MANUAL

Assessing Personality from Text (APT):
A Test of Judging Big Five Traits from Text Excerpts

Judith A. Hall    Jin X. Goh
Northeastern University

Marianne Schmid Mast
University of Lausanne

Christian Hagedorn
Northeastern University


Please address correspondence to Judith A. Hall, Dept. of Psychology, Northeastern University, Boston, MA 02115, j.hall@neu.edu.

July 2015
MANUAL

Assessing Personality from Text (APT):
A Test of Judging Big Five Traits from Text Excerpts

Overview of Test and its Validation

The APT (Assessing Personality from Text) was developed for measuring accurate identification of high versus low Big Five traits from short text excerpts. It can be used for measuring group and individual differences in this kind of accuracy. The test contains 36 excerpts, each with a binary response format (high vs. low on the criterion trait). 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.

Background

The present research can be compared to the large literature on correlates of individual differences in accuracy based on audiovisual cues such as in photographs, vocal recordings, or videotape. Although there is an abundance of such research (for a book devoted to this topic, see Hall, Schmid Mast, & West, in press), the great majority of studies of interpersonal accuracy are on judging affective states rather than personality, and none of it uses text as stimuli to be judged. The APT represents a first systematic effort to measure individual differences in

Method

Measuring Accuracy of Judging Personality from Text

Because an instrument did not exist, we created one, which is hereafter referred to as the APT, for Assessing Personality from Text. The goals for this test were (1) that it would have enough items to reflect a reasonable array of targets and excerpts while representing all Big Five traits, (2) that it would be short enough for efficient administration, and (3) that it would 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 (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 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, while 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 their 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 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 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: (1) Write a message to your mother. Talk about anything you want. (2) Write about your family pet, or the pet of a friend or neighbor. Write anything you want on this topic. (3) Write about your sister or brother, or a relative who is in your generation (such as a cousin). Describe that person. (4) Write about something scary that happened to you. (5) Write a message of complaint to a company about their product or service, such as you might in an email to their complaint department. (6) Write about a tough semester or a tough exam you had to prepare for. (7) Think of a sport you like and explain why you like it. (8) Write a thank-you note to a professor for writing letters of recommendation for you. (9) Write a short telephone dialogue between two friends who are chatting about what to do this weekend. (10) 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., Hirsch & 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 comprising the final test. Between seven and eight items represented each of the Big Five traits, approximately balanced for high or low level of the trait.

**Study 1 Method**

The participants in this study were described above ($N = 45$). Participants came singly or in pairs to the laboratory and completed the APT and all other measures individually on computers. Other measures were:

1. Self-reported Verbal and Quantitative SAT scores.
2. The TIPI (Gosling et al., 2003). The 10 items (each on 1-7 scale) were scored after appropriate reversals so that high scores indicated extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience.
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 alpha = .74).

**Study 2 Method**

This study was conducted on 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$; Cronbach’s alpha = .34), as well as 15 antonyms from the Verbal GRE that were taken from a GRE review book. An example is *Quixotic: slow, abstemious, pragmatic, benevolent, or grave* ($M$ correct = 10.37, $SD = 3.27$, range = 5-15; Cronbach’s alpha = .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:

1. TIPI and dominance items as in Study 2 (dominance scale $M = 3.80, SD = 1.48$, range = 1-7; Cronbach’s alpha = .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.*

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 alpha = .75).

5. The nine “interest in personality” self-report items mentioned above. Items were answered on a 1-7 scale from Strongly Disagree to Strongly Agree and were: *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 alpha = .80); GRE antonyms (M = 10.09, SD = 3.16, range = 2-15; Cronbach’s alpha = .75); and dominance scale (M = 3.84, SD = 1.40, range = 1-7; Cronbach’s alpha = .49).

Study 5 Method

These participants were recruited from the Participant Pool as described above, and they participated in dyads (30 dyads plus 3 students who participated singly; N = 63, 49% male, M age = 18.50). Participants also returned to the lab individually after two weeks in order to complete the APT again for test-retest purposes.

In session one, 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 alpha = .56) and felt emotions (Positive and Negative Affect Schedule or 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). Three participants who completed the first session alone followed the same procedure as in Study 4. All participants were then scheduled to return in two 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’ personality 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 did 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 Study 3. Four low-scoring outliers (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 alpha = .35); interest in personality ($M = 5.23$, $SD = .75$, range = 2.67-6.89; Cronbach’s alpha = .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 traits 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, or 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 was:

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 was:

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.

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 from low to high. Interrater reliability between the main coder and an independent coder, based on 39 writing excerpts, was $r = .68$ for personality description and $r = .54$ for vivid description of self, $r = .54$ for personality description 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: 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, 2003). All correlations were normalized using the Fisher-z transformation before meta-analytic calculations were performed and then returned to the Pearson correlation metric for presentation.

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 two-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 and their self-rated dominance. The correlations with extraversion and openness to experience were negligible and non-significant. 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 non-fiction, 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 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.

**References**


Footnotes

1. In all studies, we report 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.
Table 1. *Sample Items for Assessing Personality from Text (APT)*

**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.

---

Table 2. *Descriptive Statistics for Accuracy in Assessing Personality from Text*

<table>
<thead>
<tr>
<th>Study</th>
<th>N</th>
<th>M (SD)</th>
<th>Range</th>
<th>Cronbach’s alpha (36 items)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>.79 (.11)</td>
<td>.44-1.00</td>
<td>.69</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>.81 (.08)</td>
<td>.61-.94</td>
<td>.45</td>
</tr>
<tr>
<td>3</td>
<td>118</td>
<td>.81 (.09)</td>
<td>.56-.94</td>
<td>.50</td>
</tr>
<tr>
<td>4</td>
<td>123</td>
<td>.81 (.08)</td>
<td>.50-1.00</td>
<td>.49</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>.83 (.07)</td>
<td>.69-.97</td>
<td>.22</td>
</tr>
<tr>
<td>6</td>
<td>404</td>
<td>.80 (.11)</td>
<td>.42-1.00</td>
<td>.61</td>
</tr>
</tbody>
</table>
Table 3. Correlations of Accuracy with Gender, Big Five Traits, and Dominance

<table>
<thead>
<tr>
<th>Study</th>
<th>Gender</th>
<th>Extrav</th>
<th>Agree</th>
<th>Consc</th>
<th>Neur</th>
<th>Open</th>
<th>Dom</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.12</td>
<td>-.08</td>
<td>.03</td>
<td>.18</td>
<td>-.11</td>
<td>-.05</td>
<td>-.14</td>
</tr>
<tr>
<td>2</td>
<td>.22</td>
<td>-.34+</td>
<td>.46*</td>
<td>.17</td>
<td>-.11</td>
<td>.27</td>
<td>-.44*</td>
</tr>
<tr>
<td>3</td>
<td>.07</td>
<td>-.04</td>
<td>.18*</td>
<td>.15+</td>
<td>.01</td>
<td>.02</td>
<td>-.02</td>
</tr>
<tr>
<td>4</td>
<td>.17+</td>
<td>-.18*</td>
<td>.23**</td>
<td>.13</td>
<td>.05</td>
<td>.02</td>
<td>-.11</td>
</tr>
<tr>
<td>5</td>
<td>.10</td>
<td>.22+</td>
<td>.12</td>
<td>.09</td>
<td>-.08</td>
<td>.14</td>
<td>-.08</td>
</tr>
<tr>
<td>6</td>
<td>.18***</td>
<td>.02</td>
<td>.20***</td>
<td>.16***</td>
<td>-.15*</td>
<td>.05</td>
<td>-.10*</td>
</tr>
<tr>
<td>Mr</td>
<td>.15***</td>
<td>-.04</td>
<td>.20***</td>
<td>.15***</td>
<td>-.09*</td>
<td>.05</td>
<td>-.10**</td>
</tr>
</tbody>
</table>

Note: Extrav = extraversion, Agree = agreeableness, Consc = conscientiousness, Neur = neuroticism, Open = openness to experience, Dom = dominance, Mr = random effects mean correlation. Gender was coded 0 = male, 1 = female; positive correlations indicate that women scored higher than men.

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

+p ≤ .10  * p ≤ .05  ** p ≤ .01  *** p ≤ .001

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Masc</th>
<th>Fem</th>
<th>Fem-Masc</th>
<th>Fant</th>
<th>Persp</th>
<th>Emp</th>
<th>Dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.04</td>
<td>.22*</td>
<td>-.07</td>
<td>.21*</td>
<td>.18*</td>
<td>.21*</td>
<td>.04</td>
</tr>
<tr>
<td>4</td>
<td>-.05</td>
<td>.28**</td>
<td>-.14</td>
<td>-.04</td>
<td>.07</td>
<td>.15+</td>
<td>.00</td>
</tr>
<tr>
<td>Mr</td>
<td>-.01</td>
<td>.25***</td>
<td>-.10</td>
<td>.09</td>
<td>.12+</td>
<td>.18**</td>
<td>.02</td>
</tr>
</tbody>
</table>

Note: Masc = masculinity, Fem = femininity, Fem-Masc = femininity-masculinity (bipolar scale), Fant = fantasy, Persp = perspective-taking, Emp = empathic concern, Dist = Distress, Mr = random effects mean correlation.

+p ≤ .10  * p ≤ .05  ** p ≤ .01  *** p ≤ .001

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Enjoy reading for pleasure</th>
<th>Read for pleasure minus read for work</th>
<th>Kind of books read</th>
<th>Interest in Personality</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.17+</td>
<td>-</td>
<td>-.02</td>
<td>.29**</td>
</tr>
<tr>
<td>4</td>
<td>.14</td>
<td>.19*</td>
<td>.07</td>
<td>.19*</td>
</tr>
<tr>
<td>6</td>
<td>.10*</td>
<td>.16***</td>
<td>.13**</td>
<td>.14**</td>
</tr>
<tr>
<td>Mr</td>
<td>.12**</td>
<td>.17***</td>
<td>.09*</td>
<td>.18***</td>
</tr>
</tbody>
</table>

Note: = not applicable, Mr = random effects mean correlation. For kind of books read, high values indicate more fiction than non-fiction.

+p ≤ .10  * p ≤ .05  ** p ≤ .01  *** p ≤ .001