Journal of Personality Assessment, *96*(6), 610–618, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223891.2013.870571

Comparative Personality Judgments: Replication and Extension of Robust Findings in Personality Perception Using an Alternative Method

ANDREW BEER

Department of Psychology, University of South Carolina Upstate

The scientific study of accuracy in personality judgment typically involves the utilization of rating scales to make absolute decisions about a target individual. Although this method has many merits, it restricts some experimental options and is further removed from ecological validity than one would desire. These studies represent an attempt to develop an alternative methodology for the study of personality judgment—specifically for use in explorations of judgment process. A series of photo sets containing pictures of 3 individuals, each representing a different level of a specific personality trait, was created. The participant's task was to select high and low scorers on a dimension from the photos. Study 1 demonstrates that people can select targets with extreme scores from a photo lineup at a rate better than chance across several personality dimensions. Study 2 shows that this ability has some degree of temporal consistency. Study 3 represents an improvement on the general method via enhanced criteria for stimulus selection, incorporating both self and peer reports.

The general methodology for studying the accuracy of personality judgments under conditions of zero acquaintance—and most other conditions, for that matter—involves asking a judge to use a set of rating scales to assess a target individual, and then comparing those ratings via correlation to an accuracy criterion (often a target's self-judgments using those same rating scales). Thus, much of what we know about the accuracy of personality judgments is based on studies employing a very similar methodology. This methodology, although clearly optimal in many circumstances, could be considered limited in certain ways.

First, in a practical sense, it is somewhat difficult to study the personality judgment process under some of the restrictions presented by a rating scale-based method. Studies of social perception often involve choice paradigms, which allow researchers to examine the effects of mood (e.g., Ambady & Gray, 2002) and cognitive busyness (e.g., Patterson & Stockbridge, 1998) on the accuracy of various social judgments. Rating scale methodology makes exploration such as this much more difficult, as employment of the rating scale itself requires careful attention and most extensively validated personality measures take several minutes (at minimum) to administer. If personality judgments could be made more simply and quickly, personality researchers might be better able to situate the personality judgment process with respect to other social judgments in terms of the extent to which it is intuitive and subject to the effects of transient mood or cognitive load (for a brief review of these issues in social perception, see Ambady, 2010). Indeed, there has been recent interest in developing more intuitive (or less deliberative) methods for assessing personality in target individuals (Hirschmüller, Egloff, Nestler, & Back, 2013).

The second limitation is more general and concerns ecological validity. When people assess personality in natural settings, they are not typically assigning a number to a person's standing on a dimension. Instead, the organic process of attempting to understand stable intra-individual patterns of behavior occurs without the benefit of a formal rating system. The perceiver might employ constructs of his or her choosing, and judgments in these domains need not take any prescribed form. Often, the explicit output of the process is something akin to "Ted's a pretty reasonable guy" as opposed to "Ted's a 7 in terms of general reasonability." This sort of assessment is useful enough by and large, as it helps us to create general expectations and set our own behavioral agendas in relation to the individual in question (e.g., if I ask Ted to switch shifts with me because my child is ill, he will likely attempt to do so without complaint). However, many judgments of this nature involve not only the creation of an expectation, but a comparison of this expectation to those generated in response to others in the social environment. One chooses to ask Ted to switch shifts not only because "Ted's a reasonable guy," but perhaps because Ted's the most reasonable guy in the office or maybe because Ted is more reasonable than Frank.

Thus, in many scenarios personality is evaluated on a relative basis, often involving a choice. Some involve impactful personality-based decisions: Is candidate A smarter, more trustworthy, or friendlier than candidates B and C? Others involve more mundane personality-based decisions: Should one take the open seat on the train next to person A or person B? The social perceiver probably has more use for comparisons such as these, and perhaps is more inclined to make decisions in this fashion rather than interpreting what a point on a rating scale might mean in more general terms.

In fact, recent research indicates that relative judgments of personality might indeed be more valid than absolute judgments (Sheppard, Goffin, Lewis, & Olson, 2011). Although the relative judgment procedure outlined in these studies involves the use of ratings scales, many of the principles involved apply to this

Received June 26, 2013; Revised October 3, 2013.

Address correspondence to Andrew Beer, Department of Psychology, University of South Carolina Upstate, 132 CASB, 800 University Way, Spartanburg, SC 29303; Email: abeer@uscupstate.edu

set of studies. In particular, Goffin and Olson (2011) put forth 12 preconditions for the effective use of relative judgments, almost all of which are satisfied in personality rating situations. The first five preconditions concern what is being rated, and the authors suggested that relative judgments are best when the judgment (1) concerns a person (rather than a thing), when (2) the attributes being rated pertain to survival, when there is (3) no clear objective criterion available, and when judgments are (4) evaluative and (5) global. The authors also suggest that relative judgment procedures will work best when (6) there are multiple raters and when the referents are (7) familiar to the judge, (8) predetermined (rather than chosen by the judges) and (9) represented as groups (rather than single individuals). It is also best when these reference groups are (10) diverse in the attribute in question and (11) consistent across judges (i.e., each judge uses the same reference group). Finally, the authors suggested that (12) for relative judgments to be optimal, ratings "must be scaled in a manner that promotes their comparability across different raters" (p. 57). I contend that (a) personality judgments clearly satisfy the first five preconditions; (b) many circumstances under which we examine personality judgment satisfy preconditions 6 through 9, and (c) the studies outlined in this article generally conform to preconditions 10, 11, and 12. In sum, personality judgments—particularly personality judgments that involve comparing members of a diverse group against one another—seem quite suitable for a relative (as opposed to an absolute) rating scheme.

It is for these reasons that an alternative methodology for the study of personality perception is worthy of exploration. In these studies, I attempt to address some of the current methodological issues in the study of personality perception by investigating the utility of a comparative choice paradigm for personality judgment. As already outlined, a choice paradigm might represent a step toward greater ecological validity, and personality judgments might be an ideal circumstance for the application of relative judgment methodology. This alternative method will allow researchers to (a) conceptually replicate previous findings in personality judgment, helping to make a stronger case for the robustness of key findings, and (b) extend inquiry into personality judgment to circumstances in which traditional absolute rating paradigms might hinder study.

Overview of Studies

The studies described here represent an initial attempt at creating a relative, choice-based personality rating task. To simplify the issue, I have chosen to investigate the method in zero acquaintance situations. Each study employs a similar method to examine phenomena already addressed by extant published data obtained using rating scale methodology. In the first study, self–peer agreement is evaluated on major personality dimensions. These results are compared to the well-established zero acquaintance personality judgment literature. In Study 2, the notion that accurate personality judgment at zero acquaintance is a stable individual difference is explored. Finally, the design of Study 3 affords another opportunity to examine general accuracy in personality judgment as well as an opportunity to examine the extent to which the gender of the judge or target individual influence accuracy in this context.

Across all three studies, the procedure was relatively uniform. Participants were first given a written and oral description of a trait category, such as extraversion. After clarifying the definition of the judgment category, participants were presented with a set of three photographs. They were then instructed to indicate (a) the photograph of the individual who appeared to be the highest in the dimension relative to the others, and (b) the photograph of the individual who appeared to be lowest in the dimension. This set was typically followed by three other sets, at which point a new trait category was introduced and the judgment process repeated. Given its prominence in the personality perception literature, most judgments were focused on the Five-factor model (FFM; McCrae & Costa, 1999) of personality.

Study 1

There is a large body of research that indicates that people are generally not very adept at making personality judgments of strangers (Beer & Watson, 2008b; Borkenau & Liebler, 1992; Norman & Goldberg, 1966). The consistent exception to this rule is the trait of extraversion, which can be judged with moderate accuracy from something as simple as a photograph (Beer & Watson, 2010; Borkenau, Brecke, Möttig, & Paelecke, 2009; Borkenau & Liebler, 1992; Naumann, Vazire, Rentfrow, & Gosling, 2009), highlighting a phenomenon frequently referred to as the trait visibility effect (Funder, 1995). Would such findings hold given an entirely different methodology? Participants chose extreme scoring members of trait categories from sets of three photographs. It was hypothesized that self–peer agreement—as assessed by proportion correctly categorized—would be highest for extraversion and lowest for neuroticism, in accordance with previous findings.

Method

Participants. Participants were 90 undergraduates (25 male) between the ages of 17 and 42 ($M = 20.17$) recruited from introductory psychology courses at a midsized Southeastern U.S. university. Of these 90, 42 were Caucasian, 38 were African American, 2 were Asian, 2 were Hispanic, and 6 did not specify their ethnicity. These students participated in partial fulfillment of a course requirement.

Target Selection. Targets were selected from a group of 218 photographs (partial torso visible, "neutral" expression) of undergraduates (48 male) who had participated in a previous study (Beer & Watson, 2008b) during which each target completed a self-report instrument measuring the FFM and several other traits. Two traits in addition to the FFM—conservatism and thriftiness—were selected because they had shown significant self–stranger convergence in the original sample.

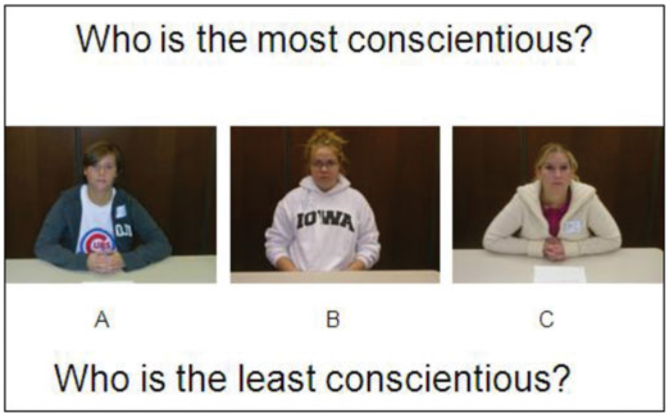
Stimulus selection depended on the prospective targets' self-reported standings on relevant trait measures. Scores on each trait dimension (FFM traits were measured using Saucier's [1994] minimarkers of Goldberg's adjective scales; thriftiness vs. extravagance and conservative versus liberal were measured using single-item bipolar scales) for each target were converted into percentile ranks from the entire initial target pool (i.e., 218 participants). To create each trial, one individual was selected who scored near the top of distribution (at least greater than the 70th percentile) on a given dimension, one individual scoring at or near the bottom of the distribution (at least lower than the 30th percentile), and one individual near the median of the distribution. This selection procedure ensured diversity in the reference group.

From the pool of 218 possible targets, 24 groups of three (4 groups of three for each of the traits comprising the FFM and 2 groups each for conservatism and thriftiness[1]) were selected to be used in Study 1. The selections were examined to ensure that the target data were consistently of high quality (no random responses, incomplete data, or low-quality photographs). The final stimuli set consisted of 72 photos (of 70 target individuals—2 individuals' photos appeared as part of a trial in more than one trait category) and mirrored the original pool in terms of gender distribution (15 male) and average standings on the primary personality dimensions of interest.

Measure. Self-reports of global personality were obtained from the judges using the Big Five Inventory (BFI; John, Naumann, & Soto, 2008). Scale reliabilities were consistent with previous work, with coefficient alphas ranging from .67 (conscientiousness) to .83 (agreeableness).

Procedure. Participants arrived at the laboratory in small groups. They first completed a brief demographic questionnaire, which assessed gender, age, and ethnicity. Then, they completed the self-report version of the BFI. After this, the experimental task was introduced.[2] Participants were told that they would be making a series of choices about people with whom they were unacquainted. The first dimension was described in detail by the experimenter, with an accompanying slide that contained a general definition of the trait construct, including what it meant to be (i.e., a high or low scorer in this domain). Participants were then shown the first set of three photos and asked to indicate on their answer sheets the highest and lowest scoring individuals from the set of photographs by circling the letter that corresponded to the chosen photograph (see Figure 1). The participants would then typically see three more sets of photographs, making six more choices. When everyone was finished, the experimenter moved to the next trait category and repeated the process. In total, there were seven trait categories: traits making up the FFM, conservatism, and thriftiness.

As mentioned previously, each FFM trait had four associated sets of photographs, and conservatism and thriftiness were each assessed using two sets of photographs. Thus, participants made a total of 48 choices across 24 sets of photographs. Because there were six possible orders in which photos could be horizontally presented in accordance with standings on the appropriate personality dimension (e.g., high-middle-low, low-middle-high, etc.), each photo order was represented four times in total across the 24 sets.


Who is the most conscientious?

Who is the least conscientious?

FIGURE 1.—Sample trial (color figure available online).

Results and Discussion

Accuracy was determined by the proportion of correct classifications across photo sets. Thus, if a participant chose the target labeled A for the question "Who is the most extraverted?" and this target was the highest scoring individual among the three pictured individuals, this would constitute a correct classification. Each set contained an extreme high scorer, an extreme low scorer, and an individual who scored near the median of the domain's distribution. In any given photo set, the expected proportion correct due to random guessing would be 1/3.[3] Obtained proportions were compared to this expected value using a z test for differences between proportions.[4]

Table 1 presents the accuracy across trait categories as represented by the proportion of correct classifications. Participants were more accurate than would be expected by chance in five of the seven traits assessed. In accordance with previous findings (e.g., Beer & Watson, 2008b, 2010; Borkenau & Liebler, 1992; Funder & Dobroth, 1987; Norman & Goldberg, 1966), extraversion showed the greatest perceiver accuracy among FFM trait dimensions in the choice task, with people accurately classifying photographs on 50% of the trials ($z = 3.43$, $p < .001$. Openness to experience followed closely at 48% correct classification ($z = 3.03$, $p = .002$); proportions for conscientiousness (45%, $z = 2.42$, $p = .016$), conservatism (43%, $z = 2.02$, $p = .043$), and thriftiness (66%, $z = 6.66$, $p < .001$) were also significantly greater than would be expected by chance (33%). These levels of accuracy are more impressive given that the stimuli in these cases were all strangers seen only in a photograph with a posed neutral expression.

[1] Due to differences in the distributions between FFM traits and the single-item indicators used to assess thriftiness and conservatism, the selection criteria for extreme scorers evenly across all judgment categories could not be applied. Thus, thriftiness and conservatism are represented by only two trials (as opposed to the four for each FFM trait). Additionally, in Study 3 the extreme scorers for positive and negative affect overlapped fairly strongly with those for extraversion and neuroticism, respectively. Thus, I was forced to similarly restrict the stimulus pool for those trials.

[2] The original study design involved an instructional manipulation. Half of the participants were instructed to think carefully and provide notes about their choices, and the other half were told to choose quickly. There were no significant differences in accuracy across the two groups, nor was there a consistent pattern in terms of direction of nonsignificant differences. Thus, the data were pooled.

[3] To explain the baseline probability, there are six orders of the three photos (H = true high, M = true middle, L = true low): HML, HLM, MHL, MLH, LHM, and LMH. Only one of these puts the true best first and the true worst last. So the probability of two correct on one set is 1/6. The probability of exactly one correct on one set is 2/6. So the expected number correct on a single pair of questions is: E(X) = 2*(1/6) + 1*(2/6) = 4/6. For all 24 pairs of questions, one would expect 24*4/6 = 16, or 1/3 of 48. Thus, all tests concerning the differences in proportions will use .33 as the baseline comparison standard.

[4] The formula for differences between observed and expected proportions is $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$ where $\hat{p}$ is the observed proportion, p is the expected proportion, q is $1 - p$ and n is the sample size.

TABLE 1.—Study 1 accuracy.

Trait	Mean Proportion Correct	z
Neuroticism	.36	.61
Extraversion	.50	3.43**
Openness	.48	3.03**
Agreeableness	.38	1.01
Conscientiousness	.45	2.42*
Conservatism	.43	2.02*
Thriftiness	.66	6.66**
Total score	.45	2.42*

Note. $N = 90$. Expected value for each proportion is .33.
$^{*}p < .05$. $^{**}p < .01$.

In an attempt to determine whether attributes of the judge were associated with ability to correctly classify photographs, I examined the correlations between self-reported personality and accuracy in each domain. No self-reported dimension of the FFM predicted overall accuracy, and BFI scale scores did not predict accuracy in any specific domain except for thriftiness. In this case, self-reported agreeableness negatively predicted accuracy in thriftiness judgments ($r = -.23$, $p < .05$), and self-reported neuroticism positively predicted accuracy in this domain ($r = .24$, $p < .05$). Thus, on the rare occasion that personality predicted accuracy, the pattern trended toward greater accuracy from more neurotic and less agreeable individuals.

A second relevant question concerned cross-domain consistency in accuracy: Do people who make accurate judgments in one category tend to do so in other categories as well? The answer was clearly "no." The average intercategory correlation was .004 with only one significant intercategory relation for conservatism and thriftiness ($r = .37$, $p < .01$). This general lack of consistency in accuracy across trait domains also mirrors findings from a study employing a more traditional judgment paradigm (Allik, Realo, Mõttus, & Kuppens, 2010).

STUDY 2

Study 1 demonstrated that some domains lend themselves to more judgmental accuracy but that the ability to make these comparative judgments across domains is not consistent within the individual. Study 2 had two primary aims. First, the study design afforded the opportunity to replicate the general accuracy findings from Study 1. Second, it afforded the opportunity to determine whether individuals' ability to correctly classify photographs was consistent within a given domain over time. Thus, the protocol from Study 1 was employed in a test–retest scenario. Participants completed the task once and then again (using the same target photo sets) exactly 2 months later.

Method

Participants. Participants were recruited from upper level psychology courses at a midsized Southeastern U.S. university at two separate points in time. A total of 112 students (21 male) between the ages of 18 and 43 ($M = 21.50$) participated in the first session. Of these 112, 54 were Caucasian, 45 were African American, 7 were Hispanic, 1 was Native American, and 5 did not specify their ethnicity. Ninety-eight students (19 male) between the ages of 18 and 43 ($M = 21.79$) participated in the second session, exactly 2 months later. Of these 98, 46 were Caucasian, 37 were African American, 3 were Asian, 5 were Hispanic, 2 were Native American, and 5 did not specify ethnicity. Participants received extra credit for their participation.

TABLE 2.—Study 2 accuracy and stability.

Trait	Time 1 Proportion[a]	Time 2 Proportion[b]	Retest Correlation[c]
Neuroticism	.37(.38)	.37(.37)	.11
Extraversion	.50**(.49**)	.54**(.55**)	.40**
Openness	.36(.37)	.41(.41)	.40**
Agreeableness	.41(.43)	.42(.43)	.35**
Conscientiousness	.44*(.43)	.44*(.43)	.19
Conservatism	.39(.37)	.43*(.45*)	.19
Thriftiness	.67**(.65**)	.71**(.72**)	.34**
Total score	.43*(.43)	.46**(.46*)	.38**

Note. Proportions for the retest sample only ($n = 66$) are in parentheses in columns 1 and 2. Asterisks in columns 1 and 2 indicate significant z scores for differences in observed proportions relative to chance, whereas asterisks in column 3 indicate correlations significantly greater than zero.
[a] $n = 112$. [b] $n = 98$. [c] $n = 66$.
$^{*}p < .05$. $^{**}p < .01$.

The retest sample consisted of the 66 participants (12 male) between the ages of 18 and 43 ($M = 21.64$) who were present at both assessment sessions. Of these 66, 29 were Caucasian, 29 were African American, 5 were Hispanic, 1 was Native American, and 2 did not specify their ethnicity. The retest sample did not differ significantly from the larger groups in terms of age, sex, ethnicity, or performance in the various domains of the choice task.

Measure. As in Study 1, self-reports of global personality were obtained using the BFI (John et al., 2008). Scale reliabilities were consistent with previous work, with coefficient alphas ranging from .70 (openness to experience) to .82 (extraversion).

Procedure. Study 2 took place in two separate large-group testing sessions 2 months apart from each other. Participants arrived in an auditorium in groups of approximately 50 and completed a demographic questionnaire and the BFI. Their attention was then directed toward the large screen at the front of the room for the remainder of the experimental session. At this point, they engaged in the same activity using the exact same stimuli as described in Study 1.

Results and Discussion

The first two data columns of Table 2 present the accuracy across trait categories at Times 1 and 2, respectively. The results are similar to Study 1, with extraversion and thriftiness showing the greatest degree of accuracy and neuroticism showing the least accuracy. In fact, the correlations between the accuracy values from Table 1 and those observed at Times 1 and 2 are .88 and .94, respectively. Finally, the strong column correlation from Time 1 to Time 2 accuracy ($r = .98$) supports the notion that trait observability as captured by this method is consistent over time.

Again, similar to Study 1, intercategory consistency was low (average intercategory correlation was .01 for Time 1 and .07 for Time 2), further demonstrating the domain specificity of judgmental accuracy in this context. At Time 1, there were no significant intercategory correlations. At Time 2, significant relations between performance in the agreeableness and conscientiousness domains were observed ($r = .25$, $p < .05$), the conscientiousness and openness domains ($r = .21, p < .05$), and as in Study 1, the conservatism and thriftiness domains ($r = .29$, $p < .01$).

The third data column of Table 2 presents the 2-month retest correlations for each trait domain. Four of the seven trait categories show significant retest correlations, suggesting that the ability to judge personality from a photo lineup has some degree of temporal stability, although it falls considerably short of the stability over the same interval of the FFM traits, for which retest correlations ranged from .70 (neuroticism) to .85 (agreeableness) in the same population. Retest correlations were stronger in trait categories for which participants made more accurate judgments (column vector correlation was .33 between Time 1 accuracy and retest correlations and .42 between Time 2 accuracy and retest correlations), likely reflecting the essentially random nature of the responses in those categories for which correct classification did not differ from chance. Given the length of the interassessment interval, one can reasonably conclude the memory effects were unlikely to explain the temporal stability of the choice task performance.

Study 3

Thus far, there have been parallel findings observed using this comparative choice method in terms of trait visibility at zero acquaintance. It also was established that the ability to make accurate comparative trait judgments within a specific domain is nonrandom and moderately stable over time. However, there has been no individual difference variable or target characteristic that predicts judgmental accuracy. In other words, some judges are better than others for some traits (and consistently so), and some traits are judged more easily by all. In Study 3, I attempted to address several limitations in the methodology of Studies 1 and 2, in turn allowing for consideration of some other phenomena in personality perception.

The target sample in Studies 1 and 2, although useful and informative on some levels, left several things to be desired. First, the gender composition was largely female. In that this mirrored the sample of judges, it was not a major problem, but it restricted the ability to examine target gender effects, a potentially interesting avenue of inquiry. In addition, targets were selected based on self-reports of personality, which are certainly flawed to some extent (for a recent discussion, see Vazire, 2010). Third, the targets were placed in a combination of single and mixed gender arrays, occasionally allowing for gender stereotypes to potentially drive some effects. Finally, targets in Studies 1 and 2 were instructed to make “neutral” facial expressions for their photographs, whereas the Study 3 targets were not given such instructions. The target sample in Study 3 was derived using more stringent selection criteria (based on a unique combination of concordant self and peer judgments), featured an equal number of same-gender trials in each trait category, and involved more spontaneous, natural target facial expressions. In addition, the existing target data allowed for observation of accuracy in positive and negative affect in addition to the FFM dimensions assessed in Study 1. This could be particularly instructive given the discordant findings regarding judgmental accuracy for these major dimensions of trait affect and the related FFM traits of neuroticism and extraversion. Specifically, FFM traits show stronger self–other agreement despite strong convergence between personality and affect measures within self-ratings (for a review, see McDade-Montez, Watson, & Beer, 2013). In all, Study 3 provides an opportunity to replicate and extend findings from Studies 1 and 2 using a new sample of targets and judges.

Method

Participants. Participants were 107 undergraduates (32 male) between the ages of 18 and 50 ($M = 20.17$) recruited from introductory psychology courses at a midsized Southeastern U.S. university. Of these 107, 65 were Caucasian, 33 were African American, 3 were Asian, 2 were Hispanic, and 4 did not specify their ethnicity. These students participated in partial fulfillment of a course requirement.

Target selection. Targets were selected from photographs of 291 married couples (mean age = 28.2 years) who had participated in a previous study (for more sample details, see Watson et al., 2004) during which each couple completed both self- and peer-report instruments measuring the FFM using the BFI (John et al., 2008) and several other traits, including positive and negative affectivity (Watson, Clark, & Tellegen, 1988). The target selection procedure was essentially identical to the one described in Study 1, except instead of basing selections on self-reports, a combination of the self- and spouse-reported dimensions was utilized. Thus, for example, the high-scoring photos for the extraversion category consisted of targets for whom both self- and spouse-reported extraversion met selection criteria. In the end, the target sample was constituted by individuals who were evaluated by their spouses very similarly to how they evaluated themselves. In fact, in most cases, the scale scores on the selection domain were identical for self- and spouse ratings. Although this does not guarantee the validity of the assessment, it represents a clear improvement over relying on self-judgments as the sole selection criterion.

Measure. As in Studies 1 and 2, self-reports of global personality were obtained using the BFI (John et al., 2008). Scale reliabilities were consistent with previous work, with coefficient alphas ranging from .76 (openness to experience) to .85 (extraversion).

Procedure. Study 3 followed almost exactly the same procedures as Study 1, except that two trait domains (conservatism and thriftiness) were replaced with new categories (positive and negative affect). Participants arrived at the laboratory in small groups, completed a brief demographic questionnaire and the self-report version of the BFI, and then made the comparative judgments in seven trait categories (neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, positive affect, and negative affect). Each category had four associated sets of photographs except for positive affect and negative affect, which each consisted of only two sets of photographs. Again, participants made a total of 48 choices across 24 sets of photographs.

Results and Discussion

Table 3 presents accuracy across trait categories as represented by the proportion of correct classifications. Once again, extraversion shows the greatest accuracy (54% correct classification, $z = 4.62, p < .001$), and neuroticism (32%) is among the most difficult categories. Unsurprisingly, negative affect is even more difficult (25%), as neuroticism is very difficult to discern

TABLE 3.—Study 3 accuracy.

Trait	Mean Proportion Correct	z
Neuroticism	.32	−.22
Extraversion	.54	4.62**
Openness	.09	−5.28**
Agreeableness	.54	4.62**
Conscientious	.36	.66
Positive affect	.48	3.30**
Negative affect	.25	−1.76
Total	.39	1.32

Note. $n = 107$.
**$p < .01$.

from photographs (Beer & Watson, 2010; Borkenau & Liebler, 1992; Naumann et al., 2009) and well-acquainted perceivers are generally less accurate in determining negative affect than in doing so for neuroticism (Watson, Hubbard, & Wiese, 2000). On a related note, participants were significantly more accurate than chance in judging positive affect (48% correct, $z = 3.30$, $p = .001$), which might also come as no surprise given its typically strong positive relation to extraversion (Lucas, Diener, Grob, Suh, & Shao, 2000; Watson & Clark, 1997).

On the other hand, there were two relatively surprising findings in Study 3. First, participants were able to correctly classify individuals in terms of agreeableness (54% correct, $z = 4.62$, $p < .001$). Typically, a still photograph is not enough on its own to generate accurate assessments of agreeableness (Beer & Watson, 2010; Borkenau & Liebler, 1992; Naumann et al., 2009). This finding is potentially important in that agreeableness is a domain that perceivers value highly yet are largely unable to discern (Ames & Bianchi, 2008). Conversely, in the only such instance across the three studies, participants performed at a level significantly worse than chance in the domain of openness to experience (9% correct, $z = -5.28$, $p < .001$). One possible explanation centers on the cohort-based age gap between judges and targets in Study 3 that was not present in Studies 1 and 2. The targets for Study 3 were typically born in the early 1970s, whereas the judges were typically born in the early 1990s. It is true that this issue does not seem to have affected other personality judgments, but perhaps openness to experience is a realm of personality judgment that relies heavily on shared interpretations of nonverbal symbols between judge and target. If so, this generational gap might hinder utilization of valid static visual cues (e.g., iconic t-shirts) to openness. On a related note, perhaps these valid cues simply are not on display in the older target sample. Many of these participants came from work as opposed to class prior to being photographed, and expressive clothing in general might be less common in this age group relative to a college student population. This, however, would not fully explain that participants performed significantly worse than would be expected by chance.[5] A closer examination of the data reveals that the accuracy for openness was particularly low for male targets (8% correct classification) whereas participants performed at around chance levels for female targets (35% correct classification). Thus, for reasons currently unknown, people had difficulty discerning differences in openness to experience regardless of target gender, and when it came to judging men, people tended to view them as opposite of their true nature.

The search for characteristics of the "good judge" once again proved relatively fruitless in Study 3, as very few self-reported personality traits predicted total or domain-specific accuracy. There were a few exceptions, however. First, extraversion was inversely correlated with overall accuracy ($r = -.22$, $p < .05$), likely driven by its inverse relation with judgmental accuracy in the domain of positive affect ($r = -.23$, $p < .05$). In addition, agreeableness was positively associated with accuracy in the domains of conscientiousness ($r = .20, p < .05$) and positive affect ($r = .26, p < .01$). This latter finding is at odds with the inverse relation between agreeableness and accuracy in thriftiness observed in Study 1. In summary, in keeping with the good judge literature as a whole, no consistent trait predictors of global or specific accuracy have been observed.

As noted previously, the target pool in Study 3 allowed for examination of target gender effects, so although these results might not be able to speak to the qualities of the good judge in this setting, perhaps they can inform some conclusions as to what constitutes a good target. Overall, it seems that making distinctions among small groups of female targets (44% correct classification, collapsed across all domains) comes more easily than making the same distinctions among groups of male targets regardless of the judge's gender (44% vs. 34% correct classification, $z = 2.42, p < .05$). No Target × Gender interactions were observed in terms of overall accuracy, as men (34%) and women (34%) judged men at approximately the same level of accuracy, and the same held for male (43%) and female (46%) judgments of women. At the domain level, however, there was one notable interactive effect in ratings of agreeableness: Women seemed to have more difficulty discerning differences among women (45% correct classification) than did women judging men (60%), men judging men (56%), or men judging women (59%).[6] Clearly, this effect requires replication, but it could help explain some of the null findings with respect to judgmental accuracy in agreeableness at zero acquaintance, given that much of this research involves samples in which women are overrepresented in both judge and target pools.

GENERAL DISCUSSION

The purpose of these studies was to investigate the utility of a comparative choice-based methodology for use in studies of personality perception. To do so, two parallel sets of target stimuli were created and tested with three different groups of judges, demonstrating a general consistency for some key findings in personality perception at zero acquaintance.

[5] A second possible explanation hinges on the fact that the two target populations were classified by measures with slightly different interpretations of openness to experience—although the descriptions provided to participants did not vary across the studies. To the extent that the description more closely tracked Goldberg's Intellect as opposed to the BFI's slightly broader Openness to Experience designation, it might have provided an advantage to participants in Studies 1 and 2. This is unlikely, however, given that the description provided was more in line with the BFI's operational definition of the construct.

[6] Due to the disparate sample sizes, I could not compare these proportions across judge gender in the same manner that the previous analyses were conducted. However, the within-female judge difference (women judged men more accurately than women judged other women) observed was statistically significant ($z = 2.61$, $p < .01$), and the within-male judge difference was not.

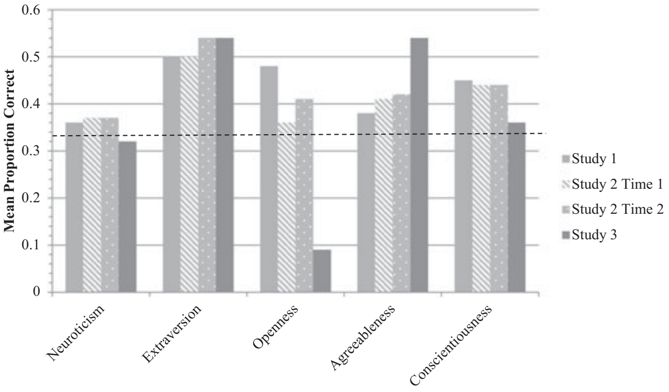


FIGURE 2.—Accuracy across studies.

Accuracy at Zero Acquaintance

Figure 2 provides a summary of the accuracy findings across studies for the five domains that appeared in each assessment. Overall, initial testing indicates that this alternative methodology yields results similar to more traditional rating scale methods in that extraversion is clearly the most observable trait in zero acquaintance situations (Beer & Watson, 2008b; Borkenau & Liebler, 1992; Norman & Goldberg, 1966; Watson, 1989). This consistency lends credence to both the effect and the soundness of the method employed in these studies. Accordingly, experimenters can feel confident that this comparative choice paradigm is tapping some of the same underlying judgment processes as the more typical rating scale-based procedures. In addition, however, there was inconsistent support for accuracy in conscientiousness, openness, conservatism, agreeableness, and positive affect, indicating that changing the format of the evaluation might lead to increased perceptual accuracy for other traits and perhaps suggesting that some feature of the task might be more conducive to accurate judgment than rating scale procedures.

Interestingly, accuracy did not seem to be substantially affected by changing the criterion from self-reports to the potentially more valid combination of self- and peer-reports. The only dramatic difference between Studies 1 and 2 and Study 3 was that accuracy for openness was significantly worse than chance in Study 3, which could be attributed to differences in target sample composition peripheral to method of selection.

In Search of the Good Judge

In general, consistent characteristics of the good judge are somewhat elusive. Since researchers first began examining the issue more than 70 years ago, very few trends have emerged (for recent discussions, see Christiansen, Wolcott-Burnam, Janovics, Burns, & Quirk, 2005; Human & Biesanz, 2011; Letzring, 2008). Some of this inconsistency might indeed be due to inconsistency in both defining accuracy and the setting in which the judgment takes place (e.g., face to face, known vs. unknown others, etc.). In keeping with this general phenomenon, little support was found for the idea that there exists a "good judge," at least as the term pertains to assessing global personality from a still photograph. There was very little consistency for accuracy within individuals across domains. There do not appear to be any studies that have demonstrated such consistency using rating scales, either, although some have shown that accuracy as measured by profile correlations based on relatively little interaction tends to be consistent within an individual (Letzring, 2008; Letzring, Wells, & Funder, 2006).

Furthermore, no consistent relations between the primary individual difference variables assessed in these studies and judgmental accuracy, general or specific, have been found. One semiconsistent finding in the good judge literature is that females make better judges than males (Chan, Rogers, Parisotto, & Biesanz, 2011; Letzring, 2010; Taft, 1955), although such a trend was not observed in these data. However, one interesting Judge × Target interaction is noteworthy in Study 3: In the domain of agreeableness, women have more difficulty delineating between other women when compared to all other judge–target gender combinations. Recent research using rating scale methodology indicates that females, in general, make better judges and targets, and that females judging other females is an optimal judge–target gender pairing (Letzring, 2010). Although this work is concordant with respect to target effects (Study 3 suggests that women do indeed make better targets), the specific all-women judge–target advantage using this comparative judgment method was not replicated.

Despite the inconsistent performance across domains and inability to detect relations between individual differences and accuracy, Study 2 did reveal moderate temporal stability for accuracy in some of these dimensions, leaving open the possibility that the ability to choose extreme scorers from a lineup might be related to consistent intra-individual cognitive processes. It is noteworthy that the trait categories for which we see little temporal stability are also those that show the least accuracy across all studies, likely due to the fact that there are little to no available cues in these photographs for those trait dimensions, and thus the choices are truly random. Future research should aim to replicate these within-domain consistencies and further explicate their origins. For instance, it is possible that trait chronicity (e.g., Bargh, Bond, Lombardi, & Tota, 1986)—the extent to which a trait category is chronically accessible—could predict performance in a domain. In addition, there are numerous individual differences that have yet to be assessed in relation to accuracy in this paradigm.

It is also important to note that the kind of accuracy assessed in these studies is highly specified. I was asking participants to make comparative judgments among targets; that is, these judgments were made relative to the comparison set only, rather than to some normative standard of behavior. Hence, it might be more appropriate to describe these findings as pertaining to differential accuracy—the ability to evaluate differences among targets—as opposed to normative accuracy—the general understanding of how the average person behaves. Recent research indicates that different individual difference variables predict normative versus distinctive accuracy, with, for example, psychological adjustment showing a positive relation with normative accuracy (Human & Biesanz, 2011).

Prospects for a Comparative Choice Methodology in Personality Perception Research

The results from these studies suggest that a comparative choice-based format could have some promise in terms of future research. The fact that this approach yielded similar general results to those observed using rating scale methodology with respect to (a) accuracy at zero acquaintance, (b) relative trait visibility, (c) consistency in accuracy across domains, and (d) general target gender effects is encouraging in terms of the viability of this method, and it also gives more weight to some of the novel findings in these studies. The work in this area is far from complete, however. First, the method itself can be modified and improved on in myriad ways. Perhaps expanding or restricting the number of alternatives could be impactful. Or perhaps the choice need not be comparative at all, but rather a binary yes–no or high–low decision task could accomplish similar research goals in a simpler fashion. It is also worth noting that the targets in these studies represented a preselected set of stimuli that conformed to a strict distribution in a given domain. Future study might involve relaxing the stimulus selection criteria or even allowing participants to generate their own comparison sets from individuals with whom they are acquainted.

That said, it could still be argued that many important decisions regarding personality perception under conditions of limited acquaintance are made on a comparative choice basis, a phenomenon well captured by the current method. In fact, some argue that all judgments of personality are relative (A. M. Wood, Brown, Maltby, & Watkinson, 2012). In addition, judgmental errors in the realm of personality perception are more impactful when they are made among highly disparate options. Failing to distinguish between two highly conscientious job candidates is far less disastrous than failing to distinguish between a very conscientious individual and her opposite. This is also well captured by this paradigm.

The quickness and simplicity of the decision task in this method represents a step toward ecological validity, edging closer to the way individuals actually think about and apply trait constructs to others in daily life. Furthermore, a task of this nature might enable formal inquiry into the process of accurate personality perception via facilitating work involving reaction time, influences of cognitive load or mood manipulations, and formalized length-of-exposure manipulations. All of these areas of inquiry are more easily and frequently explored in other areas of social judgment but currently more difficult to examine with respect to accurate personality judgment.

Beyond the improvements that can be made in zero acquaintance applications, the choice format could be utilized in judgments of known others or perhaps even self-judgments. Ozer (1993) posited that personality assessment in general could be informed by psychophysics. Indeed, it would be worthwhile to examine personality judgments attained via a series of comparisons between the self and a group of known others and whether such estimates could predict behavior as or more effectively as traditional rating-scale-based judgments.

Despite these advantages, this method is not intended as a replacement for long-standing methods in personality perception. Theoretical and computational frameworks such as the PERSON model (Kenny, 2004) and the social accuracy model (Biesanz, 2010) generally require the use of rating scales, preferably with a high volume of items, but they allow researchers to examine ideas—particularly in the realm of perceiver effects of various sorts (e.g., Beer & Watson, 2008a; Human & Biesanz, 2012; Srivastava, Guglielmo, & Beer, 2010; D. Wood, Harms, & Vazire, 2010)—that the comparative choice paradigm described in this article does not. In addition, the search for quicker and simpler methods of evaluation diminishes the precision of the judgments. Nevertheless, some popular theories about accuracy in personality judgment, such as Funder's (1995) realistic accuracy model can be further validated via the use of such methodology. In particular, experimental tests aimed at explicating moderators of accuracy (i.e., the good judge, the good target, the good trait, and good information) could benefit from a method that lends itself to quick decisions made under different judgment conditions. Best practice in this regard would involve the use of a sensible, broad-based accuracy criterion (cf. Letzring et al., 2006) in stimulus selection, multiple trials within each judgment category, and otherwise conforming to most of the conditions put forth by Goffin and Olson (2011) with respect to the optimal utility of relative judgments. I would also recommend employing the methodology in circumstances where it is clear that the construct in question is discernible from the provided stimuli. To this end, one should consult the literature when possible, or perhaps utilize a panel of experts, to determine the suitability of stimulus material to the proposed judgment task. Constructing stimuli sets such as these can be somewhat laborious in that it involves extensive data collection, but in theory, anyone with access to photographs of individuals and associated personality measures could create his or her own set. Those without ready access to such information could certainly request the appropriate materials from researchers who typically work in these areas. In the short term, those interested in using this methodology could use the sets created for these studies, but for the sake of generalizability, researchers could create several parallel sets to be made publically available.

In summary, these studies represent a humble beginning to the application of comparative choice paradigms to personality judgment. The evidence presented here indicates that such methods have promise as a supplement to—not a replacement for—more traditional methods of personality assessment and can inform us on topics that heretofore have been less accessible for study.

Acknowledgments

I would like to thank David Watson and Evan Krauter for comments on early drafts of this article.

References

Allik, J., Realo, A., Mõttus, R., & Kuppens, P. (2010). Generalizability of self–other agreement from one personality trait to another. *Personality and Individual Differences*, *48*, 128–132. doi:10.1016/j.paid.2009.09.008

Ambady, N. (2010). The perils of pondering: Intuition and thin slice judgments. *Psychological Inquiry*, *21*, 271–278. doi:10.1080/1047840X.2010.524882

Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology*, *83*, 947–961. doi:10.1037/0022-3514.83.4.947

Ames, D. R., & Bianchi, E. C. (2008). The agreeableness asymmetry in first impressions: Perceivers' impulse to (mis)judge agreeableness and how it is moderated by power. *Personality and Social Psychology Bulletin*, *34*, 1719–1736. doi:10.1177/0146167208323932

Bargh, J. A., Bond, R. N., Lombardi, W. J., & Tota, M. E. (1986). The additive nature of chronic and temporary sources of construct accessibility. *Journal of Personality and Social Psychology*, *50*, 869–878. doi:10.1037/0022-3514.50.5.869

Beer, A., & Watson, D. (2008a). Asymmetry in judgments of personality: Others are less differentiated than the self. *Journal of Personality*, *76*, 535–559. doi:10.1111/j.1467-6494.2008.00495.x

Beer, A., & Watson, D. (2008b). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment*, *90*, 250–260. doi:10.1080/00223890701884970

Beer, A., & Watson, D. (2010). The effects of information and exposure on self–other agreement. *Journal of Research in Personality*, *44*, 38–45. doi:10.1016/j.jrp.2009.10.002

Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research*, *45*, 853–885. doi:10.1080/00273171.2010.519262

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*, 703–706. doi:10.1016/j.jrp.2009.03.007

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, *62*, 645–657. doi:10.1037/0022-3514.62.4.645

Chan, M., Rogers, K. H., Parisotto, K. L., & Biesanz, J. C. (2011). Forming first impressions: The role of gender and normative accuracy in personality perception. *Journal of Research in Personality*, *45*, 117–120. doi:10.1016/j.jrp.2010.11.001

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. doi:10.1207/s15327043hup1802_2

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670. doi:10.1037/0033-295X.102.4.652

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*, 409–418. doi:10.1037/0022-3514.52.2.409

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, *6*, 48–60. doi:10.1177/1745691610393521

Hirschmüller, S., Egloff, B., Nestler, S., & Back, M. D. (2013). The dual lens model: A comprehensive framework for understanding self–other agreement of personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *104*, 335–353. doi:10.1037/a0030383

Human, L. J., & Biesanz, J. C. (2011). Through the looking glass clearly: Accuracy and assumed similarity in well-adjusted individuals' first impressions. *Journal of Personality and Social Psychology*, *100*, 349–364. doi:10.1037/a0021850

Human, L. J., & Biesanz, J. C. (2012). Accuracy and assumed similarity in first impressions of personality: Differing associations at different levels of analysis. *Journal of Research in Personality*, *46*, 106–110. doi:10.1016/j.jrp.2011.10.002

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). New York, NY: Guilford.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review*, *8*, 265–280. doi:10.1207/s15327957pspr0803_3

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

Letzring, T. D. (2010). The effects of judge–target gender and ethnicity similarity on the accuracy of personality judgments. *Social Psychology*, *41*, 42–51. doi:10.1027/1864-9335/a000007

Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology*, *91*, 111–123. doi:10.1037/0022-3514.91.1.111

Lucas, R. E., Diener, E., Grob, A., Suh, E. M., & Shao, L. (2000). Cross-cultural evidence for the fundamental features of extraversion. *Journal of Personality and Social Psychology*, *79*, 452–468. doi:10.1037/0022-3514.79.3.452

McCrae, R. R., & Costa, P. R. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). New York, NY: Guilford.

McDade-Montez, E., Watson, D., & Beer, A. (2013). Similarity, agreement, and assumed similarity in proxy end-of-life decision making. *Families, Systems, & Health*. Advance online publication. doi:10.1037/a0033372

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, *35*, 1661–1671. doi:10.1177/0146167209346309

Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology*, *4*, 681–691. doi:10.1037/h0024002

Ozer, D. J. (1993). Classical psychophysics and the assessment of agreement and accuracy in judgments of personality. *Journal of Personality*, *61*, 739–767. doi:10.1111/j.1467-6494.1993.tb00789.x

Patterson, M. L., & Stockbridge, E. (1998). Effects of cognitive demand and judgment strategy on person perception accuracy. *Journal of Nonverbal Behavior*, *22*, 253–263. doi:10.1023/A:1022996522793

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment*, *63*, 506–516. doi:10.1207/s15327752jpa6303_8

Sheppard, L. D., Goffin, R. D., Lewis, R. J., & Olson, J. (2011). The effect of target attractiveness and rating method on the accuracy of trait ratings. *Journal of Personnel Psychology*, *10*, 24–33. doi:10.1027/1866-5888/a000030

Srivastava, S., Guglielmo, S., & Beer, J. S. (2010). Perceiving others' personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology*, *98*, 520–534. doi:10.1037/a0017057

Taft, R. (1955). The ability to judge people. *Psychological Bulletin*, *52*, 1–23. doi:10.1037/h0044999

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*, 281–300. doi:10.1037/a0017908

Watson, D. (1989). Strangers' ratings of the five robust personality factors: Evidence of a surprising convergence with self-report. *Journal of Personality and Social Psychology*, *57*, 120–128. doi:10.1037/0022-3514.57.1.120

Watson, D., & Clark, L. A. (1997). Measurement and mismeasurement of mood: Recurrent and emergent issues. *Journal of Personality Assessment*, *68*, 267–296. doi:10.1207/s15327752jpa6802_4

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. doi:10.1037/0022-3514.54.6.1063

Watson, D., Hubbard, B., & Wiese, D. (2000). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, *78*, 546–558. doi:10.1037/0022-3514.78.3.546

Watson, D., Klohnen, E. C., Casillas, A., Nus Simms, E., Haig, J., & Berry, D. S. (2004). Match makers and deal breakers: Analyses of assortative mating in newlywed couples. *Journal of Personality*, *72*, 1029–1068. doi:10.1111/j.0022-3506.2004.00289.x

Wood, A. M., Brown, G. A., Maltby, J., & Watkinson, P. (2012). How are personality judgments made? A cognitive model of reference group effects, personality scale responses, and behavioral reactions. *Journal of Personality*, *80*, 1275–1311. doi:10.1111/j.1467-6494.2012.00763.x

Wood, D., Harms, P., & Vazire, S. (2010). Perceiver effects as projective tests: What your perceptions of others say about you. *Journal of Personality and Social Psychology*, *99*, 174–190. doi:10.1037/a0019390