about a forest; this was done in order to rate and control for general vividness of writing style. Vividness was rated as described in the previous paragraph. All of the ratings of these descriptions were made on a 9-point rating scale ranging from *low* to *high*. Interrater reliability between the main coder and an independent coder, based on 39 writing excerpts, was $r = .68$ for own personality description and $r = .54$ for vivid description of self, $r = .54$ for personality description of Mark and $r = .77$ for vivid description of "Mark," and $r = .79$ for the vividness of the forest excerpt. Descriptive statistics for all available writing excerpts were as follows: personality description ($M = 4.19$, $SD = 2.47$) and vivid description ($M = 3.46$, $SD = 2.08$) of self, personality description ($M = 4.47$, $SD = 1.62$) and vivid description ($M = 4.28$, $SD = 1.65$) of "Mark," and vividness ($M = 4.05$, $SD = 1.52$) of the forest excerpt.

Statistical Analysis

Within studies, ordinary descriptive and inferential statistics were performed, with all p-values based on two-tailed tests. For analysis of results across studies, the Comprehensive Meta-Analysis Software (CMA; Borenstein, Hedges, Higgins, & Rothstein, 2005) was used, also with two-tailed tests. This software yielded both random- and fixed-effects estimates of overall effect size (i.e., mean correlation), along with associated tests of whether the mean correlation was above zero. The fixed- and random-effects results were virtually identical due to a high degree of homogeneity in the distributions of correlations. We present the random-effects mean correlations only, which can be thought of as the mean correlations unweighted by sample size. A significant random-effects model yields generalization to new study designs rather than to the same study designs with new participants in them, as that would be the more limited generalization based on a fixed-effects model (Lipsey & Wilson, 2001).

All correlations were normalized using the Fisher's z transformation before meta-analytic calculations were performed and then returned to the Pearson correlation metric for presentation.

Table 2 Descriptive Statistics for Accuracy in Assessing Personality From Text

Study	*N*	*M* (*SD*)	Range	Cronbach's α (36 items)
1	45	.79 (.11)	.44–1.00	.69
2	27	.81 (.08)	.61–.94	.45
3	118	.81 (.09)	.56–.94	.50
4	123	.81 (.08)	.50–1.00	.49
5	63	.83 (.07)	.69–.97	.22
6	404	.80 (.11)	.42–1.00	.61

RESULTS

Descriptive Statistics for the APT

Table 2 shows descriptive data for the test of judging personality from text. Overall accuracy was approximately .80 (i.e., 80% of the items were answered correctly), with a broad range of scores within each study. If participants guessed on every item, the mean would be .50, and this is the value against which mean accuracy was tested. Accuracy was significantly above chance in every study, for the entire group and for the male and female subsamples analyzed separately ($p \leq .001$). The mean, standard deviation, and range were very similar across the studies. A few participants scored somewhat below chance, but they were maintained in their respective samples (extremely low scores were removed as described in earlier sections).

Cronbach's alphas, also shown in Table 2, were generally of modest magnitude with the exception of Study 5, where average accuracy was comparable to the other studies but reliability was considerably lower. Retest reliability after a 2-week interval was $r(61) = .52, p < .001$.

Gender

Table 3 shows that accuracy in judging personality from text correlated positively with gender, meaning that women scored higher than men, consistent with the general literature on accurate interpersonal judgment. The highly significant p-value associated with the random-effects mean correlation indicates that the gender difference was significantly above zero for the studies as a group. Across the six studies, the average male score was 79.17 (range = .78–.82) and the average female score was 81.50 (range = .80–.83), a small but very reliable difference.

Personality Traits

Table 3 shows the correlations of accuracy in judging personality from text with participants' own Big Five personality traits

Table 3 Correlations of Accuracy With Gender, Big Five Traits, and Dominance

Study[a]	Gender	E	A	C	N	O	Dom.
1	.12	−.08	.03	.18	−.11	−.05	−.14
2	.22	−.34†	.46*	.17	−.11	.27	−.44*
3	.07	−.04	.18*	.15†	.01	.02	−.02
4	.17†	−.18*	.23**	.13	.05	.02	−.11
5	.10	.22†	.12	.09	−.08	.14	−.08
6	.18***	.02	.20***	.16***	−.15**	.05	−.10*
M r	.15***	−.04	.20***	.15***	−.09*	.05	−.10**

Note. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness to Experience; Dom. = Dominance; *M r* = random-effects mean correlation. Gender was coded 0 = male, 1 = female; positive correlations indicate that women scored higher than men.

[a] Some sample sizes are slightly reduced due to occasional missing observations.

$^{\dagger}p \leq .10$. $^{*}p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

Table 4 Correlations of Accuracy With Masculinity, Femininity, Femininity-Masculinity, and the Interpersonal Reactivity Index Scales

Study	Masc.	Fem.	Fem.-Masc.	Fant.	Persp.	Emp.	Dist.
3	.04	.22*	−.07	.21*	.18*	.21*	.04
4	−.05	.28**	−.14	−.04	.07	.15†	.00
M r	−.01	.25***	−.10	.09	.12†	.18**	.02

Note. Masc. = Masculinity; Fem. = Femininity; Fem.-Masc. = Femininity-Masculinity (bipolar scale); Fant. = Fantasy; Persp. = Perspective Taking; Emp. = Empathic Concern; Dist. = Personal Distress; *M r* = random-effects mean correlation.

$^{\dagger}p \leq .10$. $^{*}p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

and their self-rated Dominance. The correlations with Extraversion and Openness to Experience were negligible and nonsignificant. However, the meta-analysis revealed that higher APT scores were significantly associated with more Agreeableness, more Conscientiousness, less Neuroticism, and less Dominant personality. Gender was not a confound of these relations according to partial correlations; only in Study 2 did one of the correlations change when gender was controlled for, and this was for Openness, where the correlation became substantially stronger (positively) when gender was controlled.

Table 4 shows correlations with the remaining personality variables. Empathic Concern and Femininity were significantly positively correlated with the APT. Perspective Taking was also positively related, though marginally significantly. These effects were not confounded by participant gender according to partial correlations.

Cognitive Ability

Twenty-nine of the 45 participants in Study 1 reported their Verbal and Quantitative SAT scores. SAT scores were uncorrelated with the APT, $r = -.08$ for Verbal and –.05 for Quantitative. Participants in Studies 2–4 were given the GRE antonyms test described above. These correlations with accuracy were, respectively, .18, .21 ($p < .05$), and .16 ($p < .10$). Together, the GRE correlations were significantly greater than zero according to the meta-analysis (mean correlation = .18, $p < .01$). These correlations were not confounded with gender according to partial correlations. Thus, the APT had only modest relations with verbal ability (and none with self-reported SAT scores).

Interest in Personality

Results for the nine-item interest in personality scale are shown in Table 5. Accuracy in judging personality from text was positively and highly significantly correlated with this scale across the three studies that included it, reflecting high-scoring participants' belief that personality matters in daily life and their self-report of attentiveness to personality. The most predictive individual items were "I think a lot about how people differ in their personalities" (significant in two studies); "A person's personality is reflected in how they act and in what they say" (significant in two studies); "When I meet someone new, I immediately have an impression of their personality" (significant in three studies); and "I don't spend much time comparing people I know in terms of their personalities" (reversed) and "There's no such thing as 'personality types'" (reversed; significant in one study each).

The second way that interest in personality was measured was indirect, by asking participants to write a short description of themselves and of a hypothetical person, "Mark," and by then coding their descriptions for personality content and for general stylistic vividness. This analysis yielded only suggestive evidence in favor of the hypothesis that accuracy on our test would predict more personality content in their descriptions. For the self-description, there was no relation with the APT, $r = .03$. For the "Mark" description, there was a marginally significant correlation of .08, $p = .10$, showing that more spontaneous use of personality description was associated with higher accuracy. This correlation was not appreciably changed when controlling for either gender or general vividness of the participant's writing. Thus, there was some, though limited, support for the hypothesis that skill in judging personality from text would correlate with the tendency to refer to personality more when writing about someone else.

Reading Habits

Correlations of accuracy with the reading habits items are shown in Table 5. Enjoying reading for pleasure, preferring to read fiction over nonfiction, and the difference between how much participants read for pleasure versus work were all significantly positively related to accuracy in judging personality from text according to the meta-analysis. These effects were not confounded by gender.

Other Measures of Accurate Interpersonal Perception

The five other measures of interpersonal accuracy in Study 5 did not correlate significantly with the APT. Correlations ranged

Table 5 Correlations of Accuracy With Reading Habits and the Interest in Personality Scale

Study	Enjoy Reading for Pleasure	Read for Pleasure Minus Read for Work	Kind of Books Read	Interest in Personality
3	.17†	–	−.02	.29**
4	.14	.19*	.07	.19*
6	.10*	.16**	.13**	.14**
M r	.12**	.17***	.09*	.18***

Note. — = not applicable; *M r* = random-effects mean correlation. For kind of books read, high values indicate more fiction than nonfiction.
†$p \leq .10$. *$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

from –.07 to .10, with a mean of .02 (median $r = .05$). The other five measures did not correlate well with each other either (median $r = .06$). Thus, in this study, there was no evidence of a general accuracy factor.

DISCUSSION

In six studies, correlates of individual accuracy in judging the Big Five personality traits from writing samples were examined. In the sections that follow, the results are discussed in relation to the broader literature that has tested interpersonal accuracy using audiovisual stimuli. Note must be taken, however, that most of that literature is not about judging personality but about judging affective states. Previous known studies on judging personality are mentioned where applicable.

Research Questions Previously Addressed in Audiovisual Interpersonal Accuracy Research

Women were better at judging personality from text than men. Many studies using audiovisual tests have also found this gender difference, which, though relatively small in absolute terms both in the present studies and in the audiovisual literature, is still highly reliable for tests that involve the judgment of affective cues (cf. Hall, 1984; Hall & Gunnery, 2013). The magnitude of the difference on our test (.15 when stated as a correlation effect size, and .30 when stated as Cohen's *d*) is generally comparable to the average effect size from several meta-analyses comparing adult males' and females' scores on audiovisual tests of interpersonal accuracy (Hall, 1984; Kirkland et al., 2013). This gender difference has also appeared in studies of accuracy in judging personality (Letzring, 2008; Lippa & Dietz, 2000; Vogt & Colvin, 2003).

Participants scoring higher on Empathic Concern (and, to some extent, Perspective Taking) scored higher on judging personality from text, independent of gender. Hall et al. (2009) found similar results in their meta-analysis of the broader audiovisual literature, and Letzring (2008) found, for accuracy in judging personality specifically, that greater accuracy was associated with sympathy and consideration as personality traits, and with behavioral warmth, eye contact, and enjoyment during a videotaped interaction, all of which suggest, along with the present results, a communal orientation. In the present research, the IRI scales for Fantasy and Personal Distress showed equivocal relations with accuracy, perhaps indicating that these two scales are not capturing the other-oriented, prosocial aspect of empathy (Davis, 1983; Hojat, Mangione, Kane, & Gonnela, 2005).

In terms of participants' own Big Five scores, the present studies agreed with the broader audiovisual literature in showing that more accuracy was associated with less Neuroticism (cf. Hall et al., 2009), and Letzring (2008) found this for judging personality in particular. Letzring (2008) also found, as we did, that more Agreeable people were better at judging personality. However, two of participants' Big Five traits, Extraversion and Openness, did not correlate on average with our test, unlike in previous studies on interpersonal accuracy based on audiovisual tests (Hall et al., 2009) and on judging personality in particular (correlation with participants' Openness; Christiansen et al., 2005).

The finding that more Dominant individuals scored lower on judging personality from text than less Dominant individuals is consistent with some of the research using audiovisual tests (e.g., Moeller, Ewing Lee, & Robinson, 2011), and it was also found by Letzring (2008) for judging personality in particular. However, it is discrepant from some other studies within the broader literature, which found the opposite (e.g., three studies in Schmid Mast, Jonas, & Hall, 2009). A meta-analysis (Hall et al., 2014) concluded that whether this relation is positive or negative depended on the nature of Dominance as captured in the personality scale. With personality scales that define Dominance as more egoistic and controlling, the relation was negative, showing that more Dominant individuals were less interpersonally accurate; however, when Dominance was defined as more prosocial and socially responsible, the effect was positive, showing that more Dominant individuals were more interpersonally accurate. The two Dominance items used in the present studies could be considered in the same light: "Dominant, controlling" can be considered to be the more egoistic kind of Dominance, and "a follower more than a leader, nonassertive" (reversed) can be considered to be the more prosocial kind. When these two items were examined separately in the present studies, the overall correlation for "dominant, controlling" was –.16, whereas the corresponding correlation for "a follower more than a leader, nonassertive" (*after* reversal so that both items indicate high Dominance) was –.08. Thus, although the latter correlation was not positive, as found in the Hall et al. (2014) meta-analysis, it was weaker than that found for the "dominant, controlling" item.

The finding that more Feminine (communal) individuals, controlling for gender, were better at judging personality from text is consistent with both the broader audiovisual literature (Hall et al., 2009) and with Vogt and Colvin's (2003) study of judging personality in particular. Our null finding for Masculinity was also consistent with the Hall et al. (2009) meta-analysis.

SAT scores were not related to accuracy of judging personality from text, whereas a test of vocabulary knowledge was significantly though not strongly correlated with accuracy. Together, these results support the discriminant validity of the present method of measuring accurate personality judgment, consistent with the meta-analysis of Murphy and Hall (2011) on the broader audiovisual literature; in fact, the correlation with GRE antonyms in the present studies was nearly identical to the overall result for intelligence tests in Murphy and Hall (2011). Also, Lippa and Dietz's (2000) study on accuracy of judging personality in particular found only weak correlations with cognitive ability.

The present test of judging personality from text was not at all correlated with several tests of interpersonal accuracy based on judging audiovisual stimuli. In this regard, the present test is similar to audiovisual tests in that they often have only meager correlations with each other (Hall, 2001; meta-analysis of Boone, Schlegel, & Hall, 2015). In particular, the meta-analysis showed that accurate personality judgment correlated negligibly with accuracy in judging other constructs (e.g., emotion).

Novel Questions

Finally, the present research examined some relations that have not been examined in the literature on accuracy using audiovisual stimuli. Greater accuracy in judging personality from text was positively associated with believing that personality is real and important and with self-reported attentiveness to others' personalities. This suggests that individual differences in these tendencies may be precursors of accuracy in judging personality, due to an accumulation of practice in thinking about and observing others' personalities.

The prediction that accuracy in judging personality from text would be correlated with how much participants used personality descriptions in their own writing was only weakly upheld, yet it is still worth pursuing in future research. We allowed participants to write only two or three sentences, whereas a longer writing opportunity or a spontaneous speaking opportunity might be a better basis for detecting the effect we were hoping for.

Accuracy in judging personality from text was correlated with a person's reading habits and preferences, such that people who read relatively more for pleasure than for work, enjoyed reading for pleasure more, and preferred reading fiction over nonfiction had higher judgment accuracy scores. This suggests another possible precursor of this ability: Reading a great deal of fiction is likely to expose people to countless descriptions of people's personalities as well as to the expression of personality through the characters' dialogue. Kidd and Castano (2013) found that reading literary fiction as opposed to nonfiction or popular fiction improved participants' performance on the "Reading the Mind in the Eyes" Test (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) and the DANVA (Nowicki & Duke, 1994)—both tests of accurate judgment of the meanings of nonverbal facial cues—and those authors suggested that this may be due to the rich personalities of the characters and the need to interpret their thoughts and feelings.

LIMITATIONS

Although the instrument we used for judging personality had many promising correlates and had reasonably good retest reliability, its internal consistency was modest. Modest internal consistency is not at all unusual for tests of accurate interpersonal perception, as many investigators have observed (e.g., Bänziger et al., 2011; Hall, 2001). A meta-analysis has found that accuracies for judging different personality traits correlate negligibly with each other (Hall & Boone, 2014), possibly helping to account for the only modest internal consistency of the instrument we created. Nevertheless, the present research is a promising start on measuring accuracy of judging personality from text and its correlates.

For researchers who might use, or improve on, the methodology used here, several limitations can be listed. Our method of enhancing the text excerpts to make them more strongly reflective of the original writer's actual personality takes away some of the authenticity of the communication. However, the goal was not to capture the original writers' "true" personalities in the excerpts but to measure perceivers' ability to pick up on cues to personality embedded in the excerpts. This is a subtle but important distinction. We did not, and do not, claim that the excerpts represent their authors' personalities, precisely because we did alter the excerpts to make them slightly more judgeable. Our claim is rather that the test measures whether the test taker can correctly pick up on cues to personality traits that are embedded in textual material. To draw an analogy from the emotion judgment field, many of the valid tests in use contain stimuli that are posed by actors; the researchers do not claim that the test items represent the actors' true emotions, only that the test items contain cues that signal certain emotions (according to criteria). Similarly, if a personality researcher hired actors to use valid cues to portray different traits in video clips and turned those clips into a personality judgment test, the claim would not be that the clips represent those actors' real personalities, only that the cues they enact represent how people with those traits behave. In both of these examples, the test is not robbed of its validity by virtue of having posed cues in it. The test of validity is whether the test can pass the standards of construct validation. The same situation applies to the APT.

Although there would be value in developing accuracy instruments based on entirely spontaneous, nonselected stimuli, such tests would not necessarily be more valid for measuring individual differences in accuracy; that is an empirical question that only a program of construct validation could answer. It may be relevant, however, that in the audiovisual interpersonal accuracy literature, tests based on spontaneous cues (e.g., Buck, 1976; Costanzo & Archer, 1989) appear not to have greater validity than tests based on posed cues (e.g., Bänziger et al., 2011; Nowicki & Duke, 1994).

Another potential criticism is that the total number of target people represented in the test was only nine. It is important to keep in mind, however, that the goal was not to represent or generalize across many targets; if that had been the goal, a very different approach to test construction would have been used. The goal was to create a test that could distinguish among perceivers. For that purpose, there is no theoretical or psychometric necessity to represent many targets. Indeed, the well-established and validated PONS test that measures accuracy of decoding situationally based nonverbal cues of affect has only one target person (Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979).

It must also be acknowledged that the self-report scale used to classify the writers of the text excerpts as high or low on the Big Five traits (the TIPI) was not as psychometrically sound nor likely as valid as a longer personality inventory would be, and also that an even more solid personality criterion would include gathering converging reports from friends or family members. However, the TIPI does correlate well with longer measures of the Big Five Traits (Gosling et al., 2003).

CONCLUSION

In the present research, individual differences in accuracy of judging personality from text excerpts were measured. This skill was shown to have a broad range of correlates, most of which supported the construct validity of the measurement and some of which suggested possible paths to gaining such accuracy. To date, there is little known about precursors of interpersonal accuracy—in any content domain for any kind of stimuli. However, many correlates from the audiovisual test literature are candidates as causal precursors, such as being a dancer or an athlete, early socialization in the family, having a preverbal toddler, learning American Sign Language, and having a range of personality attributes and attitudes (Hall et al., 2009). In addition, formal training in accurate interpersonal perception is an established precursor, although it is not known whether these effects last beyond an immediate testing context and they have not been aimed much at personality judgment accuracy (Blanch-Hartigan, Andrzejewski, & Hill, 2012).

The present research begins to fill a gap in the personality judgment field. Most personality judgment research involves people judging the personality of their live interaction partner(s) (e.g., Letzring, 2008). Although there are many good reasons why researchers choose this face-to-face paradigm, it indicates that the development of standardized stimuli and validated tests of accuracy in personality judgment has not been a goal of personality judgment research. Furthermore, although researchers of personality judgment do use videotaped behavior or photographs as stimuli that could, in principle, be developed into standardized instruments (e.g., Letzring, 2015; Lippa & Dietz, 2000; Vazire, Naumann, Rentfrow, & Gosling, 2008), it appears researchers have not taken that next step. If researchers who study the communication of personality would make standardized tests, many additional research questions could be asked and in a more economical and efficient way. Additionally, the use of standardized tests serves the field not only because more studies could be done but also because comparability across studies is enhanced, promoting better accumulation of knowledge.

The present research dealt only with judging personality using written text excerpts as the stimuli. Obviously, people use others' words to make countless other inferences, including what emotional state they are in, whether the person has high or low status, what sexual orientation they have, how intelligent they are, and whether they are romantically interested, to name just a few. "Reading" the nuances of linguistic style and word choice—reading between the lines, as the phrase goes—should become a standard topic for researchers interested in the accuracy with which people perceive their social world. Important though nonverbal cues are, they are only part of how people communicate.

Declaration of Conflicting Interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Note

[1] In all studies, we report the mean, standard deviation, range, and Cronbach's alpha for all measures not previously developed by other researchers, with the exception of the dyadic accuracy scores from Study 5: Because these scores were profile correlations across rated items, Cronbach's alpha cannot be calculated.

References

Back, M. D., Schmukle, S. C., & Egloff, B. (2008). How extraverted is honey.bunny77@hotmail.de? Inferring personality from e-mail addresses. *Journal of Research in Personality*, **42**, 1116–1122.

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**, 189–204.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, **42**, 241–251.

Beukeboom, C. J., Tanis, M., & Vermeulen, I. E. (2013). The language of extraversion: Extraverted people talk more abstractly, introverts are more concrete. *Journal of Language and Social Psychology*, **32**, 191–201.

Blanch-Hartigan, D., Andrzejewski, S. A., & Hill, K. M. (2012). The effectiveness of training to improve person perception: A meta-analysis. *Basic and Applied Social Psychology*, **34**, 483–498.

Boone, R. T., Schlegel, K., & Hall, J. A. (2015, February). *Meta-analysis of correlations between tests of interpersonal accuracy: Drawing the map of the interpersonal skill domain*. Talk given at annual meeting of the Society for Personality and Social Psychology, Long Beach, CA.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis* (2nd ed.). Englewood, NJ: Biostat.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Buck, R. (1976). A test of nonverbal receiving ability: Preliminary studies. *Human Communication Research*, **2**, 162–171.

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**, 123–149.

Cole, J. S., & Gonyea, R. M. (2010). Accuracy of self-reported SAT and ACT test scores: Implications for research. *Research in Higher Education*, **51**, 305–319.

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

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, **44**, 113–126.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, **128**, 203–235.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, **94**, 334–346.

Frauendorfer, D., Schmid Mast, M., Nguyen, N., Gatica-Perez, D., & Odobez, J. M. (2014). *Can recruiters accurately predict applicants' job performance based on thin-slices of applicant job interview behavior? Yes, they can!* Manuscript in preparation.

Funder, D. C. (2001). Accuracy in personality judgment: Research and theory concerning an obvious question. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace: Decade of behavior* (pp. 121–140). Washington, DC: American Psychological Association.

Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology*, **77**, 746–761.

Gifford, R., & Hine, D. W. (1994). The role of verbal behavior in the encoding and decoding of interpersonal dispositions. *Journal of Research in Personality*, **28**, 115–132.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, **37**, 504–528.

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style*. Baltimore: Johns Hopkins University Press.

Hall, J. A. (2001). The PONS test and the psychometric approach to measuring interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 143–160). Mahwah, NJ: Erlbaum.

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior*, **33**, 149–180.

Hall, J. A., & Bernieri, F. J. (Eds.). (2001). *Interpersonal sensitivity: Theory and measurement*. Mahwah, NJ: Erlbaum.

Hall, J. A., Bernieri, F. J., & Carney, D. R. (2005). Nonverbal behavior and interpersonal sensitivity. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 237–281). Oxford: Oxford University Press.

Hall, J. A., & Gunnery, S. D. (2013). Gender differences in nonverbal communication. In J. A. Hall & M. L. Knapp (Eds.), *Nonverbal communication: Vol. 2. Handbooks of communication science* (pp. 639–669). Berlin: deGruyter Mouton.

Hall, J. A., & Schmid Mast, M. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion*, **7**, 438–446.

Hall, J. A., Schmid Mast, M., & Latu, I. M. (2014). The vertical dimension of social relations and accurate interpersonal perception: A meta-analysis. *Journal of Nonverbal Behavior.*

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality*, **43**, 524–527.

Hojat, M., Mangione, S., Kane, G. C., & Gonnela, J. (2005). Relationships between scores of the Jefferson Scale of Physician Empathy and the Interpersonal Reactivity Index. *Medical Teacher*, **7**, 625–628.

Holleran, S. E., & Mehl, M. R. (2008). Let me read your mind: Personality judgments based on a person's natural stream of thought. *Journal of Research in Personality*, **42**, 747–754.

Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality*, **45**, 92–99.

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, **342**, 377–380.

Kirkland, R. A., Peterson, E., Baker, C. A., Miller, S., & Pulos, S. (2013). Meta-analysis reveals adult female superiority in "Reading the Mind in the Eyes" Test. *North American Journal of Psychology*, **15**, 121–146.

Küfner, A. C. P., Back, M. D., Nestler, S., & Egloff, B. (2010). Tell me a story and I will tell you who you are! Lens model analyses of personality and creative writing. *Journal of Research in Personality*, **44**, 427–435.

Letzring, T. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality*, **42**, 914–932.

Letzring, T. D. (2015). Observer judgmental accuracy of personality: Benefits related to being a good (normative) judge. *Journal of Research in Personality*, **54**, 51–60.

Lippa, R. A., & Dietz, J. K. (2000). The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior*, **24**, 25–43.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Marsh, A. A., & Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: A meta-analysis. *Neuroscience and Biobehavioral Reviews*, **32**, 454–465.

Moeller, S. K., Ewing Lee, E. A., & Robinson, M. D. (2011). You never think about my feelings: Interpersonal dominance as a predictor of emotion decoding accuracy. *Emotion*, **11**, 816–824.

Murphy, N. A., & Hall, J. A. (2011). Intelligence and nonverbal sensitivity: A meta-analysis. *Intelligence*, **39**, 54–63.

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**, 9–34.

Nowicki, S., & Duke, M. (2013). Accuracy in interpreting nonverbal cues. In J. A. Hall & M. L. Knapp (Eds.), *Handbooks of communication science: Vol. 2. Nonverbal communication* (pp. 441–470). Berlin: DeGruyter Mouton.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, **54**, 547–577.

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

Savla, G. N., Vella, L., Armstrong, C. C., Penn, D. L., & Twamley, E. W. (2013). Deficits in domains of social cognition in schizophrenia: A meta-analysis of the empirical evidence. *Schizophrenia Bulletin*, **39**, 979–999.

Schmid Mast, M., Jonas, K., & Hall, J. A. (2009). Give a person power and he or she will show interpersonal sensitivity: The phenomenon and its why and when. *Journal of Personality and Social Psychology*, **97**, 835–850.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, **8**(9): e73791.

Shanley, L. A., Walker, R. E., & Foley, J. M. (1971). Social intelligence: A concept in search of data. *Psychological Reports*, **29**, 1123–1132.

Spence, J. T., & Helmreich, R. L. (1978). *Masculinity and femininity: Their psychological dimensions, correlates and antecedents*. Austin: University of Texas Press.

Stopfer, J. M., Egloff, B., Nestler, S., & Back, M. D. (2014). Personality expression and impression formation in online social networks: An integrative approach to understanding the processes of accuracy, impression management, and meta-accuracy. *European Journal of Personality*, **28**, 73–94.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, **29**, 24–54.

Thompson, A. E., & Voyer, D. (2014). Sex differences in the ability to recognize non-verbal displays of emotion: A meta-analysis. *Cognition and Emotion*, **28**, 1164–1195.

Tskhay, K. O., & Rule, N. O. (2014). Perceptions of personality in text-based media and OSN: A meta-analysis. *Journal of Research in Personality*, **49**, 25–30.

Vazire, S., Naumann, L. P., Rentfrow, P. J., & Gosling, S. D. (2008). Portrait of a narcissist: Manifestations of narcissism in physical appearance. *Journal of Research in Personality*, **42**, 1439–1447.

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

Walker, R. E., & Foley, J. M. (1973). Social intelligence: Its history and measurement. *Psychological Reports*, **33**, 839–864.

Wall, H. J., Taylor, P. J., Dixon, J., Conchie, S. M., & Ellis, D. A. (2013). Rich contexts do not always enrich the accuracy of personality judgments. *Journal of Experimental Social Psychology*, **49**, 1190–1195.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, **54**, 1063–1070.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, **44**, 363–373.

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion*, **9**, 478–487.