Psychological Bulletin
2010, Vol. 136, No. 6, 1092–1122

0033-2909/10/$12.00 DOI: 10.1037/a0021212

An Other Perspective on Personality: Meta-Analytic Integration of Observers' Accuracy and Predictive Validity

Brian S. Connelly
University of Toronto

Deniz S. Ones
University of Minnesota

The bulk of personality research has been built from self-report measures of personality. However, collecting personality ratings from other-raters, such as family, friends, and even strangers, is a dramatically underutilized method that allows better explanation and prediction of personality's role in many domains of psychology. Drawing hypotheses from D. C. Funder's (1995) realistic accuracy model about trait and information moderators of accuracy, we offer 3 meta-analyses to help researchers and applied psychologists understand and interpret both consistencies and unique insights afforded by other-ratings of personality. These meta-analyses integrate findings based on 44,178 target individuals rated across 263 independent samples. Each meta-analysis assessed the accuracy of observer ratings, as indexed by interrater consensus/reliability (Study 1), self–other correlations (Study 2), and predictions of behavior (Study 3). The results show that although increased frequency of interacting with targets does improve accuracy in rating personality, informants' interpersonal intimacy with the target is necessary for substantial increases in other-rating accuracy. Interpersonal intimacy improved accuracy especially for traits low in visibility (e.g., Emotional Stability) but only minimally for traits high in evaluativeness (e.g., Agreeableness). In addition, observer ratings were strong predictors of behaviors. When the criterion was academic achievement or job performance, other-ratings yielded predictive validities substantially greater than and incremental to self-ratings. These findings indicate that extraordinary value can gained by using other-reports to measure personality, and these findings provide guidelines toward enriching personality theory. Various subfields of psychology in which personality variables are systematically assessed and utilized in research and practice can benefit tremendously from use of others' ratings to measure personality variables.

Keywords: personality, meta-analysis, observers, informants, consensus

Supplemental materials: http://dx.doi.org/10.1037/a0021212.supp

Research in personality psychology has made substantial contributions across many domains of psychology. In health psychology, personality traits have been closely linked both to adopting health-promoting lifestyles and to providing resiliency to health threats (Bogg & Roberts, 2004; Friedman & Booth-Kewley, 2003; Lahey, 2009; Munafò et al., 2003). In clinical psychology, personality traits predict susceptibility to many disorders (Cassin & von Ranson, 2005; Malouff, Thorsteinsson, & Schutte, 2005), and some argue that the gap between "normal" and "abnormal" personality traits is much smaller than previously supposed (Krueger, Caspi, Moffitt, Silva, & McGee, 1996; Krueger & Tackett, 2003; Saulsman & Page, 2004). In behavioral genetics and neuroscience, researchers are increasingly linking individual differences in physiology and anatomy to stable personality traits (Ebstein et al., 1996; Munafò et al., 2003). In industrial and organizational psychology and educational psychology, personality traits strongly predict individuals' motivation, performance, advancement, and attitudes (Barrick, Mount, & Judge, 2001; Hough, 1992; Judge, Heller, & Mount, 2002; Ones, Dilchert, Viswesvaran, & Judge, 2007; Poropat, 2009). Last, personality traits are strong predictors of major life outcomes, including occupational attainment, divorce, and mortality (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). With such robust influences of personality traits across fields of psychology, theories in a broad and diverse set of research are increasingly adjusting to incorporate the influences of personality.

Developing the five-factor model of personality (Costa & McCrae, 1992; Digman, 1990; Goldberg, 1993) has laid much of the groundwork for the proliferation of personality research across so many areas of psychology. The five-factor model posits that the five basic dimensions—Emotional Stability, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness—underlie personality. Though contemporary theorists view personality traits as hierarchically organized, with both narrower and broader traits than these five factors (Costa & McCrae, 1995; DeYoung, 2006; Digman, 1997), the array of particular trait scales within many personality inventories can be meaningfully aligned with these five factors. The five factors have emerged across

Brian S. Connelly, Department of Management, University of Toronto Scarborough, Toronto, Ontario, Canada; Deniz S. Ones, Department of Psychology, University of Minnesota.

This research was supported by a research grant from the College Board. We thank Matthew J. Borneman, Colin G. DeYoung, David A. Kenny, Fred L. Oswald, and SiSi Tran for their comments on an earlier draft of this paper.

Correspondence concerning this article should be addressed to Brian S. Connelly, Department of Management, University of Toronto Scarborough, Toronto, Ontario M1C 1A4, Canada. E-mail: brian.connelly@utoronto.ca

personality inventories, genders, ages, raters, languages, and cultures (Dilchert, Ones, Van Rooy, & Viswesvaran, 2006; McCrae & Costa, 1987, 1997), suggesting that the five-factor model is a strongly generalizable framework for describing personality traits. Though some have argued for alternate conceptualizations of personality (e.g., Block, 1995; Eysenck, 1993) and for the merit of studying trait compounds that span multiple Big Five factors (Ones & Viswesvaran, 2001), the five-factor model is the organizing taxonomy for personality traits most frequently used by meta-analysts across these fields (e.g., Barrick et al., 2001; Heller, Watson, & Hies, 2004; Malouff et al., 2005; Steel, 2007). Adopting this common, meaningful language has allowed integration of research findings from an otherwise jangling sprawl of personality theories and measures.

Though these veins of personality research have benefited from the availability of a common language, most researchers have not taken advantage of the diversity of methods for measuring personality traits. Instead, research in these fields typically measures personality with only a single, self-report questionnaire. Despite the urging of some (e.g., Funder, 1999; Klonsky, Oltmanns, & Turkheimer, 2002; Vazire, 2006), fields outside of personality or social psychology rarely collect personality measures from informants who have observed targets ("other-raters," e.g., spouses, friends, even complete strangers). If these other-ratings overlap so strongly with self-report measures that they are redundant and locate personality traits similarly within relevant nomological nets, psychologists could confidently proceed using only self-reports and conducting business as usual. If other-ratings contain some degree of unique, trait-relevant information, however, other-ratings may enrich the theoretical understanding of personality traits' role in areas that would otherwise be neglected.

The capacity for other-ratings to contribute incrementally to personality research in these fields is contingent on those ratings being accurate. Fortunately, decades of research and debate within personality and social psychology have addressed the accuracy of other-ratings. Research has ranged from specifying the conditions necessary for accurate judgment to be possible (Funder, 1995; Kenny, 1991) to searching for the most accurate judges (Bernieri, Zuckerman, Koestner, & Rosenthal, 1994; Vernon, 1933) to questioning whether accurate judgments about others' personality are even possible (Shrauger & Schoeneman, 1979). Moreover, cross-observer accuracy can occur only if the target has behaved in a fairly consistent way across situations in which trait perceptions are formed. As a result, research on observers' accuracy has been at the heart of the person–situation debate (Kenrick & Funder, 1988). Finding such accuracy foretells a more favorable outlook on trait psychology as a whole by implying cross-situational consistency in behavior.

This research base contains a wealth of information addressing how accurate other-ratings are, when other-ratings will be more or less accurate, and why other-rating accuracy may vary. Unfortunately, empirical integration of this information has been limited, and many "whens" and "whys" of other-rating accuracy have yet to be addressed. In this paper, we present three meta-analyses that bring cohesion and clarity to a broad and divergent set of research on the accuracy of others' judgments of personality. For personality and social psychology, these meta-analyses synthesize and extend existing understanding of the extent and process of achieving accurate perceptions of others' personality traits. For the broader community of psychologists incorporating personality traits in particular disciplines, these meta-analyses build a basis for evaluating the potential contribution of other-ratings and prescribe guidelines about which traits are most easily rated, which others are most able to rate, and which trait cues are most critical. In the pages that follow, we first review the general purposes for which research has measured personality by using other-ratings and discuss how the streams of research can be integrated within an accuracy framework. Next, we describe findings from three meta-analyses, each focusing on evaluating a particular accuracy criterion: interrater reliability/consensus (Study 1), correspondence with self-ratings (Study 2), and behavioral predictions (Study 3). Finally, we close by describing how our findings inform theory about the accuracy of other-ratings and personality measurement more broadly.

Previous Research Examining the Accuracy of Other-Ratings of Personality

Research has generally examined other-ratings of personality for one of three purposes. First and most commonly, researchers have used other-ratings to replicate major findings about personality across measurement methods. Such studies have examined other-ratings' temporal stability (Costa & McCrae, 1988), factor structure (e.g., McCrae & Costa, 1987), and heritability (e.g., Riemann, Angleitner, & Strelau, 1997), as well as simply contributed to validation of self-report measures. Second, other-ratings played a central role in the now legendary person–situation debate waged between the mid 1960s and 1980s (for a historical overview, see Kenrick & Funder, 1988). Evaluating the correspondence between personality trait ratings from different sources was one approach to testing the stability of behavior across situations. That is, ratings of two individuals observing a target in different settings can correspond only if both (a) that target has behaved in a consistent manner across settings and (b) those observers have interpreted the target's behavior in a similar way. Ensuing research showed that cross-context observers do agree when they are well acquainted with the target, even when they are separated by substantial time gaps (Costa & McCrae, 1988). Last, a stream of research in social psychology has examined the process and accuracy of person perception when judgments are made by relatively unacquainted other-raters. For example, the weighted accuracy model (Kenny, 1991) and PERSON model (Kenny, 2004) depict the effects on agreement between observers' trait ratings from increased observation of targets, overlapping opportunities for observation, stereotypes, and shared meaning. The general finding across studies in social psychology examining person perception has been moderate interrater and self–other correlations ($\approx$.30) for Extraversion and, to a slightly lesser extent, Conscientiousness. Accuracy for the other Big Five traits generally emerges only in designs that clearly elicit trait-relevant behavior. Time-series studies of these previously unacquainted other-raters and targets have not shown interrater reliability to increase with increased observation of the target (Kenny, Albright, Malloy, & Kashy, 1994). Thus, accuracy does not appear to improve dramatically simply with a greater quantity (amount) of observation, though stronger quality of observation (e.g., being a spouse or a close friend of the target) does improve accuracy.

These separate research areas developed with considerably different agendas and hold differentiated underlying philosophies about which method of measuring personality uncovers "truth." Self-rating replication/validation research presumes other-rating accuracy to be strong. In research on the accuracy of limited acquaintances, the accuracy of other-ratings is the focus of interest, and self-ratings serve as a criterion with presumably high accuracy. Finally, in research on the consistency of personality across situations, researchers have debated whether accuracy in self-ratings or other-ratings is even possible. Given the breadth of research across these three research streams, comprehensive review of research on other-ratings must shift from asking "Are they accurate?" to asking "how much?" "by which criteria?" "through what process?" "for which traits?" and "under what conditions?"

Fortunately, three well-recognized criteria have emerged for evaluating other-rating accuracy (Funder & West, 1993). Accurate personality judgments from an other-rater should predict judgments from a different other-rater (i.e., interrater reliability), self-ratings (i.e., self–other accuracy), and relevant behaviors and outcomes (i.e., criterion-related validity). Although factors other than accuracy may affect any one of these criteria (e.g., interrater reliability may reflect commonly held stereotypes and not purely accurate trait perception), finding that other-ratings satisfy this triad of accuracy criteria strongly supports their accuracy.

The Process of Accurate Judgment: Funder's Realistic Accuracy Model

Funder's (1995) realistic accuracy model presents an integrative conceptual framework addressing accuracy questions of "through what process?" and "under what conditions?" First, the realistic accuracy model posits a particular process for making accurate judgments of a target's personality traits. This process is first a function of the environment in which an other-rater observes the target. The environment must allow the target to express the trait (relevance) and allow the observer to perceive this trait expression (availability). However, accuracy depends not only on trait expression but also on accurate trait perception. Thus, observers must notice trait-relevant cues (detection) and appropriately assemble these cues to form an impression of the target (utilization). Accurate judgments can be formed only when all four conditions of the process are satisfied.

Funder (1995) noted that personality psychologists have traditionally focused on relevance and availability (RA) because they are fundamentally interested in how traits result in the expression of trait-relevant behavior. In contrast, social psychologists have traditionally focused on detection and utilization (DU) because these stages reflect the person perception process that forms an other-rater's judgment. Figure 1 diagrams the effects of RA and DU processes on the measurement of an other-rater's perception of a target's trait. First, RA processes determine the extent to which traits are generally expressed to others. What this expression is and how strongly it is linked to the underlying trait may vary across observers on the basis of the context of acquaintance (i.e., RA). Next, DU processes determine how strongly particular other-raters' perceptions align with that trait expression. Finally, measurement error affects measurement of an other-rater's perception. Note that the accuracy of self-perceptions of traits is similarly determined by RADU processes (though these may operate differently for self- vs. other-raters). Because self-perceptions can be gathered from only one rater (the target), however, separating the effects RA and DU processes is not possible for self-ratings.

The three accuracy criteria identified by Funder and West (1993) are important because they illustrate how strongly RADU occurs. Interrater reliability (the correspondence between other-raters from the same context of acquaintance) can be strong only if (a) the trait has been meaningfully expressed through RA, (b) DU processes have linked raters' perceptions strongly to trait expressions, and (c) other-raters' perceptions have been measured with relatively little measurement error.[1] Similarly, correlations between self-ratings and other-ratings can be strong only if (a) the trait has been expressed to others through strong RA, (b) others' perceptions align with the trait expression through strong DU, (c) self-perceptions also align with the trait through strong RADU, and (d) self- and other-perceptions have been measured with relatively little measurement error. Thus, accuracy criteria are critical for ascertaining the relative strength of RADU processes.

Additionally, for 90 years accuracy researchers have proposed, studied, and found many moderators ("under what conditions?") of this process of forming accurate personality judgments. Funder (1995) argued that these moderators can be grouped into four overarching determinants of accuracy: good judge, good target, good trait, and good information. "Good judge" describes individual differences in judges' ability to judge targets, whereas "good target" describes the relative ease with which others can rate a particular target. "Good trait" posits differences across traits in how easily they can be observed and interpreted. Finally, "good information" refers to the accuracy of the cues available for a particular trait. Note that these four accuracy moderators may function through different parts of the RADU process. For example, some traits may be good traits because they are more strongly linked to particular observable behaviors, such that the strength of these good traits lies in their relevance and availability. In contrast, traits may be "good" because they are easier for observers to perceive accurately. In these cases, the strength of these traits lies in their ease of detection and utilization.

Research on accuracy moderators has produced differential support for the four categories. Studies have been relatively successful in identifying good traits and good information for accuracy. However, comparison of studies of good judges and good targets is difficult because of difficulties in measuring who is a good judge and who is a good target. In addition to facing difficulties indexing other-rating accuracy itself, researchers in these fields have struggled to measure qualities predictive of being a good judge or a good target (Chaplin & Goldberg, 1984; Vernon, 1933). Thus, we focus here on the most consistent accuracy moderators whose research also lends itself to integration across studies: good trait and good information moderators.

[1] As one anonymous reviewer noted, raters from the same context may observe slightly different target behavioral sets (e.g., two coworkers do not observe the target in all of the same situations). Such differences in trait expression across observers could marginally reduce detection or utilization parameters because of actual differences in relevance or availability. However, these effects are generally theorized to diminish as opportunity to observe targets increases (Kenny, 2004), and studies have generally found little difference in agreement between those observing high and low overlap in behaviors (Kenny et al., 1994; Sullivan, 1995).

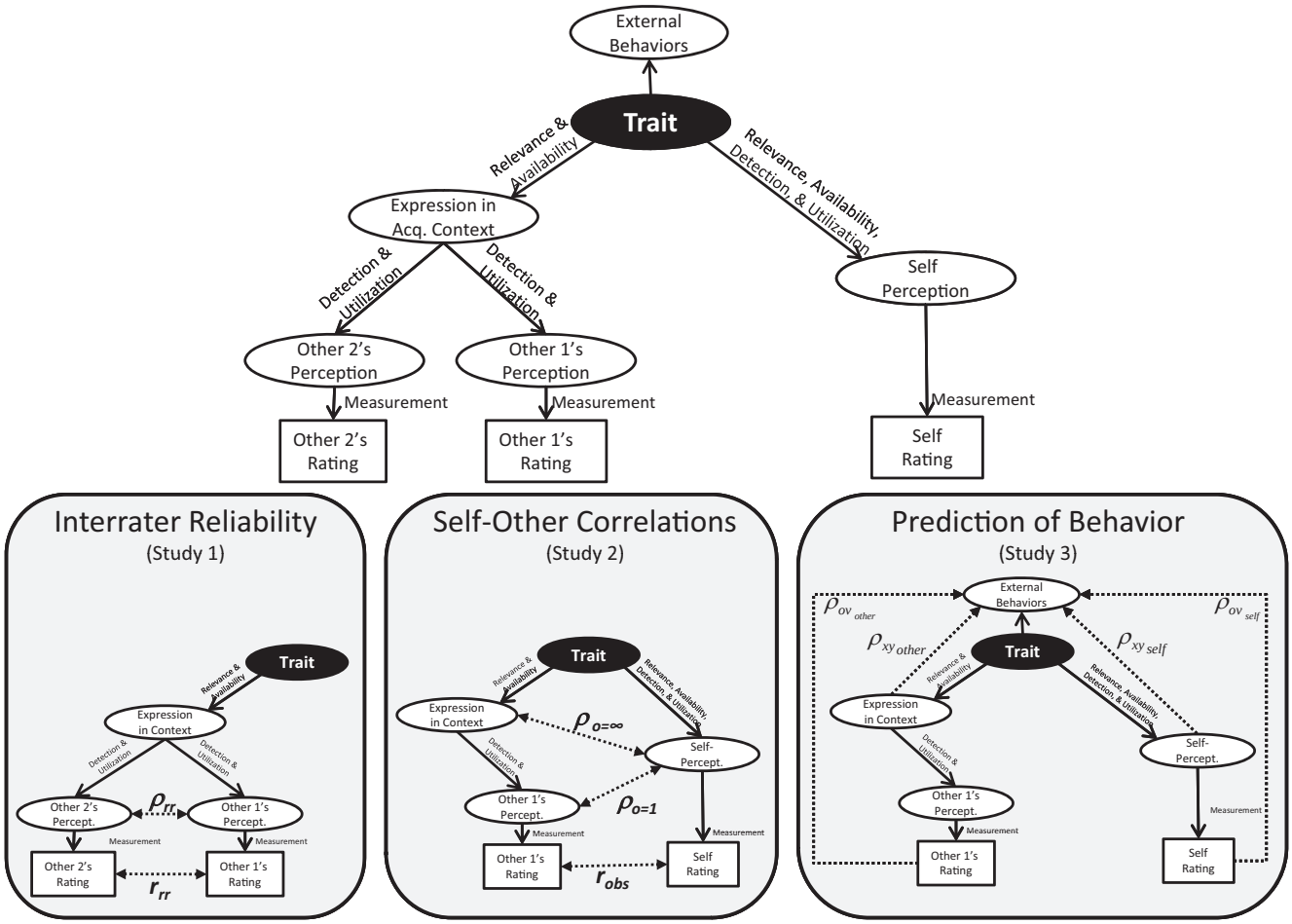

Figure 1. General measurement model aligning the realistic accuracy model's relevance, availability, detection, and utilization processes with Study 1, 2, and 3 accuracy criteria. Dashed lines indicate accuracy correlations. r_{rr} = observed interrater reliabilities at the scale level; ρ_{rr} = interrater reliabilities at the level of two observers' latent perception of the target; r_{obs} = correlation between self-ratings and an other's rating; $\rho_{0=1}$ = correlation between latent self-perceptions and the latent perception of one other-rater; $\rho_{o=\infty}$ = correlation between latent self-perceptions and trait expression observable to all other-raters (i.e., as though detection and utilization occurred perfectly); ρ_{ov} = operational validities for predicting latent external behaviors from self- or other-rating scales; ρ_{xy} = correlation of latent external behaviors with latent trait expressions perceived by self- and other-raters.

The Good Trait

Researchers have long believed some traits to be easier for other-raters to report accurately (John & Robins, 1993), and Funder's (1995) inclusion of the good trait moderator reflects this perspective. Funder suggested two critical dimensions for determining accuracy across ratings of traits: high visibility of traits and low evaluativeness of traits. Highly visible traits comprise tendencies externally expressed (e.g., behavior), whereas low-visibility traits comprise more internal tendencies not directly accessible to others (e.g., thoughts and feelings). Among the five factors, Extraversion stands out as a high-visibility trait: Tendencies to be socially outgoing, dominant, and energetic are linked to expressive social behaviors, and Extraversion measures describe more tendencies in behaviors than in thoughts or feelings (Zillig, Hemenover, & Dienstbier, 2002). Emotional Stability and Openness to Experience have the opposite orientation, predominantly describing tendencies in internal thoughts (Openness) and affective states (Emotional Stability) that are aspects of personality low in visibility (Zillig et al., 2002).

Traits high in evaluativeness are those for which great social value is placed on an individual's standing on the trait; social norms impose less judgment for nonevaluative traits. Because targets may try to conceal undesirable behaviors and highlight desirable behaviors, observers are likely to observe fewer genuine trait cues for evaluative traits (e.g., Agreeableness and the intellect component of Openness). In contrast, behaviors related to less evaluative traits (e.g., Extraversion and the experiencing component of Openness) are more likely to be attributed to differences in interests. Although traits' desirability and evaluativeness may vary as social and cultural norms vary, research in North American

samples has generally shown Extraversion to be the least evaluative and Agreeableness and the intellect aspect of Openness to be the most evaluative of the Big Five (John & Robins, 1993).

Empirical research has shown that accuracy differences across the five factors correspond closely to differences in visibility and evaluativeness. Extraversion, high in visibility and high in nonevaluativeness, has typically been the most accurately rated personality trait, especially when the other-raters had no prior knowledge of the target (e.g., Borkenau & Liebler, 1992; Connolly, Kavanagh, & Viswesvaran, 2007; John & Robins, 1993; Kenny et al., 1994). In addition, this research has shown that others' ratings of Agreeableness (high evaluativeness) typically have lower interrater reliability and lower self–other correlations.

Good Information

Funder's (1995) realistic accuracy model also specifies that accuracy in judging others' personality depends on having good information about the traits being rated. The effects of information quality have typically been studied in two ways. First, researchers have manipulated information by choosing others with specific relationships with the target (referred to here as information source moderators). For example, the information that a parent draws on to describe his or her child's personality represents a different behavioral set than that available to the targets' coworkers. Table 1 presents a taxonomy of information sources typically used in other-rating research, with six general types of information sources: family members, friends, cohabitators, work colleagues, incidental acquaintances, and strangers. These information source categories conform to lay distinctions between types of relationships, and they depict both different contexts for observing the target and different levels of acquaintance. Starzyk, Holden, Fabrigar, and MacDonald (2006) examined six dimensions proposed to underlie levels of acquaintance: duration, frequency of interaction, knowledge of goals, physical intimacy, self-disclosure, and social network familiarity. These dimensions reflect important differences in how target-related information is conveyed to other-raters. Table 1 shows the relative standing of information source categories both on overall acquaintance and on five of the six dimensions as rated by two independent raters (duration was measured as the median number of months of acquaintance reported in studies). Note that information sources are similarly rank ordered across most dimensions (frequency was the exception), an outcome mirrored in Starzyk et al.'s original findings. Thus, these six dimensions of acquaintance may be more parsimoniously described as two dimensions: frequency of interaction and interpersonal intimacy. Convention suggests that information sources with greater acquaintance across these dimensions would rate targets with greater accuracy. However, it is unclear which dimension is most critical for providing accurate trait information.

When other-ratings are collected from strangers, researchers have greater control over the type of information presented to other-raters. Table 2 presents a categorization of these differences in stimulus information presented to strangers (information types), indicating differences in media and procedures researchers have used to present target-related information (visual cues only, audio cues only, audio and visual, text/electronic communication, and personal object). Again, it is presumed that greater information across the stimuli would be related to greater accuracy, but it is unclear which set of information is most critical.

The previous sections have reviewed major ways in which researchers have manipulated accuracy moderators and suggested ways that these manipulations may enhance or decrease accuracy. We focus our analyses primarily on studying the magnitude and mechanisms through which two accuracy moderators (good traits and good information) affect the three accuracy criteria (interrater reliability, self–other correlations, and predictive validity). Representing the breadth of potential traits, information, and accuracy criteria is clearly beyond the purview of a single study. Even the strongest individual studies are unlikely to examine other-rating accuracy with more than two types of information sources or information types or to study the full range of accuracy criteria. The rare exceptions typically come from large-scale behavioral genetics databases or major personality research centers. Even in such large-scale data collection efforts, the span of targets, traits, and information sources/types cannot be fully represented, and many important research questions about the accuracy of other-ratings remain unanswered. Thus, single data collection efforts cannot afford a comprehensive understanding of accuracy, an observation Funder (1995) noted in proposing the realistic accuracy model. As a result, pursuing the best answers to these research questions necessitates a meta-analytic approach.

Meta-Analytic Database and Approach

To begin this series of meta-analyses, we used seven search strategies to locate studies collecting personality ratings from a nonself source: (a) using a search string in PsycINFO ([*personality or trait or temperament*] *and* [peer *or informant or spouse or friend or roommate or stranger or consensus or consensual validity or consensual validation or self–other agreement or zero-acquaintance or thin slices of behavior*]); (b) hand searching through a collection of over 200 psychological test manuals; (c) reviewing research bibliographies of three personality inventories that have other-report forms (the NEO Personality Inventory—Revised, the Personality Research Form, and the Six Factor Personality Questionnaire); (d) reviewing the reference sections of existing meta-analyses and summary articles on other-ratings of personality (Connolly et al., 2007; Kenny et al., 1994; McCrae et al., 2004); (e) manually searching relevant existing meta-analytic databases; (f) contacting researchers who have frequently used other-ratings to request unpublished data; and (g) reviewing the reference sections of articles located through Strategies 1–6 for potential contributing data sources. This process produced 596 studies that were read for potential inclusion in the database.

A study had to meet several criteria to be included in the database. First, the study must have collected ratings from a nonself source describing a "normal" personality trait (i.e., traits with greater variation in the general population than in clinical populations). Because abnormal personality traits tend not to be normally distributed in nonclinical populations and do not generally align with a single Big Five trait, omitting studies measuring only abnormal traits (e.g., Oltmanns, Turkheimer, & Strauss, 1998) yields results that more purely reflect normal person-perception processes. Second, the study must have presented the study's sample size and one of the following: (a) interrater reliability of other-ratings of personality traits, (b) correlations be-

Table 1
Hierarchical Information Source Taxonomy of Other-Raters

Information source	*k*	Defining characteristics	Overall acquaintance	Duration (months)	Frequency	Knowledge of goals	Physical intimacy	Self-disclosure	Social network familiarity	Example
Family	50	Long-term, close relationship with target; generally linked to target genetically or through marriage	High	211.20	Medium	Medium high	High	High	Medium high	Children/significant others (Hoffman et al., 2007)
Spouse/life partner	31	Is married to target or in a long-term committed, cohabitating relationship	Very high	228.00	Very high	Very high	Very high	Very high	Very high	Spouse (Watson & Humrichouse, 2006)
Parents/elder relatives	13	Is a parent/older relative who has known the target all of his or her life; authority over target as a child	High	216.83	Low	Medium	High	Medium high	Medium	Parents and grandparents (Malloy et al., 2004)
Siblings	4	Brother/sister; has known target for most of his/her life	High	202.60	Low	Medium	High	Medium high	Medium	Same-sex twin (Heath et al., 1992)
Friends	86	Relationship with target based on mutual liking	High	39.42	Medium	Medium high	Medium high	Medium high	High	Friend, roommate, or boyfriend/girlfriend (Colvin & Funder, 1991)
Dating partners	4	Romantic relationship with target, shorter duration/less committed than spouses/life partners	High	54.00	Medium	High	Very high	High	High	Significant other (O'Connor et al., 2001)
Best/close friend	11	Friends who are specified as being especially close to target	Very high	35.24	Medium	Very high	High	Very high	Very high	Peer most familiar to target (Ino-Oko & Matsui, 1977)
Friend/close acquaintance	57	General friends, no other criteria specified	High	31.08	Medium low	Medium	Medium high	Medium high	High	Close acquaintance (Piedmont et al., 2000)
Peer at school	8	Peers of targets who are also students	Medium	30.48	Medium	Medium	Medium	Medium low	Medium high	Peer (Kenrick & Stringfield, 1980)
Cohabitators	39	Live with the target; nonfamily members	Medium	7.00	High	Medium	Medium high	Medium high	High	Roomate/acquaintance on same floor of dorm (Hase & Goldberg, 1967)
Roommate	17	Share room/close living quarters with the target	Medium high		Very high	Medium high	Medium high	High	High	Cohabitor (couples or roomates; Borkenau & Liebler, 1993)
Dorm/housemate	17	Live in the same residence as target	Medium low		Medium high	Medium low	Medium high	Medium	Medium high	Dorm neighbors (Ashton & Goldberg, 1973)
Work colleague	12	Know target based on employment context	Medium low	30.12	Medium high	Medium	Medium low	Low	Medium low	Colleague (Cooke et al., 1987)
Supervisor	3	Target reports directly to other	Medium low	40.80	High	Medium	Low	Very low	Medium low	Work supervisors (Nilsen, 1995)
Coworker	6	Colleague at same level as target	Medium high	28.80	High	Medium high	Medium low	Medium	Medium high	Coworker (police officer partners; Mellor, 1986)
Subordinate	1	Other reports directly to target	Medium low	21.00	High	Medium	Low	Low	Medium low	Work subordinates (Nilsen, 1995)

(*table continues*)

Table 1 (*continued*)

Information source	*k*	Defining characteristics	Overall acquaintance	Duration (months)	Frequency	Knowledge of goals	Physical intimacy	Self-disclosure	Social network familiarity	Example
Customer	3	Individuals soliciting professional services from target	Low	15.37	Medium	Low	Low	Very low	Low	Customer (Mount et al., 1994)
Reference	2	Individuals listed as a reference by target	Medium high	8.28	Low	High	Medium	Very low	Low	Reference (Bradley, 1997)
Incidental acquaintance	18	Short-term acquaintance with target, based on common activity	Low	1.75	Medium	Low	Medium low	Medium low	Low	
Assigned classmate	12	Knew target only through common coursework	Low	3.00	Medium	Low	Medium low	Medium low	Medium low	Classmate (Bratko et al., 2006)
Group comember	3	Acquainted with target through co-membership in a particular group	Low	1.75	Medium	Low	Medium	Medium	Very low	Expedition leader (Watts et al., 1994)
Stranger	64	Unacquainted with target prior to study beginning	Very low	.00	Very low	Very low	Very low	Low	Very low	Confederates (Borkenau et al., 2004)

Note. Ratings of information source categories on overall acquaintance and Starzyk et al.'s (2006) dimensions based on rankings by two independent raters. *k* = number of samples in our database contributing data for each information source type. Duration (months) of acquaintance is based on the median number of months of acquaintance reported by studies using other-ratings of personality.

tween self-ratings and other-ratings of personality traits, or (c) correlations of other-ratings of personality with indicators of academic performance, work performance, or trait first impressions conveyed to strangers (or information allowing for computing a, b, or c). Studies in which targets were young children (age 14 or younger) or in which samples were chosen on the basis of psychopathology, trauma, or brain damage were excluded to avoid including targets for whom personality development/change was more likely. Finally, we excluded studies that collected other-ratings through peer nomination procedures in which a group of raters (e.g., a sorority) nominated the group members who was highest or lowest on a particular trait. The number of nominations received became the target's score on the trait. These peer nomination measurement procedures provide useful information only for individuals at the extreme ends of the distribution, and including correlations from such procedures would downwardly bias meta-analytic estimates.

In considering accuracy criteria, one further distinction must be made. The most common and most familiar approach to indexing rank-order consistency is to correlate ratings from the two sources for one trait across many targets. In contrast, some researchers have used profile correlations to index accuracy, with ratings across many traits correlated across rating sources for one target. Such profile correlations reflect consistency in the rank ordering of traits for a particular individual but are typically averaged across individuals in different conditions (e.g., the mean profile correlation for acquainted vs. unacquainted other-raters). Kenny and Winquist (2001) referred to the former approach as "nomothetic accuracy" (accuracy across a sample of individuals) and the latter approach as "idiographic accuracy" (accuracy for one individual across a sample of traits). Nomothetic and idiographic accuracy have substantively different interpretations, and the two approaches are not interchangeable. Because other-rating accuracy for the particular trait being rated is of more substantive interest to personality researchers across many domains of psychology, nomothetic accuracy is appropriate. Thus, we included only nomothetic forms of accuracy in our analyses.

In total, the database consisted of data on 44,718 target individuals assessed as part of 263 independent samples and presented in 188 published and unpublished sources. For each effect size, the particular information source was coded according to the taxonomy in Table 1, the information type for strangers was coded according to the taxonomy in Table 2, and the personality trait being rated was coded according to the Big Five taxonomy presented in Hough and Ones (2001). Hough and Ones' taxonomy, which aligns specific personality measures with the five factors, has been used in numerous other personality meta-analyses (e.g., Dudley, Orvis, Lebiecki, & Cortina, 2006; Foldes, Duehr, & Ones, 2008; Trapmann, Hell, Hirn, & Schuler, 2007). All coding was conducted by the first author (a junior professor whose primary research interests are personality and meta-analysis) and was subsequently reviewed by the second author (a full professor with extensive experience in conducting meta-analyses involving personality variables, many of which were published in peer-reviewed scientific literature). Any discrepancies were resolved by discussion.

We used Hunter and Schmidt's (2004) psychometric meta-analytic approach for synthesizing study results. Hunter–Schmidt meta-analysis is a random-effects meta-analysis model, meaning

Table 2
Taxonomy of Types of Information Given to Strangers

Information type category	*k*	Defining characteristics	Example
Visual cues only	27		
Still visual	10	Minimal actions; appearance cues only	• Photograph of head and shoulders plus left/right profile (Shevlin et al., 2003) • Video of still target (Borkenau & Liebler, 1992)
Silent nonverbal	19	Observe target behavior with no oral communication	• Completed measures in the same room; no talking (Ambady et al., 1995) • Muted videotapes of targets being interviewed (Gangestad et al., 1992)
Audio cues only	10	No visual information presented; hear target's voice	• Audio of targets reading alphabet (Berry, 1991) • Three 20-s audio recordings of targets in mock jury discussion (Scherer, 1972)
Activity (audio and visual)	30	Observe both visual and audio information	
Prescribed behavior (scripted/posed)	3	Behavior is completely instructed by experimenter	• 15-s video stating name and study design number (Barrick et al., 2000) • Film of targets entering room, walking, sitting, reading a weather forecast, standing, and leaving (Borkenau & Liebler, 1993)
Natural behavior (unscripted)	30	Behavior occurs relatively naturally, though experimenters may give general instructions (e.g., talk about a particular topic)	• Define neologism and argue why that definition is appropriate (Borkenau et al., 2004) • Talk with opposite-sex participant and debate capital punishment (Kolar et al., 1996)
Personal object	5	Do not actually observe target's behavior; observe some representative object of the target	• View personal living spaces (Gosling et al., 2002) or personal web pages (Marcus et al., 2006) • Listen to target's 10 favorite songs (Rentfrow & Gosling, 2006)
Text/electronic communication	7	Do not directly observe target; read text from target or exchange writing with target	• Read text about targets' deepest thoughts and feelings about themselves (Bosson et al., 2000) • Online chat session with target participating in two online activities (Rouse & Haas, 2003)

Note. *k* indicates the number of independent samples in our database using each information source category.

that it does not assume an identical population parameter across all studies (fixed-effects models do assume identical population parameters). Such random-effects models more accurately estimate confidence intervals around point estimates and better avoid Type I errors in detecting moderators than do fixed-effects models (Hunter & Schmidt, 2000; Schmidt, Oh, & Hayes, 2009). In addition, Hunter–Schmidt meta-analysis accounts for the effects of statistical artifacts (e.g., measurement error) that reduce study correlations and introduce variability across study results not attributable to true variability in population correlations. That is, correlations vary across studies not only because studies may come from different populations or because of sampling error but also because studies differ in the reliability of their measures. In Hunter–Schmidt meta-analysis, means and variances for these effects of measurement artifacts are corrected for with artifact distributions. Just as a meta-analytic database culls a distribution of correlations reported in studies, an artifact distribution is a set of artifact statistics (e.g., reliabilities) reported in studies. These artifact distributions effectively model both the mean and the variability in measurement artifacts across studies. In Hunter–Schmidt meta-analysis these artifact distributions are used to correct mean observed correlations (ρ) to more accurately reflect the correlations among constructs. Similarly, artifact distributions are used to correct standard deviations of correlations (SD_ρ) to reflect the variability in study correlations due to true variability in population parameters and not to measurement artifacts.[2]

The ability to correct correlations for measurement artifacts is an especially important feature of Hunter–Schmidt meta-analysis for this research domain (Schmidt & Hunter, 1999; Schmidt, Le, & Ilies, 2003). Figure 1 directly illustrates how different forms of measurement error affect observed correlations seen in accuracy research. In particular, low test–retest reliability weakens the link between other-raters' perceptions and ratings, and low interrater reliability weakens the link between trait expression and an other-rater's perception. Whereas other forms of meta-analysis would estimate only observed correlations that are affected by these

[2] The meta-analyses applied in Studies 1, 2, and 3 produced a common set of meta-analysis statistics. First, the number of independent samples (*k*) and the total sample size summed across independent samples (*N*) serve as indicators of the amount of data contributing to the meta-analysis. The observed mean correlation ($\bar{r}$) and standard deviation of correlations (SD_{obs}) describe basic properties of the observed distribution of correlations drawn from independent samples. Estimates of variance in correlations that would be expected due to sampling error and unreliability are calculated. Variance due to these artifacts is subtracted from the observed variance to calculate the residual standard deviation in correlations (SD_{resid}). In addition, observed correlations are corrected for attenuation due to statistical artifacts. These corrections are used to estimate the mean true score correlation (ρ) and standard deviation of true score correlations (SD_ρ). Thus, although both $\bar{r}$ and ρ represent estimates of the population parameter from pooling across samples' correlations, ρ has been corrected for statistical artifacts whereas $\bar{r}$ has not. Around these mean true score correlations (ρ), confidence intervals can be calculated that indicate boundaries within which the population correlation can be expected to fall ($Conf_L$ and $Conf_U$).

forms of measurement error, Hunter–Schmidt meta-analysis permits estimation of correlations at multiple levels of Figure 1's measurement model. Such estimates better disentangle the effects of RA, DU, and test reliability on other-rating accuracy, as we describe within each meta-analysis.

To detect cross-study moderator effects, Hunter–Schmidt meta-analysis compares point estimates for moderator groups (e.g., comparing corrected mean self–other correlations for friend other-raters vs. stranger other-raters). Meta-analytic investigations of true moderator effects should show (a) initially large SD_ρ values when moderator groups are pooled together, (b) differences across moderator groups' corrected mean correlations (ρ), and (c) reductions in SD_ρ values when moderator groups are separated. Because we sought to evaluate other-rater accuracy with a meta-analytic model that allowed population parameters to vary across studies and that corrected observed correlations for statistical artifacts, Hunter–Schmidt methods were most appropriate. Further methodological detail specific to each study follows within each study's Method section.

Study 1: Other-Rater Accuracy as Consensus Among Raters (Interrater Reliability)

In Study 1, we focused on the extent to which ratings by other-raters correspond when rating a common target. Two existing quantitative reviews have summarized a small set of interrater reliabilities. Kenny et al. (1994) meta-analyzed interrater reliabilities of other-ratings of personality measures from 32 studies using Kenny's social relations model (Kenny & La Voie, 1984; Malloy & Kenny, 1986). The social relations model apportions variance in trait ratings that is associated with variance due to targets (analogous to interrater reliability), variance due to raters, and variance due to the unique relationship between targets and raters. Single-rater interrater reliabilities for strangers were generally quite low (usually less than .10) for all traits except Extraversion (approximately .30). These reliabilities were higher for long-term acquaintances, with all traits showing single-rater interrater reliabilities in the .25–.30 range. In another, more recent summary, Connolly et al. (2007) reported a brief table of interrater reliabilities for each of the Big Five as a sidebar to their meta-analysis of self–other convergence of personality traits. However, because these reliability coefficients reflect the reliability of a composite of others' ratings (with reliabilities generally ranging from .69 to .81) rather than the interrater reliability of a single other-rater, these values are not comparable to those presented in Kenny et al. (1994).

Beyond studies included in the Kenny et al. (1994) and Connolly et al. (2007) meta-analyses, a considerably larger set of studies examining interrater reliability is available that is yet to be analyzed, synthesized, and interpreted. In total, Study 1 is based on 1,510 interrater reliability coefficients from 114 independent samples. Perhaps more noteworthy, this larger set of studies permits finer distinctions among information sources. For example, among the long-term acquaintances category in Kenny et al., family members (who may have had a lifespan to observe and form impressions about the target) would likely have stronger interrater reliabilities than would work colleagues. In addition, this larger set of studies permits distinctions among limited acquaintance studies based on the type of stimulus (information type) presented about targets as well. As yet, no research synthesis has examined the effect that differences in the type of information presented have on interrater reliability. Thus, in Study 1 information source and information type were more broadly examined as potential moderators of interrater reliability.

Study 1: Method

We collected interrater reliabilities if raters came from the same Level 1 information source category (e.g., no reliabilities were included when raters were a mix of friends or colleagues). Studies reported interrater reliabilities with a variety of statistics, including intraclass correlations, coefficient alphas (in which raters were treated as items), and percentages of variance accounted for by target effects in social relations model analyses (Kenny & La Voie, 1984). Such heterogeneity of reliability statistics was inevitable, given the breadth of types of other-raters used in this research domain, and is common to nearly all fields in which interrater reliabilities have been meta-analyzed. For example, it would be impossible in a study examining the interrater reliability of parents' ratings of personality to have a sample size of any magnitude if a fixed set of parents rated every target (i.e., a block design; Kenny & Albright, 1987). Although these reliability statistics vary in their calculation procedures, single-rater reliabilities derived from each of these statistics closely reflect the correlation that would be observed if ratings from two randomly chosen raters were correlated. Given this common core, these forms of interrater reliability coefficients were included to allow for coverage of the full breadth of other-rater categories.

Meta-analyses were conducted to examine and explain variability across studies in interrater reliabilities. However, a major determinant of the reliability of a composite of other-raters is the number of raters used, and studies varied widely in the number of raters describing each target. The number of raters must be held constant to facilitate comparisons in examining variability in interrater reliabilities. Therefore, we meta-analyzed single-rater reliabilities in the database (reliabilities either reported directly by authors or estimated during study coding with the Spearman–Brown formula from the interrater reliability of k raters). These single-rater reliability estimates reflect the correlation that would be expected between two randomly selected raters' descriptions of a target. Thus, these single-rater reliability estimates were treated as correlations in estimating sampling error variance and in making corrections for instability of measurement.

Interrater reliabilities reflect the effects of two types of measurement error aside from purely random error: error due to true discrepancies in how raters view the target (rater-specific error, or, in Figure 1, errors in detection and utilization) and error due to rater factors unique to the particular time at which ratings are collected, such as the rater's mood (transient error). Framed another way, if two ratings from the same other-rater at two different time points do not agree precisely, it is unlikely that ratings from two different other-raters will agree precisely. Because interrater reliabilities reflect these sources of measurement error, they have been referred to as coefficients of equivalence and stability (Schmidt et al., 2003).

However, in comparing interrater reliabilities across moderators, the interest is primarily in understanding the extent of other-raters' truly discrepant views of the target (i.e., only rater-specific error). This true rater-specific error can be estimated by correcting

observed interrater reliabilities for test–retest unreliability in both other-raters. Indicated as ρ_{rr} in Figure 1, these corrections estimate the correlation between other-raters' latent perceptions rather than observed measures. These corrected correlations reflect the correlations that would be observed between two other-raters if the perceptions of targets' traits could be measured at many (infinite) time points (i.e., with measurement error due to instability removed from these correlations). To apply these corrections, we created artifact distributions of test–retest reliabilities of other-ratings separately for each of the Big Five.[3]

Study 1: Results and Discussion

Good trait moderators. First, we conducted single-rater interrater reliability meta-analyses of each personality trait with all categories of raters included. The results of this meta-analysis are presented in Table 3. With across-information sources combined, the average single-rater reliabilities were modest, typically ranging from $\bar{r}_{rr} = .32$ to $\bar{r}_{rr} = .43$. These estimates increased when corrected for test–retest reliability ($.39 < \rho_{rr} < .51$). The nonoverlapping confidence intervals indicate that Extraversion, the most visible trait, yielded the highest interrater reliability, followed by Conscientiousness. However, interrater reliabilities for Emotional Stability, Openness, and Agreeableness were lower and produced mostly overlapping confidence intervals. This finding for Extraversion is consistent with results of Kenny et al. (1994), where Extraversion was the trait that showed the highest proportion of target variance among ratings, but it highlights that Conscientiousness may also be an especially good trait.

Note also that the magnitude of these single-rater interrater reliabilities, even when corrected for test–retest unreliability, is considerably smaller than those shown in other domains. For example, ratings of performance and abilities generally yield single-rater interrater reliabilities between .52 and .85 (Connelly & Ones, 2008; Conway, Jako, & Goodman, 1995; Salgado & Moscoso, 1996; Viswesvaran, Ones, & Schmidt, 1996). These findings highlight the particular importance in personality research of combining multiple raters to overcome rater idiosyncrasies that reduce interrater reliabilities.

Good information moderators. There was moderate variability around SD_ρ values, suggesting some potential room for moderators to affect interrater reliabilities. Thus, information source was examined as a potential moderator of interrater reliabilities for each personality trait. Across the five factors, family and friends (sources high on frequency and intimacy dimensions) generally had the strongest interrater reliabilities, mostly between $\rho_{rr} = .40$ and $\rho_{rr} = .55$. The lower bound of family and friends' confidence intervals generally excluded less intimately acquainted information sources, namely, cohabitators (who, for interrater reliabilities, were typically dorm or housemates rather than roommates), work colleagues, and incidental acquaintances. For most traits, interrater reliabilities for cohabitators and incidental acquaintances did not even exceed the lower bound of the confidence interval for strangers. This is likely because strangers in the included studies had observed targets in a limited but common set of situations, whereas cohabitators and incidental strangers typically observed the target behaving regularly in a variety of situations that may not have been common across observers. Thus, for rating many traits, this common set of situations for strangers appears to yield stronger interrater reliabilities than do the broader observation spectra of cohabitators and incidental strangers. To some extent, strangers' ratings likely also reflect a convergence on commonly held stereotypes rather than true personality. Indeed, although it would seem more likely for strangers' ratings to be the least accurate, the weighted accuracy model (Kenny, 1991) and PERSON model (Kenny, 2004) predict that both this overlap in behaviors observed and the use of stereotypes would enhance the interrater reliability. Still, these overlap and stereotype effects may not enhance accuracy as measured by other criteria (such as correlations with self-ratings and predictions of behavior).

Good trait × Good information moderators. The pattern of higher interrater reliabilities for family and friends was less pronounced for two of the Big Five: Extraversion and Agreeableness. In the case of Extraversion, interrater reliabilities were relatively high for all information sources, with even strangers and incidental acquaintances both showing interrater reliabilities of $\rho_{rr} = .48$. This pattern of findings suggests that cues to rate Extraversion are readily apparent to observers, and increased acquaintance with the target may only minimally improve the correspondence between two observers. On the other hand, interrater reliabilities for Agreeableness (the most evaluative trait) were relatively uniformly low, with the typical advantage for family and friends being somewhat less pronounced for Agreeableness. Thus, it may be the case that how "likable" (Agreeable) a rater finds a target to be depends largely on how well the rater "likes" the target, suggesting that Agreeableness perceptions may be more idiosyncratic to the rater's particular relationship with and appeal to the target.

Several important trends in interrater reliabilities can be noted across these moderator analyses for information source categories. First, interrater reliabilities were generally highest when the others providing ratings were family members or friends. When trait ratings came from cohabitators, work colleagues, or incidental acquaintances, interrater reliabilities tended to be lower. In many cases, they were lower than those for ratings from strangers. This pattern of reliabilities is consistent with interpersonal intimacy rather than interaction frequency as the critical dimension of acquaintance for accuracy. However, the findings for strangers indicate that substantial overlap in behaviors observed can buffer the effects of low intimacy on interrater reliabilities. Separating analyses by information source generally reduced the variability around mean interrater reliability estimates, suggesting information source to be a true moderator. However, considerable variability (SD_ρ) around average interrater reliabilities for strangers remained.

Information-type moderators of strangers' interrater reliabilities. Given this remaining variability in strangers' interrater reliabilities, the large number of independent samples for strangers

[3] These artifact distributions had the following means and standard deviations of the square root of test–retest reliabilities: Emotional Stability: mean $\sqrt{r_{xx}} = .91$, $SD_{\sqrt{r_{xx}}} = .03$ ($k = 8$; $N = 733$); Extraversion: mean $\sqrt{r_{xx}} = .92$, $SD_{\sqrt{r_{xx}}} = .03$ ($k = 8$; $N = 733$); Openness: mean $\sqrt{r_{xx}} = .90$, $SD_{\sqrt{r_{xx}}} = .03$ ($k = 5$; $N = 663$); Agreeableness: mean $\sqrt{r_{xx}} = .90$, $SD_{\sqrt{r_{xx}}} = .05$ ($k = 8$; $N = 733$); Conscientiousness: mean $\sqrt{r_{xx}} = .90$, $SD_{\sqrt{r_{xx}}} = .03$ ($k = 5$; $N = 663$). Test–retest reliabilities were collected from studies collecting other-ratings from the same observer at two time points less than one year apart.

Table 3
Meta-Analysis of Single-Rater Interrater Reliabilities, by Information Source

Trait and source	k	N	$\bar{r}_{rr}$	SD_{obs}	SD_{resid}	ρ_{rr}	SD_{ρ}	$Conf_L$	$Conf_U$	FS k
Emotional Stability	72	13,458	.33	.14	.13	**.40**	.15	.37	.41	403
Family	5	774	.37	.16	.14	**.44**	.17	.37	.51	32
Friends	16	3,102	.38	.11	.08	**.45**	.10	.42	.49	106
Cohabitators	4	1,021	.20	.07	.04	**.24**	.04	.17	.31	12
Work colleagues	5	682	.28	.12	.08	**.34**	.10	.25	.41	23
Incidental acquaintances	5	338	.18	.07	.00	**.22**	.00	.09	.33	13
Strangers	41	3,723	.23	.15	.12	**.27**	.14	.24	.31	148
Extraversion	82	12,438	.43	.13	.11	**.51**	.13	.49	.52	623
Family	5	774	.45	.08	.04	**.53**	.05	.46	.59	40
Friends	16	3,111	.46	.08	.05	**.55**	.06	.51	.57	131
Cohabitators	7	1,101	.28	.08	.03	**.34**	.03	.26	.39	32
Work colleagues	6	1,238	.37	.12	.10	**.44**	.12	.38	.49	38
Incidental acquaintances	7	466	.40	.13	.07	**.48**	.08	.38	.56	49
Strangers	49	4,238	.40	.17	.14	**.48**	.16	.44	.50	343
Openness	53	7,990	.32	.13	.11	**.39**	.14	.38	.42	286
Family	2	185	.38	.07	.00	**.47**	.00	.31	.62	13
Friends	9	2,077	.43	.05	.00	**.53**	.00	.49	.58	68
Cohabitators	3	939	.21	.03	.00	**.26**	.00	.19	.34	10
Work colleagues	5	928	.29	.11	.08	**.36**	.10	.29	.43	24
Incidental acquaintances	5	338	.20	.09	.00	**.25**	.00	.12	.38	15
Strangers	31	3,601	.30	.17	.15	**.37**	.18	.34	.41	155
Agreeableness	83	10,689	.32	.14	.12	**.40**	.15	.37	.42	448
Family	5	774	.25	.18	.16	**.31**	.20	.23	.39	20
Friends	20	3,263	.34	.11	.08	**.43**	.09	.38	.46	116
Cohabitators	8	1,172	.33	.06	.00	**.41**	.00	.34	.47	45
Work colleagues	6	1,238	.29	.07	.02	**.37**	.03	.29	.42	29
Incidental acquaintances	5	338	.24	.07	.00	**.30**	.00	.17	.42	19
Strangers	48	4,094	.27	.16	.13	**.33**	.15	.30	.37	211
Conscientiousness	64	11,523	.36	.13	.11	**.44**	.14	.42	.46	397
Family	5	774	.35	.17	.15	**.43**	.19	.35	.51	30
Friends	20	3,394	.37	.08	.04	**.46**	.04	.42	.49	128
Cohabitators	8	1,071	.26	.06	.00	**.32**	.00	.25	.39	34
Work colleagues	18	2,400	.32	.09	.04	**.40**	.05	.35	.44	97
Incidental acquaintances	5	338	.26	.17	.13	**.32**	.16	.19	.44	21
Strangers	35	3,466	.28	.15	.11	**.34**	.14	.31	.38	161

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Corrected interrater reliabilities (ρ_{rr}) are presented in boldface for emphasis. k = number of independent samples contributing data; N = total sample size; $\bar{r}_{rr}$ = mean observed interrater reliability coefficient; SD_{obs} = observed standard deviation of interrater reliability coefficients; SD_{resid} = standard deviation of interrater reliability coefficients after accounting for variability due to sampling error and test–retest unreliability; ρ_{rr} = mean interrater reliability corrected for the test–retest unreliability of both raters; $SD_{\rho} = SD$ of corrected interrater reliability accounting for variability from sampling error and test–retest unreliability; $Conf_L$ and $Conf_U$ = lower and upper bounds of 95% confidence interval around ρ_{rr}; FS k = fail-safe k, the number of studies where $r = .00$ that must be located to make $r = .05$.

permitted follow-up analyses examining information-type moderators. Table 4 presents these results. Audio plus visual activities tended to have the strongest interrater reliabilities (the lower bounds of their confidence intervals almost always excluded other cues). These higher interrater reliabilities were more pronounced when behaviors were relatively natural rather than prescribed. Similarly, across all traits, text and electronic communication produced the lowest interrater reliabilities. These findings suggest that these media are likely the worst for creating impressions of targets' personality, which is particularly concerning given the rise of electronic communication among relative strangers as a basis for participating in social networks and developing romantic relationships. Visual cues, audio cues, and personal object information types tended to have interrater reliabilities in between those of audio plus visual activities and text/electronic communication.

Particular information types had relative advantages for particular traits. For Emotional Stability, Extraversion, and Agreeableness, audio cues had interrater reliabilities falling within the confidence interval of the best cues of activities (audio plus visual). These findings are consistent with research on the "loud voice of Extraversion" and the "shaky voice of Neuroticism" (Scherer, 1978) and suggest this stream may be well complemented by studying the gruff voice of Disagreeableness.

For Openness and Conscientiousness, personal object cues had interrater reliabilities falling within the confidence interval of activities (audio + visual). For Openness, personal object cues illustrate how individuals decorate offices and bedrooms or choose favorite songs. These cues yield information about the targets' aesthetic sensibilities, traditionalism, and thoughtfulness. For Conscientiousness, personal object cues provide salient indicators of the orderliness component of Conscientiousness in how targets have organized rooms and web pages. Thus, visibility in Openness and Conscientiousness may not occur as much in direct behavior but rather in how an individual creates his or her environment.

Table 4
Information Type Moderators Meta-Analysis of Strangers' Single-Rater Interrater Reliabilities

Trait and information source	*k*	*N*	$\bar{r}_{rr}$	SD_{obs}	SD_{resid}	ρ_{rr}	SD_{ρ}	$Conf_L$	$Conf_U$	FS *k*
Emotional stability	*41*	*3,723*	*.23*	*.15*	*.12*	***.27***	*.14*	.24	.31	148
Visual cues only	18	1,202	.15	.11	.00	**.18**	.00	.11	.24	36
Still visual	8	371	.25	.28	.24	**.29**	.29	.18	.41	32
Silent nonverbal	11	926	.13	.09	.00	**.16**	.00	.08	.23	18
Audio cues only	9	315	.32	.14	.00	**.38**	.00	.26	.50	49
Activity (audio + visual)	17	2,336	.30	.15	.13	**.36**	.15	.31	.40	85
Prescribed behavior	3	267	.22	.06	.00	**.26**	.00	.12	.39	10
Natural behavior	15	2,136	.32	.16	.14	**.38**	.16	.34	.43	81
Personal object	5	411	.15	.05	.00	**.18**	.00	.06	.29	10
Text/electronic communication	4	243	.09	.09	.00	**.11**	.00	−.04	.25	3
All Extraversion	*49*	*4,238*	*.40*	*.17*	*.14*	***.48***	*.16*	.44	.50	343
Visual cues only	20	1,331	.30	.12	.03	**.35**	.04	.29	.41	100
Still visual	9	393	.30	.11	.00	**.35**	.00	.24	.46	45
Silent nonverbal	14	1,187	.30	.12	.07	**.36**	.08	.29	.41	70
Audio cues only	10	393	.45	.25	.21	**.53**	.24	.43	.62	80
Activity (audio + visual)	19	2,388	.48	.11	.09	**.57**	.10	.53	.60	163
Prescribed behavior	3	267	.45	.06	.00	**.53**	.00	.41	.64	24
Natural behavior	16	2,124	.50	.10	.07	**.59**	.09	.55	.62	144
Personal object	5	411	.30	.07	.00	**.35**	.00	.25	.45	25
Text/electronic communication	4	243	.23	.10	.00	**.28**	.00	.13	.41	14
All Openness	*31*	*3,601*	*.30*	*.17*	*.15*	***.37***	*.18*	.34	.41	155
Visual cues only	15	1,270	.19	.16	.13	**.23**	.16	.17	.30	42
Still visual	3	249	.23	.03	.00	**.29**	.00	.14	.43	11
Silent nonverbal	12	999	.14	.15	.11	**.18**	.13	.10	.25	22
Audio cues only	3	245	.19	.12	.06	**.24**	.08	.08	.38	8
Activity (audio + visual)	13	2,129	.28	.13	.11	**.35**	.14	.30	.40	60
Prescribed behavior	2	200	.27	.07	.00	**.33**	.00	.17	.49	9
Natural behavior	11	1,929	.29	.14	.12	**.36**	.14	.31	.41	53
Personal object	5	412	.42	.12	.07	**.52**	.09	.42	.62	37
Text/electronic communication	3	237	.11	.08	.00	**.14**	.00	−.02	.29	4
All Agreeableness	*48*	*4,094*	*.27*	*.16*	*.13*	***.33***	*.15*	.30	.37	211
Visual cues only	24	1,373	.17	.14	.04	**.22**	.05	.15	.27	58
Still visual	8	357	.23	.04	.00	**.29**	.00	.16	.40	29
Silent nonverbal	19	1,265	.15	.15	.08	**.19**	.10	.12	.25	38
Audio cues only	10	393	.35	.28	.24	**.43**	.29	.32	.54	60
Activity (audio + visual)	19	2,424	.31	.12	.09	**.39**	.11	.34	.43	99
Prescribed behavior	3	267	.24	.03	.00	**.30**	.00	.15	.43	11
Natural behavior	16	2,160	.31	.12	.09	**.39**	.10	.34	.43	83
Personal object	5	412	.27	.06	.00	**.34**	.00	.22	.44	22
Text/electronic communication	3	237	.18	.14	.09	**.23**	.11	.07	.37	8
All Conscientiousness	*35*	*3,466*	*.28*	*.15*	*.11*	***.34***	*.14*	.31	.38	161
Visual cues only	14	939	.19	.13	.06	**.24**	.07	.16	.31	39
Still visual	3	249	.2	.14	.09	**.25**	.11	.10	.39	9
Silent nonverbal	14	939	.19	.13	.05	**.24**	.06	.16	.31	39
Audio cues only	5	229	.25	.08	.00	**.30**	.00	.15	.45	20
Activity (audio + visual)	15	2,260	.35	.13	.11	**.43**	.13	.39	.48	90
Prescribed behavior	3	267	.27	.10	.00	**.34**	.00	.19	.47	13
Natural behavior	12	1,996	.35	.13	.11	**.43**	.14	.38	.48	72
Personal object	5	412	.33	.11	.05	**.41**	.06	.30	.51	28
Text/electronic communication	3	237	.13	.10	.00	**.16**	.00	.00	.31	5

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Corrected correlations are presented in boldface for emphasis. Overall analyses are presented in italics, followed by moderator analyses (indented). *k* = number of independent samples contributing data; *N* = total sample size; $\bar{r}_{rr}$ = mean observed interrater reliability coefficient; SD_{obs} = observed standard deviation of interrater reliability coefficients; SD_{resid} = standard deviation of interrater reliability coefficients after accounting for variability due to sampling error and test–retest unreliability; ρ_{rr} = mean interrater reliability corrected for the test–retest unreliability of both raters; SD_{ρ} = standard deviation of corrected interrater reliability accounting for variability from sampling error and test–retest unreliability; $Conf_L$ and $Conf_U$ = lower and upper bounds of 95% confidence interval around ρ_{rr}; FS *k* = fail-safe *k*, the number of studies where *r* = .00 that must be located to make *r* = .05.

Summary. Study 1 focused specifically on the first of Funder and West's (1993) three accuracy criteria: interrater reliability or, more commonly, rater consensus. The values reported in Study 1 yield two general but clear messages. The magnitude of single-rater reliabilities (generally between uncorrected $\bar{r}_{rr}$ = .30 and $\bar{r}_{rr}$ = .45) suggests that there is clear correspondence across raters. At the same time, however, the magnitude of reliabilities clearly indicates the need for researchers to continue soliciting other-

ratings from multiple other-raters. Ratings from a single other-rater contain substantially idiosyncratic views of the target. These idiosyncratic perspectives of the target's personality traits will substantially attenuate the correlations between other-ratings and external variables (such as self-ratings or behavioral criteria). Use of multiple raters, however, offers a method of overcoming these rater idiosyncrasies. If the interrater reliability for a single rater is $\bar{r}_{rr} = .45$ (as was the case for the best information sources, family and friends, rating the best trait, Extraversion), five other-raters are needed to produce a composite with interrater reliability greater than $\bar{r}_{rr} = .80$. If the interrater reliability for a single rater is a more typical value of $\bar{r}_{rr} = .30$, 10 other-raters are needed to produce a composite with interrater reliability greater than $\bar{r}_{rr} = .80$. Thus, these results point to a clear need for researchers to collect other-ratings from multiple others to overcome the drastic idiosyncrasies of a single other-rater.

Study 2: Other-Rater Accuracy as Strong Self–Other Correlations

Study 1 focused on the overlap between two different other-raters in describing the personality of a particular target person. In contrast, Study 2 focused on the correspondence between an other's rating and the target's self-rating and meta-analytically examined how closely self- and other-perceptions are aligned. Such research is central to understanding the uniqueness and redundancy of self- and other-reports of personality. As correlations between self-ratings and other-ratings near zero, the constructs assessed by self- and other-rating measurement methods become increasingly independent. Because personality traits are conceptualized as being stable characteristics of individuals across situations, meager convergence across these methods would indicate that at least one (and maybe both) is deficient or contaminated as a measurement source. Moreover, examining trait and information moderators of self–other correlations provides insight about potential causes of this deficiency or contamination. On the other hand, as self–other correlations near unity, the information contained in each rating source becomes increasingly redundant, and neither rating source is likely to afford any unique perspective on the target's relative standing on the trait. Should self- and other-ratings of personality traits show especially strong redundancy, little incremental validity could be expected from combining self- and other-ratings to predict behavior. Indeed, the choice between measuring personality traits would simply be a matter of the relative convenience of administration. Thus, establishing where self- and other-ratings fall on this continuum between uniqueness and redundancy is a necessary first step for evaluating their deficiency or contamination and for forecasting the unique contribution of each rating type.

A considerable amount of research has studied this question of the overlap between self- and other-ratings of personality traits, and some of this research has been summarized in a previous meta-analysis by Connolly et al. (2007). Connolly et al. found generally strong correlations between self- and observer-ratings of traits. Close relatives' ratings generally had stronger correlations with self-ratings than did ratings from peers. The advantage for close relatives' ratings for predicting self-ratings was even more pronounced when compared to strangers' ratings. Strangers showed low correspondence with self-ratings for Emotional Stability ($\rho = .08$), Openness ($\rho = .22$), and Agreeableness ($\rho = -.01$) and only modest self–other correlations for Extraversion ($\rho = .39$) and Conscientiousness ($\rho = .34$). By adding a number of previously unavailable or omitted studies, Study 2 represents an increase from Connolly et al. in independent samples by a factor of two to four. This larger sampling of studies allowed for better approximation of true self–other correlations and a finer analysis of potential moderators.

Study 2: Method

Study 2 was a meta-analysis of correlations between self-ratings of a personality trait and other-ratings on the same trait. However, studies frequently based these self–other correlations on the relationship between a self-rating and a composite of multiple others-ratings. All else equal, studies using composites of more raters will produce stronger self–other correlations than studies using just one rater, due to higher interrater reliabilities of composites. Because different information sources tend to use different numbers of raters (e.g., only one rater for self–spouse correlations but frequently multiple raters for self–stranger correlations), it was desirable initially to estimate self–other correlations with the number of other-raters held constant. Thus, self–other correlations reported in studies were adjusted to estimate the correlation that would have been observed had a single other-rater been used. To estimate these self–other correlations for a single other, we first disattenuated these self–other correlations based on a multi-other composite for interrater unreliability for r raters. Next, these disattenuated self–other correlations were reattenuated for the interrater unreliability of 1 rater. For studies providing other-rater interrater reliability, we used the reliabilities reported in the study to estimate the correlation for a single–other pair. For studies that did not report other interrater reliabilities, we selected interrater reliability values from Study 1 that matched the study on the Big Five trait measured, the type of other-rater used, and the type of cuing stimuli.

In many cases, a single study would contribute multiple self–other correlations for the same analysis. To avoid violating assumptions about the independence of correlations from samples, we combined these correlations in one of two ways within samples before meta-analyzing them across samples. In some cases, several correlations were reported because studies used multiple measures assessing the same Big Five trait (e.g., a study reports self–other correlations for two Conscientiousness facet measures, achievement and dependability). In such cases, self–other correlations were composited (Nunnally, 1978) if intercorrelations among the scales could be estimated. Composite correlations represent the correlation that would have been observed had the scales been summed (e.g., summing achievement and dependability scores to better represent the full Conscientiousness domain) prior to calculating a correlation between self- and other-ratings. Compositing is not possible when scale intercorrelations are not reported; in such instances, self–other correlations were averaged within the study. In other cases, several correlations were reported within one sample because studies used multiple types of other-raters (e.g., self–sibling correlations and self–stranger correlations). Such correlations were averaged for the purposes of overall analyses but separated for moderator analyses.

In some cases, self-ratings and other-ratings of personality traits were not made on the same scale. The most common scenario in such cases was that others rated the target on a shorter form developed specifically to parallel the inventory completed by targets. We compared self–other correlations of samples as a potential moderator using the same measures for self- and other-raters to those using different measures. In nearly all analyses, studies using different inventories for self- and other-raters yielded comparable self–other correlations to those using the same inventory. Contrary to what might have been expected, the small differences in correlations were more likely to favor studies using different scales than those using identical scales. Finding such small differences in the opposite direction of what would be expected suggests that these differences are likely the result of second-order sampling error. Thus, we report the results combining same-measure and different-measure self–other correlations.

Statistical artifact distributions. To reduce bias due to statistical artifacts, we corrected mean and variance estimates of self–other correlations for sampling error and unreliability in both the self-rating and the other-rating. As in Study 1, no range restriction corrections were made, but we corrected for unreliability in self- and other-ratings. For self-ratings, we used artifact distributions of test–retest reliabilities from Viswesvaran and Ones (2000).

Two separate methods were used independently to correct the self–other correlations for measurement in the other-ratings. In the first method, the correlations between self- and other-ratings were corrected for test–retest unreliability. These test–retest reliability corrected values indicate the stable overlap between a self-rating and a single other-rating (i.e., with transient error removed from both self- and other-ratings). If the self-rater is indeed simply equivalent to another other-rater, these values should be directly comparable to the corrected meta-analytic single-other interrater reliabilities from Study 1.

The second reliability correction method for other-ratings was to use single-rater interrater reliabilities to correct for unreliability in other-ratings. Single-rater interrater unreliability distributions from Study 1 were matched to self–other correlation meta-analyses on traits measured, information source, and information type. In analyses spanning across relationship types (e.g., Emotional Stability self–other correlations for all types of others), some information source categories were represented in the Study 1 artifact distributions disproportionately to their representation in the self–other correlations. For example, many more interrater reliabilities were presented for strangers' ratings than for friends' ratings, but fewer self–other correlations were obtained for strangers than for friends. Such differences were important to account for, given the finding in Study 1 that information source moderates interrater reliability. Thus, interrater unreliability distributions for overall analyses combining across information source categories were synthetically created by weighting the frequencies of interrater reliabilities from each information source category by their proportional representation in the self–other correlations.

Each set of self–other correlations was meta-analyzed twice: In one set, other-ratings were corrected for test–retest unreliability; in the other set, other-ratings were corrected for interrater reliability for a single observer. Meta-analytic results representing these two analyses are described throughout using the notation $\rho_{o=1}$ for test–retest corrected self–other correlations and $\rho_{o=\infty}$ for interrater-reliability corrected self–other correlations.

These two methods for correcting self–other correlations have substantively different meanings. Recall from Figure 1 that the realistic accuracy model (Funder, 1995) posits that personality ratings can be accurate because of a four-component process: relevance, availability, detection, and utilization. Funder noted that accuracy correlations reflect the extent to which these four components are satisfied, and, correspondingly, $\rho_{o=1}$values are affected by the extent to which trait information is relevant, available, detected, and utilized accurately by an other-rater. In contrast, however, because $\rho_{o=\infty}$ values have been corrected for interrater unreliability in other-ratings, $\rho_{o=\infty}$ values estimate the overlap of targets' self-perceptions with all information available to others. In other words, $\rho_{o=\infty}$ values estimate accuracy if detection and utilization had occurred perfectly.

This difference in meanings associated with $\rho_{o=1}$ and $\rho_{o=\infty}$ values has implications for testing for good trait and good information moderators. Good trait and good information moderators of $\rho_{o=1}$ correlations may affect relevance, availability, detection, or utilization processes. In contrast, because $\rho_{o=\infty}$ values estimate accuracy with perfect detection and utilization, good trait and good information moderators of $\rho_{o=\infty}$ correlations indicate that the moderator can affect only relevance or availability. Much can be gained by sequentially comparing the differences across trait and information source moderators that are found for $\rho_{o=\infty}$ correlations versus those found for $\rho_{o=1}$ correlations. Finding that an information or trait moderator affects $\rho_{o=\infty}$ correlations (e.g., one information source yields higher $\rho_{o=\infty}$ correlations than another information source) indicates that the moderator operates through the RA component of accuracy. This is apparent with $\rho_{o=\infty}$ correlations because (a) $\rho_{o=\infty}$ correlations correct for errors in DU processes and (b) the RADU of self-perceptions is held constant across self–other correlations. Even if $\rho_{o=\infty}$ correlations show no differences indicative of a moderator effect, that moderator may affect $\rho_{o=1}$ correlations. Finding differences in $\rho_{o=1}$ correlations without differences in $\rho_{o=\infty}$ correlations would indicate that the moderator affects other-rating accuracy entirely through DU processes. Thus, examining both sets of correlations makes it possible to determine the mechanisms through which moderators affect accuracy.

Study 2: Results and Discussion

Good trait moderators. First, self–other correlations were meta-analyzed across information source types for each personality trait (see Table 5). Examining the $\rho_{o=\infty}$ correlations suggests that there is considerable overlap between self-perceptions and the information available to others for all of the Big Five traits (i.e., cues for judging each of the Big Five appear to generally be relevant and available to observers). These corrected correlations ranged from $\rho_{o=\infty} = .71$ for Agreeableness to $\rho_{o=\infty} = .82$ for Conscientiousness, suggesting that there is strong but not complete overlap between self-perceptions and others' perceptions when analyzed across all observer types. In addition, these $\rho_{o=\infty}$ correlations cluster tightly, suggesting that good trait moderators do not strongly affect other-rating accuracy through RA processes.

Table 5
Self–Other Consensus Correlations: Information Source Moderators

Trait and information source	*k*	*N*	$\bar{r}$	SD_{obs}	$\rho_{o=1}$	$SD_{\rho o=1}$	$Conf_L$	$Conf_U$	$\rho_{o=\infty}$	$SD_{\rho o=\infty}$	FS *k*
Emotional Stability	*148*	*27,341*	*.34*	*.15*	***.43***	*.16*	*.39*	*.41*	***.72***	*.19*	*858*
Family	37	6,501	.43	.11	**.54**	.10	.48	.53	**.80**	.00	281
Friend	54	7,358	.33	.12	**.42**	.11	.36	.41	**.63**	.13	302
Cohabitator	16	2,777	.32	.08	**.40**	.05	.33	.41	**.70**	.00	86
Work colleague	8	981	.14	.10	**.18**	.05	.09	.23	**.35**	.06	14
Incidental acquaintance	8	1,054	.17	.17	**.21**	.18	.13	.27	**.45**	.37	19
Stranger	33	3,835	.08	.09	**.10**	.00	.06	.13	**.22**	.00	20
Extraversion	*186*	*28,957*	*.41*	*.15*	***.51***	*.16*	*.47*	*.49*	***.77***	*.17*	*1,339*
Family	38	6,834	.48	.10	**.61**	.10	.54	.58	**.83**	.10	327
Friend	69	9,091	.40	.12	**.51**	.11	.44	.49	**.70**	.13	483
Cohabitator	27	3,144	.38	.13	**.48**	.12	.41	.48	**.81**	.07	178
Work colleague	11	1,647	.24	.17	**.30**	.18	.23	.33	**.49**	.27	42
Incidental acquaintance	11	1,270	.34	.15	**.43**	.15	.34	.45	**.63**	.20	64
Stranger	40	4,328	.22	.10	**.27**	.04	.22	.29	**.43**	.00	136
Openness	*105*	*20,036*	*.34*	*.17*	***.45***	*.21*	*.40*	*.43*	***.79***	*.20*	*609*
Family	25	3,924	.43	.12	**.57**	.12	.49	.56	**.84**	.15	190
Friend	35	5,542	.37	.14	**.50**	.16	.42	.48	**.70**	.21	224
Cohabitator	13	2,144	.35	.19	**.47**	.23	.38	.47	**.99**	.45	78
Work colleague	6	1,396	.20	.17	**.27**	.21	.18	.30	**.48**	.35	18
Incidental acquaintance	8	799	.11	.10	**.15**	.04	.05	.22	**.30**	.02	10
Stranger	23	3,266	.12	.10	**.16**	.07	.10	.19	**.31**	.00	32
Agreeableness	*151*	*22,389*	*.29*	*.14*	***.39***	*.16*	*.34*	*.37*	***.71***	*.12*	*725*
Family	32	5,113	.37	.13	**.50**	.14	.42	.48	**.91**	.00	205
Friend	63	8,224	.29	.11	**.39**	.09	.33	.38	**.60**	.10	302
Cohabitator	18	2,634	.26	.10	**.34**	.08	.27	.36	**.65**	.10	76
Work colleague	11	1,647	.23	.11	**.31**	.11	.22	.34	**.53**	.17	40
Incidental acquaintance	10	1,080	.17	.13	**.23**	.12	.14	.28	**.41**	.21	24
Stranger	32	3,852	.09	.09	**.12**	.03	.07	.15	**.23**	.00	26
Conscientiousness	*145*	*23,907*	*.37*	*.15*	***.50***	*.17*	*.44*	*.46*	***.82***	*.19*	*928*
Family	33	5,154	.42	.10	**.57**	.09	.48	.54	**.85**	.00	244
Friend	56	7,102	.38	.12	**.51**	.10	.44	.48	**.76**	.13	370
Cohabitator	25	3,333	.38	.15	**.51**	.16	.43	.50	**.84**	.22	165
Work colleague	11	1,647	.18	.10	**.24**	.08	.16	.27	**.42**	.12	29
Incidental acquaintance	9	1,054	.24	.17	**.32**	.20	.22	.36	**.57**	.30	34
Stranger	25	3,264	.13	.07	**.18**	.00	.12	.20	**.34**	.00	40

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Corrected correlations are presented in boldface for emphasis. Overall analyses are presented in italics, followed by moderator analyses (indented). k = number of independent samples contributing data; N = total sample size; $\bar{r}$ = mean observed self–other correlation; SD_{obs} = observed standard deviation of self–other correlations; $\rho_{o=1}$ = self–other correlation corrected for test–retest reliability in self and other personality rating (i.e., corrected self–single other correlation); $SD_{\rho o=1}$ = standard deviation of $\rho_{o=1}$, correcting for variance due to sampling error and test–retest unreliability; $Conf_L$ and $Conf_U$ = lower and upper bounds of 95% confidence interval around $\rho_{o=1}$; $\rho_{o=\infty}$ = self–other correlation corrected for test–retest reliability in self and interrater reliability in other personality rating (i.e., corrected self–all other correlation); $SD_{\rho o=\infty}$ = standard deviation of $\rho_{o=\infty}$, correcting for variance due to sampling error, self test–retest unreliability, and other interrater reliability; FS k = fail-safe k, the number of studies where $r = .00$ that must be located to make $r = .05$.

In examining these overall analyses for $\rho_{o=1}$ correlations, we found relatively little overlap in confidence intervals across the traits. Results generally paralleled those for interrater reliability coefficients for Study 1 (see Table 3). That is, self–other correlations were highest for the most visible trait of Extraversion ($\rho_{o=1} = .51$), followed by Conscientiousness ($\rho_{o=1} = .50$), the low-visibility traits of Openness and Emotional Stability ($\rho_{o=1} = .43$ and $\rho_{o=1} = .45$; overlapping confidence intervals), and finally the highly evaluative trait of Agreeableness ($\rho_{o=1} = .39$). Note that the rank ordering of the $\rho_{o=1}$ correlations did not match the rank ordering of $\rho_{o=\infty}$ correlations, which were quite close in magnitude across the five factors. This pattern of findings suggests that general differences in accuracy for particular traits are likely more due to differences in how easy it is to detect and utilize trait-related cues rather than differences across traits in the extent to which they are relevant and available in affecting trait expression.

Good information moderators. Next, self–other correlations for each trait were analyzed separately for other-raters who were family members, friends, cohabitators, work colleagues, incidental acquaintances, and strangers. When we examined $\rho_{o=\infty}$ correlations, a pattern similar to that observed in Study 1 emerged. Family members, who had the greatest intimacy, were consistently the other-raters with the strongest $\rho_{o=\infty}$ self–other correlations. Friends' and cohabitators' ratings had the next strongest $\rho_{o=\infty}$ correlations for most traits. Work colleagues and incidental acquaintances (high frequency but low ratings on intimacy dimensions) showed only small advantages in $\rho_{o=\infty}$ over strangers, who were generally the least accurate information sources. With the exception that the lowest self–other correlations were yielded by

strangers, this pattern is identical to that for interrater reliabilities. Thus, these results indicate that information source indeed affects accuracy by affecting relevance and availability.

When we examined $\rho_{o=1}$ correlations, a similar pattern emerged: Family members were highest, followed by friends and cohabitators, followed by work colleagues and incidental acquaintances, closely followed by strangers. However, $\rho_{o=1}$ correlations are even more (proportionally) distinct across information sources, and the confidence intervals generally overlap very little across rater types (overlap between friends and cohabitators being the exception). This finding suggests that differences in accuracy across information source moderators are due to differences in both RA and DU. Although others' ratings can converge with self-ratings for some traits after only short exposure to targets, considerable increments in self–other correlations occur only with the increased interpersonal intimacy that comes with friendship and being part of the family. With this increased intimacy, accuracy of other-raters improves mostly because (a) observers have more opportunities to observe trait-relevant cues (RA) but also because (b) other-raters draw trait inferences from these cues in a way that is more consistent with the way targets form trait inferences about themselves (DU). Thus, other-rater's information source appears to be an important moderator of self–other correlations through effects on both RA and DU.

Good trait × Information moderators. The advantages for more intimately acquainted other-raters were more pronounced for some traits than for others. The different information source categories showed the greatest discrepancies in $\rho_{o=\infty}$ for Emotional Stability and Openness, two traits especially low in visibility. In contrast, Extraversion (a trait high in visibility) showed high $\rho_{o=\infty}$, with the smallest differences across the information source moderators. This pattern for Extraversion held when examining $\rho_{o=1}$ correlations: Differences across information sources were still least pronounced for Extraversion. These results suggest that, compared to those for other traits, information source differences in rating Extraversion are due less to differences in relevance and availability and more to differences in detection and utilization. High-visibility traits such as Extraversion are already high in relevance and availability, so increased intimacy has less of an effect on accuracy. For low-visibility traits, this increased intimacy of information source is necessary for improving trait relevance and accuracy by gaining access to the individual's internal thoughts and feelings. As a result, low trait visibility appears to be less of a barrier for well-acquainted others in creating accurate ratings, because trait visibility functions primarily through its effects on relevance and availability.

For Agreeableness (the trait highest in evaluativeness), the magnitude and dispersion across information sources for $\rho_{o=\infty}$ correlations were comparable to those for Emotional Stability, Openness, and Conscientiousness. However, $\rho_{o=1}$ correlations were generally lower than other traits and showed limited advantages for more closely acquainted information sources. Finding this inconsistency suggests that smaller differences for Agreeableness across information sources' $\rho_{o=1}$ correlations are due to differences across information sources in detection and utilization. When detection and utilization occur perfectly (as represented in $\rho_{o=\infty}$ correlations), differences across information sources are as pronounced for Agreeableness as for other traits. This pattern of findings suggests that well-acquainted other-raters are more idiosyncratic (weaker DU) in rating Agreeableness than they are for other traits. These idiosyncrasies specific to Agreeableness affect accuracy because of differences in the way individuals detect and utilize Agreeableness-related information. Thus, $\rho_{o=1}$ correlations may be lower for well-acquainted others for Agreeableness, because other-raters' judgments of the target's likability is more a matter of personal preference (DU) than a result of the target's attempt to disguise unlikable behaviors (RA). Thus, these results do not suggest that a highly evaluative trait such as Agreeableness is weaker in accuracy because of self-presentation effects but rather because ratings of evaluative traits may be more prone to idiosyncrasies in rater perceptions, even when the other-rater is intimately acquainted with the target.

Our examination of information source moderators in Table 5 generally showed reductions in variability around meta-analytic estimates when separating by information source, consistent with information source being an important accuracy moderator. However, the SD_ρ values in Table 5 for family, friends, and cohabitators showed remaining variability in self–other correlations that potentially signal the presence of additional moderators within information sources. This was not the case for strangers' ratings, however: SD_ρ values for self–stranger correlations were consistently near zero. Thus, regardless of the information type presented to strangers, self–other correlations were generally of the same magnitude. Finding a general lack of variability in self–stranger correlations is perhaps quite surprising. Despite the wide variety of study designs to expose strangers to information about targets, stranger-ratings converged with self-ratings quite consistently across studies, regardless of the corrections applied. Thus, Study 1's finding that information type was an important moderator of strangers' interrater reliabilities was not paralleled in Study 2's self–other correlations.

Information source moderators of self–other correlations: Hierarchical moderator analyses. Further moderator analyses were conducted within information source types for family, friends, and cohabitators. For example, self–family member correlations were separated into self–parent, self–spouse, and self–sibling correlations. Given the small number of interrater reliabilities for nonstranger information sources in Study 1, separate interrater reliability corrections were not available for different information source Level 2 moderators. Because we did not use different interrater reliability artifact distributions within a particular information source, we do not make attributions to relevance, availability, detection, or utilization here and focus on differences across Level 2 information source $\rho_{o=1}$ correlations. These moderator analyses are described in turn within each information source.

Self–family information sources. Table 6 presents self–other correlations separated by type of family member. Across all traits for family members, $\rho_{o=1}$ correlations were higher when the other-rater was a spouse than a parent. The advantage for self–spouse correlations was even more pronounced for some traits. Extraversion (the trait typically most easily observed accurately) showed the greatest differences between self–parent and self–spouse/sibling correlations. These findings are perhaps to be expected. Parents typically have little opportunity to observe children in social circles with their peers and have a built-in dominance structure in their relationship with the target. Relatively few self–sibling correlations were available and only for Emotional Stabil-

Table 6
Moderators of Self–Other Consensus Correlations: Types of Family Members

Trait and family member	*k*	*N*	$\bar{r}$	SD_{obs}	$\rho_{o=1}$	$SD_{\rho o=1}$	$Conf_L$	$Conf_U$	$\rho_{o=\infty}$	$SD_{\rho o=\infty}$	FS *k*
					Emotional Stability						
All family	*37*	*6,501*	*.43*	*.11*	***.54***	*.10*	.51	.56	***.80***	*.00*	281
Spouse	22	2,970	.43	.10	**.54**	.08	.50	.58	**.80**	.00	167
Parent	11	1,458	.34	.13	**.43**	.13	.37	.49	**.62**	.12	64
Sibling	3	1,607	.45	.05	**.57**	.03	.52	.62	**.83**	.00	24
					Extraversion						
All family	*38*	*6,834*	*.48*	*.10*	***.61***	*.10*	.59	.63	***.83***	*.10*	327
Spouse	22	2,901	.50	.07	**.63**	.00	.59	.66	**.86**	.00	198
Parent	12	1,730	.36	.09	**.45**	.07	.40	.50	**.61**	.07	74
Sibling	2	1,372	.56	.01	**.70**	.00	.65	.74	**.95**	.00	20
					Openness						
All family	*25*	*3,924*	*.43*	*.12*	***.57***	*.12*	.54	.60	***.84***	*.15*	190
Spouse	16	1,999	.44	.10	**.58**	.08	.53	.63	**.85**	.09	125
Parent	8	1,186	.36	.15	**.48**	.17	.41	.54	**.70**	.24	50
Sibling											
					Agreeableness						
All family	*32*	*5,113*	*.37*	*.13*	***.50***	*.14*	.47	.53	***.91***	*.00*	205
Spouse	19	2,527	.40	.11	**.53**	.09	.49	.57	**.98**	.00	133
Parent	10	1,515	.28	.14	**.37**	.15	.31	.43	**.69**	.00	46
Sibling											
					Conscientiousness						
All family	*33*	*5,154*	*.42*	*.10*	***.57***	*.09*	.54	.60	***.85***	*.00*	244
Spouse	19	2,527	.46	.10	**.62**	.08	.58	.66	**.93**	.00	156
Parent	11	1,556	.36	.11	**.49**	.10	.43	.55	**.73**	.00	68
Sibling	1	240	.42		**.56**		.41	.69	**.84**		7

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Corrected correlations are presented in boldface for emphasis. Overall analyses are presented in italics, followed by moderator analyses (indented). k = number of independent samples contributing data; N = total sample size; $\bar{r}$ = mean observed self–other correlation; SD_{obs} = observed standard deviation of self–other correlations; $\rho_{o=1}$ = self–other correlation corrected for test–retest reliability in self and other personality rating (i.e., corrected self–single other correlation); $SD_{\rho o=1}$ = standard deviation of $\rho_{o=1}$, correcting for variance due to sampling error and test–retest unreliability; $Conf_L$ and $Conf_U$ = lower and upper bounds of 95% confidence interval around $\rho_{o=1}$; $\rho_{o=\infty}$ = self–other correlation corrected for test–retest reliability in self and interrater reliability in other personality rating (i.e., corrected self–all other correlation); $SD_{\rho o=\infty}$ = standard deviation of $\rho_{o=\infty}$, correcting for variance due to sampling error, self test–retest unreliability, and other interrater reliability; FS k = fail-safe k, the number of studies where $r = .00$ that must be located to make $r = .05$.

ity, Extraversion, and Conscientiousness. However, these self–sibling correlations were generally of comparable magnitude to self–spouse correlations.

Finding relatively lower self–parent correlations is quite intriguing. A person's parents are almost always the individuals who have known the target the longest. However, self–parent correlations were of approximately comparable magnitude to self–friend correlations. The additional, small-*k* findings showing self–sibling correlations to be of comparable magnitude to self–spouse correlations raises questions about why self–parent correlations are not higher. There are several possible explanations for these findings. First, parents may wear rose-colored glasses when describing their children, such that parents are less able to acknowledge where their children's personality traits may be somewhat undesirable. Second, spouses and siblings may simply be more interpersonally intimate with targets than are parents. Given that parent-raters were most commonly used family-member other-raters when targets were undergraduates, the personality of many targets may be at a critical developmental stage as they explore (and exploit) newfound identity and freedoms of living independently from their parents. A final component of this decreased intimacy may be that parents are prone to recalling impressions of targets formed during targets' childhood, and these impressions may be insensitive to true personality changes during the course of adult development.

Self–friend information sources. Moderator analyses comparing correlations between self-ratings and ratings from different types of friends are presented in Table 7. These analyses show several trends. First, dating partners had the highest self–other $\rho_{o=1}$ correlations across all sets of friends. These correlations were generally comparable to self–spouse correlations. The small number of studies specifically focusing on dating partners (and the large standard deviations around the meta-analytic means), however, makes these analyses potentially subject to second-order sampling error. That is, due to the small k for dating partners, the studies included in the present analyses may have by chance been studies producing especially large self–other correlations. Thus, there is a need for additional research on self–other correlations among dating couples. Studies specifying that friends be a best/

Table 7
Moderators of Self–Other Consensus Correlations: Types of Friends

Trait and friend type	*k*	*N*	$\bar{r}$	SD_{obs}	$\rho_{o=1}$	$SD_{\rho o=1}$	$Conf_L$	$Conf_U$	$\rho_{o=\infty}$	$SD_{\rho o=\infty}$	FS *k*
Emotional Stability											
All friends	*54*	*7,358*	*.33*	*.12*	***.42***	*.11*	*.39*	*.45*	***0.63***	*.13*	*302*
Dating partner	3	496	.38	.13	**.48**	.13	.38	.57	**0.74**	.16	20
Best/close friend	11	987	.29	.12	**.37**	.08	.30	.44	**0.57**	.07	53
Friend/close acquaintance	36	5,092	.32	.13	**.41**	.12	.38	.44	**0.62**	.16	194
Peer at school	5	803	.33	.07	**.41**	.00	.33	.48	**0.63**	.00	28
Extraversion											
All friends	*69*	*9,091*	*.40*	*.12*	***.51***	*.11*	*.49*	*.53*	***0.70***	*.13*	*483*
Dating partner	2	430	.42	.13	**.53**	.14	.43	.62	**0.73**	.17	15
Best/close friend	10	743	.37	.19	**.47**	.19	.39	.55	**0.65**	.25	64
Friend/close acquaintance	49	6,482	.40	.11	**.51**	.10	.48	.54	**0.70**	.12	343
Peer at school	6	1,049	.39	.06	**.49**	.00	.42	.55	**.68**	.00	41
Openness											
All friends	*35*	*5,542*	*.37*	*.14*	***.50***	*.16*	*.47*	*.53*	***0.70***	*.21*	*224*
Dating partner	2	430	.51	.05	**.69**	.00	.59	.78	**0.96**	.00	18
Best/close friend	3	327	.21	.14	**.28**	.14	.14	.42	**0.39**	.19	10
Friend/close acquaintance	27	4,269	.37	.14	**.49**	.16	.46	.52	**0.69**	.21	173
Peer at school	4	648	.31	.16	**.42**	.19	.32	.51	**0.59**	.26	21
Agreeableness											
All friends	*63*	*8,224*	*.29*	*.11*	***.39***	*.09*	*.36*	*.42*	***0.60***	*.10*	*302*
Dating partner	3	496	.42	.08	**.56**	.03	.46	.65	**0.86**	.00	22
Best/close friend	10	725	.25	.14	**.33**	.10	.24	.42	**0.51**	.13	40
Friend/close acquaintance	46	6,128	.28	.11	**.38**	.08	.35	.41	**0.58**	.09	212
Peer at school	3	738	.29	.09	**.39**	.07	.30	.48	**0.61**	.06	14
Conscientiousness											
All friends	*56*	*7,102*	*.38*	*.12*	***.51***	*.10*	*.48*	*.54*	***0.76***	*.13*	*370*
Dating partner	2	430	.50	.04	**.67**	.00	.57	.76	**1.00**	.00	18
Best/close friend	9	632	.38	.14	**.51**	.12	.42	.60	**0.76**	.16	59
Friend/close acquaintance	41	5,327	.36	.12	**.48**	.10	.45	.51	**0.72**	.13	254
Peer at school	4	643	.38	.10	**.51**	.07	.42	.60	**0.77**	.07	26

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Corrected correlations are presented in boldface for emphasis. Overall analyses are presented in italics, followed by moderator analyses (indented). *k* = number of independent samples contributing data; *N* = total sample size; $\bar{r}$ = mean observed self–other correlation; SD_{obs} = observed standard deviation of self–other correlations; $\rho_{o=1}$ = self–other correlation corrected for test–retest reliability in self and other personality rating (i.e., corrected self–single other correlation); $SD_{\rho o=1}$ = standard deviation of $\rho_{o=1}$, correcting for variance due to sampling error and test–retest unreliability; $Conf_L$ and $Conf_U$ = lower and upper bounds of 95% confidence interval around $\rho_{o=1}$; $\rho_{o=\infty}$ = self–other correlation corrected for test–retest reliability in self and interrater reliability in other personality rating (i.e., corrected self–all other correlation); $SD_{\rho o=\infty}$ = standard deviation of $\rho_{o=\infty}$, correcting for variance due to sampling error, self test–retest unreliability, and other interrater reliability; FS *k* = fail-safe *k*, the number of studies where *r* = .00 that must be located to make *r* = .05.

especially close friend or a peer at school generally produced self–other correlations similar to more generic instructions to nominate a friend or close acquaintance to provide ratings. These findings do not necessarily mean that increased intimacy among friends has no effect on self–other correlations. Rather, it is more likely that participants automatically look first to nominate their closest friends when studies ask them to nominate any friend or a peer at school.

Self–cohabitator information sources. Next, we compared self–other $\rho_{o=1}$ correlations when cohabitators were roommates versus when they were dorm or housemates. These data are presented in Table 8. The type of cohabitator made a substantial difference in self–other correlations. Across all of the five factors, roommates' ratings converged with self-ratings much more strongly than did dorm- or house-mates' ratings. This advantage for roommates' ratings was most pronounced for Openness traits ($\rho_{o=1}$ = .60 for roommates vs. $\rho_{o=1}$ = .15 for dorm/housemates) and Conscientiousness traits ($\rho_{o=1}$ = .59 for roommates vs. $\rho_{o=1}$ = .32 for dorm/housemates). Generally, self–roommate correlations were similar to self–friend correlations, whereas self–dorm/housemates correlations were closer to self–colleague correlations.

Comparing self–other accuracy among friends versus cohabitators gives additional insight into the role of setting for observing a target. Roommates differed from friends in that they have access to observing targets within the "private sphere" of targets' living spaces, whereas friends' private sphere observation is irregular and less frequent. Roommate self–other accuracy correlations were generally quite comparable to those for friends. Two explanations may account for these findings. One possibility is that roommates

Table 8
Moderators of Self–Other Consensus Correlations: Types of Cohabitators

Trait and cohabitator type	k	N	$\bar{r}$	SD_{obs}	$\rho_{o=1}$	$SD_{\rho o=1}$	$Conf_L$	$Conf_U$	$\rho_{o=\infty}$	$SD_{\rho o=\infty}$	FS k
Emotional Stability											
All cohabitators	*16*	*2,777*	*.32*	*.08*	***.40***	*.05*	.36	.44	***0.70***	*.00*	35
Roommate	12	2,288	.32	.09	**.41**	.07	.36	.46	**0.72**	.05	26
Dorm/housemate	3	407	.29	.04	**.37**	.00	.25	.48	**0.65**	.00	6
Extraversion											
All cohabitators	*27*	*3,144*	*.38*	*.13*	***.48***	*.12*	.44	.52	***0.81***	*.07*	76
Roommate	18	1,984	.43	.12	**.55**	.11	.50	.60	**0.93**	.00	59
Dorm/housemate	7	877	.27	.07	**.34**	.00	.26	.42	**0.58**	.00	12
Openness											
All cohabitators	*13*	*2,144*	*.35*	*.19*	***.47***	*.23*	.42	.52	***0.99***	*.45*	33
Roommate	9	1,559	.45	.11	**.60**	.12	.55	.65	**1.00**	.07	32
Dorm/housemate	3	384	.12	.05	**.15**	.00	.03	.27	**0.32**	.00	1
Agreeableness											
All cohabitators	*18*	*2,634*	*.26*	*.10*	***.34***	*.08*	.29	.39	***0.65***	*.10*	29
Roommate	10	1,574	.33	.06	**.43**	.00	.37	.49	**0.82**	.00	23
Dorm/housemate	7	978	.16	.06	**.21**	.00	.13	.29	**0.39**	.00	4
Conscientiousness											
All cohabitators	*25*	*3,333*	*.38*	*.15*	***.51***	*.16*	.47	.55	***0.84***	*.22*	70
Roommate	18	2,412	.44	.13	**.59**	.13	.55	.63	**0.98**	.12	61
Dorm/housemate	5	638	.24	.06	**.32**	.00	.22	.42	**0.53**	.00	7

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Corrected correlations are presented in boldface for emphasis. Overall analyses are presented in italics, followed by moderator analyses (indented). k = number of independent samples contributing data; N = total sample size; $\bar{r}$ = mean observed self–other correlation; SD_{obs} = observed standard deviation of self–other correlations; $\rho_{o=1}$ = self–other correlation corrected for test–retest reliability in self and other personality rating (i.e., corrected self–single other correlation); $SD_{\rho o=1}$ = standard deviation of $\rho_{o=1}$, correcting for variance due to sampling error and test–retest unreliability; $Conf_L$ and $Conf_U$ = lower and upper bounds of 95% confidence interval around $\rho_{o=1}$; $\rho_{o=\infty}$ = self–other correlation corrected for test–retest reliability in self and interrater reliability in other personality rating (i.e., corrected self–all other correlation); $SD_{\rho o=\infty}$ = standard deviation of $\rho_{o=\infty}$, correcting for variance due to sampling error, self test–retest unreliability, and other interrater reliability; FS k = fail-safe k, the number of studies where $r = .00$ that must be located to make $r = .05$.

have less intimacy with targets or less motivation to attend to targets' trait-related cues, but their observation of targets in private settings compensates for this lack of intimacy and motivation. Alternatively, observation of private spheres may yield no increase in accuracy, and the similarity between self–cohabitator and self–friend correlations is simply explained as being because targets choose to live with friends. Kurtz and Sherker (2003) specifically studied previously unacquainted roommates at two time points (2 and 15 weeks after moving in together). Self–roommate correlations were initially modest (near that of work colleagues and incidental acquaintances) for all traits but Conscientiousness, but the self–other correlations increased by Week 15 (of comparable magnitude to our estimates for self–friend or self–family member correlations). No moderating effect was found for relationship quality on self–other correlations, suggesting private sphere observation may indeed compensate in roommates' self–other correlations. Still, further research separating friend and nonfriend cohabitators is needed to inform this question.

Summary. These results showed self-ratings and other-ratings to generally overlap substantially, especially when corrected for other-rater interrater unreliability ($\rho_{o=\infty}$). Such a finding offers strong support that personality traits are relevant and available. That is, personality traits strongly affect behavior in ways that can be apparent to observers. Still, this overlap is not complete, and observers are likely to still see targets in a slightly different way than targets see themselves.

Good traits similar to those in Study 1 emerged in Study 2. In particular, self–other accuracy was generally highest for Extraversion and lowest for Emotional Stability and Openness to Experience. This pattern of findings was most pronounced for information sources not intimately acquainted with the target. It is likely that in interpersonally intimate relationships, there are more circumstances in which expressing traits typically low in visibility is relevant and available. Put another way, a person's worries and self-doubt (low Emotional Stability) or musings about art and ideas (Openness) are not likely to be poured out to a colleague or a new acquaintance. Indeed, religious and political attitudes are stigmatized topics to bring up in conversation with unfamiliar peers, and both are strongly related to Openness. Thus, traits' low visibility may in part reflect conforming to social norms to not express these traits. In cultures where social norms differ, however, these traits may be more frequently expressed and thereby more visible. Research comparing self–other correlations across cultures varying on such social norms may be especially fertile ground for studying the effects of trait visibility.

Good information moderators of accuracy are also apparent throughout the results of Study 2. First, these findings point to the importance of interpersonal intimacy with the target for self–other accuracy. Spouses and dating partners—whose intimacy with the target is greatest—showed the highest self–other correlations for all traits. Friends, roommates, and parents formed the next tier of self–other correlations, followed by dorm/housemates, work colleagues, and incidental acquaintances. Indeed, self–other correlations were only slightly greater for work colleagues and incidental acquaintances than they were for strangers. This is particularly surprising: The perceptions of a colleague who works beside the target day in and day out may align with target self-perceptions only slightly more than those of someone who has just met the target. Thus, these findings suggest that the quality of interactions with a target plays a much stronger role than pure quantity of interactions in self–other accuracy.

In addition, the magnitude of self–stranger correlations in Study 2 was substantially weaker than the stranger interrater reliabilities in Study 1. Such limited exposure as afforded to stranger ratings shows clear deficits when the accuracy criterion is self–other correlations, suggesting that stranger ratings are considerably less valid than ratings from other information source categories. The only consistent exception to this came when strangers rated targets' Extraversion. Although it is fascinating that strangers show any evidence for validity, the modest self–stranger correlations suggest that researchers should be wary of interpreting the present enthusiasm for zero-acquaintance and thin slice studies as evidence that the depth of personality traits are readily apparent to even a casual observer.

Study 3: Validity of Other-Ratings for Predictions of Behavior

In Study 2, we examined the relationship between self-ratings and other-ratings of traits. The results from Study 2 suggest that there is clear overlap in self- and other-perceptions of personality traits among those who are close with the target, a finding that is encouraging for studying the validity of other-ratings for predicting behaviors. This overlap implies that the validity of self-ratings of personality traits for predicting behavior is likely to generalize to traits measured by other-ratings.

Despite the overlap found in Study 2, self- and other-ratings are not completely redundant, even after correcting for interrater reliability in the other-rater ($\rho_{o=\infty}$). Such findings indicate that some distinctiveness among self- and other-perceptions of targets' personality traits likely remains, raising the question "Which perspective is more accurate?" Evaluating the relative accuracy of self- and other-ratings of personality traits necessitates a comparison of how each predicts theoretically related external criteria. Only a handful of studies have examined the validity of other-ratings for predicting such external criteria. Study 3 meta-analyzed these studies across three criterion domains: first impressions of traits conveyed to strangers, academic achievement, and job performance.

There is a long-standing tradition of concerns about self-presentation effects in self-report measures that would argue that other-ratings predict target behavior better than do self-reports. Self-presentation effects include impression management and self-deception (Paulhus, 1984; Paulhus & Reid, 1991; Paulhus & Trapnell, 2008). Although traditional scales intended to assess self-presentation have a history of disappointing results in detecting distortion (Ellingson, Sackett, & Hough, 1999; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; McCrae & Costa, 1983; Ones, Viswesvaran, & Reiss, 1996), concerns that individuals may intentionally or unintentionally misrepresent themselves on self-report personality measures are prominent. Such self-report biases would introduce trait-irrelevant variance into measures and would diminish self-reports' predictive validity. Other-reports, presumably free of such self-presentation effects, could potentially assess traits more directly and may thus produce higher validities for other-ratings in predicting behavior.

In contrast, it is entirely possible that other-ratings might have lower predictive validities than self-report measures. Simply put, no other-raters have the same degree of opportunity as does the self to observe a target's behavior, nor the direct access to a target's thoughts, feelings, and values (i.e., RA is strongest for self-ratings). This may be the case particularly when other-raters are not close acquaintances of the target. Furthermore, there may be response distortion effects on other-ratings. Through friendship biases, some others may be unwilling or unable to disclose negative information about the target. Thus, distortion effects similar to those thought to affect self-reports could affect other-ratings. In addition, judgment biases (e.g., overlooking the press of social roles and situations on the behavior of others) might potentially weaken others' judgments of personality (Jones, 1979; Nisbett & Ross, 1980). To the extent that differences in the opportunity to observe or differences in rating biases create differences between self- and other-ratings' accuracy, self-ratings may have an advantage in predicting behavior.

Finally, an alternate perspective is to view the self as simply another other-rater. In a particularly interesting study, McCrae, Stone, Fagan, and Costa (1998) administered self-report and spouse-report measures to a group of couples. In follow-up interviews about items with disagreements between self- and spouse-reports, the most common reasons were interpreting items differently or considering different specific instances. These reasons are consistent with traditional conceptualizations of measurement error and would be reduced by administering multiple items. Disagreements from considering the target in different contexts, roles, or time periods were considerably less frequent, and intentional self-reported faking, perceived contrast, and assumed similarity were almost never listed as reasons for disagreements. If this is truly and broadly the case, self- and other-ratings should have comparable validities.

Thus, in addition to establishing the accuracy of other-ratings for predicting behavior, Study 3 explicitly compared the validities of other-ratings to self-ratings. The available studies contributing relevant data of other-ratings predicting behaviors are a small pool of studies in only a few domains: trait first impressions (how an other-rating of a particular trait corresponds to ratings of the same trait made by strangers), academic achievement (typically indicated by grades), and job performance (individuals' contribution to organizational effectiveness). Therefore, Study 3 serves as a summary of available research predicting behavioral outcomes and a progress report intended to direct future research.

Study 3: Method

Studies included in the Study 3 meta-analysis presented a correlation between other-ratings of a personality trait and a measure of a stranger's first impression of the target's trait, academic achievement, or job performance. These behavioral measures had to be made by an independent rating source. For example, some of the correlations reported by Mount, Barrick, and Strauss (1994) were between supervisor ratings of personality traits and the same supervisor's ratings of job performance. Because of common method bias, including such correlations would upwardly bias validities for other-ratings. Thus, we excluded such nonindependent correlations.

Meta-analytic procedures similar to those described in Studies 1 and 2 were followed. As before, samples varied in the number of other-raters used to measure personality. To facilitate comparisons with self-ratings, we adjusted all correlations between multi-other composite ratings of traits and criteria to the level of a single other-rater using procedures identical to those in Study 2. That is, validities were individually disattenuated for interrater unreliability of *r* raters and then reattenuated for interrater unreliability of a single rater. Similarly, when samples contributed several correlations to a single analysis as a result of using multiple measures of the predictor or criterion, these correlations were composited where possible and otherwise averaged.

Artifact distributions. Across all criteria, we corrected predictor unreliability in other-ratings using interrater-reliability coefficients from Study 1 most closely matching the samples contributing predictive validities. For predicting trait first impressions, friends' interrater reliability distributions were used; for predicting academic achievement, overall (across rater categories) interrater reliability distributions were used; for predicting job performance, work colleagues' interrater reliability distributions were used. Comparing these interrater reliability corrected validities for other-raters to true score validities for self-raters (typically corrected using internal consistency reliability) roughly compares the validity of all possible trait information gathered from others to all possible trait information gathered from self. In Figure 1, these validities are labeled as $\rho_{xy_{self}}$ and $\rho_{xy_{other}}$. In addition, operational validities are presented that are corrected for other artifacts but not for unreliability in other- or self-ratings. These operational validities can be used to compare the validity of a single other-rating with that of a self-rating ($\rho_{ov_{self}}$ and $\rho_{ov_{other}}$ in Figure 1).

We made corrections for criterion unreliability and range restriction to match meta-analyses of self-ratings as closely as possible. For predicting trait first impressions, criterion unreliability was corrected using interrater unreliability distributions for strangers from Study 1, and no corrections were made for range restriction. For predicting academic achievement, criteria were corrected for unreliability using the internal consistency reliability distribution for grades described in Kuncel, Hezlett, and Ones (2001), and no corrections were made for range restriction in the predictor or the criterion. For predicting job performance, interrater unreliability distributions from Viswesvaran et al. (1996) were used to create an unreliability artifact distribution. The values presented in Viswesvaran et al. were estimated as the interrater reliability for a single job performance rater. Thus, the distribution from Viswesvaran et al. was used to synthetically create a criterion reliability artifact distribution matching the number of job performance raters in the study. Although meta-analyses typically correct correlations between personality self-ratings and job performance for modest range restriction, the few sources providing validities for other-ratings did not provide adequate information for estimating the extent of range restriction. Thus, no range restriction corrections were made for other-ratings' predictive validities, and, as a result, comparing other-report validities to self-report validities for predicting job performance may somewhat favor self-report values.

Study 3: Results and Discussion

Throughout these results, we compare the validities for other-ratings and for self-ratings. For predicting academic achievement, meta-analytic self-report validities were drawn from Hough (1992) and from Poropat (2009); for predicting job performance, meta-analytic self-report validities were drawn from Barrick et al. (2001).[4] For predicting trait first impressions, the validities for other-ratings can be compared to the self–stranger correlations reported in Study 2. In addition, we estimated self–stranger correlations from the same studies contributing other–stranger data in Study 3. We report these values in Table 9 to add closer comparison for a set of targets and study methods. Unfortunately, self-report data were not available for many studies contributing other-rater predictions of academic achievement or job performance. Thus, comparisons are made to meta-analytic values only.

Predicting trait first impressions.

Zero-order correlations. First, we examined correlations for self-ratings and other-ratings with the criterion of first impressions made on strangers (see Table 9). Self-report comparison values come from the same studies providing other-rating validities. Note first that the observed correlations between self-ratings and other-ratings increase considerably when these correlations are corrected for interrater unreliability in the strangers' first impression ratings (ρ_{ov}). This is because the interrater reliability values in Study 1 for strangers that were used in correcting for criterion unreliability were typically modest. However, self-ratings and other-ratings provided comparable and strong operational validity for predicting trait first impressions, with self-ratings and other-ratings yielding point estimates with consistently overlapping confidence intervals. Openness showed the strongest operational validities (ρ_{ov}) for both self- and other-ratings in predicting trait first impressions. The advantages for Openness held even when we examined true score validities (ρ), in which differences across traits in interrater reliabilities are controlled. Operational validities and true score validities were somewhat lower for Agreeableness, Emotional Stability, and Conscientiousness for self- and for other-raters. On the whole, though, these findings show quite comparable self- and other-rater validities when the criterion is first impressions of traits made on strangers.

Self and one other combined. The incremental validity from combining a self-rating and an other-rating to predict trait first impressions was estimated. Note that, across all traits, combining a self-rating and one other-rating produces increments in validity ($R_{ov} - \rho_{ov}$) beyond self- or other-ratings alone. Thus, when self-report and other-report ratings are combined to predict trait

[4] Self-report validities from Poropat (2009) are based on a sample-size weighted average of correlations from secondary and tertiary schools.

Table 9
Meta-Analysis of Other-Ratings and Self-Ratings Validities for Predicting First Impressions Made on Strangers

	Zero-order meta-analytic results										Combined: Self + 1 other	
Trait and rating type	k	N	$\bar{r}$	SD_{obs}	SD_{resid}	ρ_{ov}	$SD_{\rho_{ov}}$	$CI_{\rho_{ov}}$	ρ_{xy}	SD_{ρ}	R_{ov}	β
Emotional Stability											**.26**	
Other-ratings	7	1,013	.18	.07	.00	**.24**	.00	[.16, .32]	**.41**	.00		.20
Self-ratings	7	1,013	.13	.06	.00	**.18**	.00	[.10, .26]	**.20**	.00		.11
Extraversion											**.38**	
Other-ratings	7	1,013	.26	.04	.00	**.31**	.00	[.24, .38]	**.46**	.00		.21
Self-ratings	7	1,013	.28	.07	.00	**.33**	.00	[.26, .40]	**.37**	.00		.25
Openness											**.44**	
Other-ratings	5	989	.25	.09	.00	**.37**	.00	[.28, .45]	**.58**	.00		.27
Self-ratings	5	989	.25	.09	.00	**.36**	.00	[.28, .45]	**.42**	.00		.26
Agreeableness											**.26**	
Other-ratings	7	1,013	.14	.03	.00	**.20**	.00	[.11, .28]	**.34**	.00		.15
Self-ratings	7	1,013	.16	.06	.00	**.22**	.00	[.14, .30]	**.26**	.00		.18
Conscientiousness											**.30**	
Other-ratings	7	1,013	.19	.08	.00	**.25**	.00	[.17, .33]	**.42**	.00		.19
Self-ratings	7	1,013	.18	.09	.00	**.24**	.00	[.16, .32]	**.27**	.00		.17

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Corrected correlations (ρ_{ov} and ρ_{xy}) and multiple correlations (R_{ov}) are presented in boldface for emphasis. k = number of independent samples; N = total sample size; $\bar{r}$ = mean observed correlation; SD_{obs} = observed standard deviation; SD_{resid} = standard deviation of correlations after accounting for variability from sampling error and unreliability; ρ_{ov} = operational validity, corrected for unreliability in the criterion only; $SD_{\rho_{ov}}$ = standard deviation of operational validities, corrected for variability due to sampling error and criterion unreliability; $CI_{\rho_{ov}}$ = 95% confidence interval around ρ_{ov} estimates; ρ_{xy} = true score validity, correcting for unreliability in the predictor and criterion; SD_{ρ} = standard deviation of true validities, corrected for variability due to sampling error and predictor and criterion unreliability; R_{ov} = operational multiple correlation from combining self- and one other-rating; β = standardized beta-weight in the multiple regression for other- or self-rating of the trait.

first impressions, self- and other-ratings explain comparable unique variance in trait first impressions. These results are consistent with considering self- and other-raters as essentially equivalent rating sources, with increments in validity stemming from the increased predictor reliability of adding a second, equivalent rater.

Predicting academic achievement.

Zero-order correlations. Table 10 presents validities for self- and other-reports of personality traits for predicting academic performance. Consistent with the self-report findings, Conscientiousness and Emotional Stability traits had strong validities for predicting academic achievement. However, the operational validities for other-reports of these traits were considerably larger than those for self-reports (ρ_{ov} = .41 vs. ρ_{ov} = .25 and .18 for Conscientiousness; ρ_{ov} = .27 vs. ρ_{ov} = .22 and .00 for Emotional Stability), with no overlap in confidence intervals for these traits. The interrater reliability corrected validities for other-ratings of Conscientiousness and Emotional Stability are substantial (ρ = .69 and ρ = .46, respectively), particularly when compared to the true score correlations for self-reports (ρ = .31 and .22 and ρ = .25 and .00, respectively). These results clearly indicate that Conscientiousness and Emotional Stability are traits especially relevant to academic achievement, and studies using self-reports or other-reports from only one rater will underestimate the importance of these traits.

It is interesting that other-ratings of Extraversion had particularly strong correlations for predicting academic achievement (ρ_{ov} = .35), whereas self-reports of Extraversion had considerably lower validity (ρ_{ov} = .08 and −.02). However, there was considerable variability around this average estimate for other-ratings ($SD_{\rho_{ov}}$ = .28). The small number of independent samples contributing data precluded any formal moderator analyses, but the strongest validities came from studies in which the other-rater was a high school principal, reference, or interviewer. In contrast, ratings from friends and classmates had considerably lower (and even negative) correlations with academic achievement. It may be that what principals, references, and interviewers perceive as Extraversion may be partially conflated with Conscientiousness. That is, individuals who are sociable, dominant, and energetic with principals and interviewers are pursuing goals of excelling and achieving rather than socializing. Thus, these raters' perceptions of students' Extraversion may be "contaminated" with other academic achievement-related traits, like Conscientiousness. Further research on the predictive validity of other-ratings of Extraversion in educational settings is clearly merited.

Self and one other combined. Next, we estimated the validity from combining a trait self-rating and a single other-rating for predicting academic achievement based on zero-order operational validities (ρ_{ov}) and observed self–single other correlations ($\bar{r}$) from Study 2. When other-ratings are added to self-ratings, the multiple-regression results show considerable increases in validity. These gains are most pronounced for Conscientiousness, Emotional Stability, and Extraversion. The only exception to these findings was for Openness, where self-ratings were somewhat more predictive than other-ratings. Nonetheless, these data suggest that other-ratings not only are accurate and valid for predicting academic achievement but are also more accurate than self-ratings.

Predicting job performance.

Zero-order correlations. In the domain of industrial and organizational psychology, the validity of many predictors of job performance has been meta-analytically documented (Schmidt &

Table 10
Meta-Analysis of Other-Ratings and Self-Ratings Validities for Predicting Academic Achievement

	Zero-order meta-analytic results									1 other + self (Hough)		1 other + self (Poropat)	
Trait and rating type	k	N	$\bar{r}$	SD_{obs}	SD_{resid}	ρ_{ov}	$SD_{\rho ov}$	$CI_{\rho ov}$	ρ_{xy}	SD_{ρ}	R_{ov}	β	R_{ov}	β
Emotional Stability											**.30**		**.29**	
Other-ratings	6	2,940	.25	.12	.10	**.27**	.11	[.24, .31]	**.46**	.19		.22		.31
Self-ratings—Hough	162	70,588	.20			**.22**		[.21, .23]	**.25**			.14		
Self-ratings—Poropat	104	54,462	.00			**.00**		[−.01, .01]	**.00**					−.10
Extraversion											**.36**		**.39**	
Other-ratings	7	3,081	.32	.29	.29	**.35**	.31	[.31, .38]	**.52**	.47		.38		.43
Self-ratings—Hough	128	63,057	.07			**.08**		[.07, .09]	**.09**			−.08		
Self-ratings—Poropat	103	54,072	−.02			**−.02**		[−.03, −.01]	**−.02**					−.20
Openness											**.20**		**.18**	
Other-ratings	4	1,278	.17	.14	.13	**.18**	.14	[.12, .24]	**.29**	.22		.15		.17
Self-ratings—Hough	8	3,628	.13			**.14**		[.11, .18]	**.17**			.09		
Self-ratings—Poropat	102	54,380	.07			**.08**		[.07, .09]	**.09**					.02
Agreeableness											**.01**		**.05**	
Other-ratings	6	1,460	.01	.08	.05	**.01**	.05	[−.05, .07]	**.02**	.09		.01		.00
Self-ratings—Hough	15	7,330	.01			**.01**		[−.01, .04]	**.01**			.01		
Self-ratings—Poropat	99	53,432	.04			**.05**		[.04, .06]	**.06**					.05
Conscientiousness											**.42**		**.41**	
Other-ratings	9	3,609	.37	.14	.13	**.41**	.14	[.38, .44]	**.69**	.24		.37		.40
Self-ratings—Hough	42	18,661	.23			**.25**		[.24, .27]	**.31**			.11		
Self-ratings—Poropat	127	64,867	.17			**.18**		[.17, .19]	**.22**					.03

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Meta-analytic correlations for self-ratings drawn from Hough (1992) are designated as "Self-ratings—Hough," and meta-analytic correlations for self-ratings drawn from Poropat (2009) are designated as "Self-ratings—Poropat." Corrected correlations (ρ_{ov} and ρ_{xy}) and multiple correlations (R_{ov}) are presented in boldface for emphasis. k = number of independent samples; N = total sample size; $\bar{r}$ = mean observed correlation; SD_{obs} = observed standard deviation; SD_{resid} = standard deviation of correlations after accounting for variability from sampling error and unreliability; ρ_{ov} = operational validity, corrected for unreliability in the criterion only; $SD_{\rho ov}$ = standard deviation of operational validities, corrected for variability due to sampling error and criterion unreliability; $CI_{\rho ov}$ = 95% confidence interval around ρ_{ov} estimates; ρ_{xy} = true score validity, correcting for unreliability in the predictor and criterion; SD_{ρ} = standard deviation of true validities, corrected for variability due to sampling error and predictor and criterion unreliability; R_{ov} = operational multiple correlation from combining self- and one other-rating; β = standardized beta-weight in the multiple regression for other- or self-rating of the trait.

Hunter, 1998). This research has typically found general mental ability to be the best predictor of job performance (unreliability-corrected $\rho = .50$), with other predictors such as work sample tests and structured interviews generally showing unreliability-corrected correlations between .20 and .40. In this study, we examined the validity for predicting job performance from other-ratings of personality. Validities for personality traits in predicting job performance were obtained from Barrick et al. (2001; values from independent sources only) and used as a self-rating comparator. These values for self- and other-reports are presented in Table 11.

Table 11 confirms for other-reports the typical finding from self-report studies showing that Conscientiousness predicts job performance. However, the predictive power of other-ratings of Conscientiousness is considerably greater than that of self-ratings ($\rho_{ov} = .29$ vs. $\rho_{ov} = .20$; nonoverlapping confidence intervals). In addition, other-ratings of Emotional Stability, Openness, and Agreeableness, unlike those of self-reports, showed considerable validity for predicting job performance. These results suggest that other-reports may indeed provide stronger validities for predicting job performance than do self-report measures. Note also that the true score validities expected from combining large numbers of other-raters for rating Conscientiousness, Emotional Stability, Agreeableness, or Openness are extremely high ($\rho = .55$, $\rho = .37$, $\rho = .31$, and $\rho = .45$, respectively). Indeed, these considerably exceed validities for predicting job performance from personality ratings reported in any past, large-scale research. Moreover, the validity for Conscientiousness is even larger than that for general mental ability's capacity to predict job performance (cf. Salgado, Viswesvaran, & Ones, 2002). This suggests that past research relying on a single self-rating of personality traits has by far underestimated the true importance of personality for workplace behavioral outcomes.

Self and one other combined. Again, we examined the incremental validity of combining self-reports and other-reports through multiple regression. The results in Table 11 closely mirror those found for predicting grade point average. The incremental validity from adding other-ratings to self-ratings was typically substantial. The converse, however, did not hold: Self-ratings added relatively little prediction beyond that of other-ratings alone. This pattern of findings was particularly true for Conscientiousness, Agreeableness, and Openness, where the traits were substantially related to job performance. Thus, these data lend further support to the accuracy of other-ratings and suggest they may be even more powerful than self-ratings for predicting job performance.

Summary. Study 3 summarized research examining the accuracy of other-ratings of personality traits through their capacity to predict three behaviorally based criteria: trait first impressions,

Table 11
Meta-Analysis of Other-Ratings and Self-Ratings Validities for Predicting Job Performance

	Zero-order meta-analytic results										Combined: Self + 1 other	
Trait and rating type	k	N	$\bar{r}$	SD_{obs}	SD_{resid}	ρ_{ov}	$SD_{\rho ov}$	$CI_{\rho ov}$	ρ_{xy}	SD_{ρ}	R_{ov}	β
Emotional Stability											**.19**	
Other-ratings	7	1,190	.14	.06	.00	**.17**	.00	[.10, .25]	**.37**	.00		.16
Self-ratings, Barrick et al. (2001)	224	38,817	.06			**.11**		[.09, .12]	**.12**	.08		.09
Extraversion											**.14**	
Other-ratings	6	1,135	.08	.10	.07	**.11**	.09	[.03, .18]	**.18**	.15		.09
Self-ratings, Barrick et al. (2001)	222	39,432	.06			**.11**		[.09, .12]	**.12**	.12		.09
Openness											**.22**	
Other-ratings	6	1,135	.18	.08	.00	**.22**	.00	[.15, .30]	**.45**	.00		.22
Self-ratings, Barrick et al. (2001)	143	23,225	.03			**.04**		[.02, .06]	**.05**	.11		.00
Agreeableness											**.18**	
Other-ratings	7	1,190	.13	.07	.00	**.17**	.00	[.09, .24]	**.31**	.00		.15
Self-ratings, Barrick et al. (2001)	206	36,210	.06			**.11**		[.09, .13]	**.13**	.09		.07
Conscientiousness											**.31**	
Other-ratings	7	1,190	.23	.07	.00	**.29**	.00	[.22, .36]	**.55**	.00		.25
Self-ratings, Barrick et al. (2001)	239	48,100	.12			**.20**		[.19, .22]	**.23**	.10		.11

Note. Personality measures developed outside the theoretical framework of the Big Five were coded according to the working Big Five trait taxonomy presented in Hough and Ones (2001). Meta-analytic correlations for self-ratings drawn from Barrick et al. (2001) are designated as "Self-ratings—Barrick et al." Corrected correlations (ρ_{ov} and ρ_{xy}) and multiple correlations (R_{ov}) are presented in boldface for emphasis. k = number of independent samples; N = total sample size; $\bar{r}$ = mean observed correlation; SD_{obs} = observed standard deviation; SD_{resid} = standard deviation of correlations after accounting for variability from sampling error and unreliability; ρ_{ov} = operational validity, corrected for unreliability in the criterion only; $SD_{\rho ov}$ = standard deviation of operational validities, corrected for variability due to sampling error and criterion unreliability; $CI_{\rho ov}$ = 95% confidence interval around ρ_{ov} estimates; ρ_{xy} = true score validity, correcting for unreliability in the predictor and criterion; SD_{ρ} = standard deviation of true validities, corrected for variability due to sampling error and predictor and criterion unreliability; R_{ov} = operational multiple correlation from combining self- and one other-rating; β = standardized beta-weight in the multiple regression for other- or self-rating of the trait.

academic achievement, and job performance. Although the number of independent samples contributing data to these analyses was typically small, pooling across these studies yielded total sample sizes generally between 1,000 and 4,000. The findings from these studies are encouraging indicators of the accuracy of other-reports of personality. Fairly consistently across traits and criteria, other-ratings predicted at least as well as self-ratings. In addition, other-ratings' correlations with criteria mostly followed the same basic pattern across traits as did those of self-ratings. These results suggest that other-ratings have discriminant validity in predicting criteria (i.e., other-ratings tend not to predict theoretically unrelated criteria as well as they do theoretically related criteria).

It is most interesting that criterion-related traits predicted academic achievement and job performance considerably better when ratings came from others than when they came from the self. Correlations of this degree are beyond what has ever been observed for single factors of the Big Five. Since Mount et al. (1994) published their study of other-ratings' validity for predicting job performance, researchers have frequently cited this research as presenting fascinating possibilities for applied measures of personality. But for a few exceptions, however, replication and expansion of these findings has been relatively overlooked, especially by recent challenges to the validity of personality measures in organizational settings (Morgeson et al., 2007; Murphy & Dzieweczynski, 2005; cf. Ones et al., 2007).

Several explanations may drive these advantages for other-raters' predictive power. First, it may indeed be that other-ratings are not contaminated with the response biases claimed to plague personality self-reports and that, as a result, they realize higher validities for predicting academic achievement and job performance. The absence of a parallel effect for predicting trait first impressions, however, casts some degree of doubt on this interpretation as the sole explanation.

Alternatively, these effects may be explained by the specific context in which other-raters typically knew the target. When predicting academic achievement, other-raters were individuals who knew the target primarily in an academic context (e.g., classmates, principals); when predicting job performance, other-raters generally knew the target primarily in a work context (work colleagues). Thus, this contextualized knowledge basis may enhance validity beyond self-reports for these others, whereas self-impressions are likely formed from a variety of contexts. This logic is paralleled in frame of reference approaches to measuring personality, in which personality items are contextualized by adding "at work" suffixes (Schmit, Ryan, Stierwalt, & Powell, 1995). However, support for such measurement approaches has been mixed (cf. Lievens, De Corte, & Schollaert, 2008; Small & Diefendorff, 2006). We are aware of only one study predicting job performance from other-ratings from outside the workplace: In a study of 97 targets, Hülsheger and Connelly (2010) found that

friends' ratings predicted job performance considerably more than did self-reports and to a magnitude comparable to that of meta-analytic values. Though this finding merits replication, it does not appear that other-ratings lose their advantage over self-ratings when removed from the criterion context. Rather, knowledge of work colleagues and incidental acquaintances specific to the criterion context may compensate for the generally weaker other-rating accuracy, but other-raters may still yield stronger predictive validities than self-raters. Closer research scrutiny on context-specific knowledge of personality is clearly warranted to disentangle this explanation from alternates.

Finally, Hogan's (1996; Hogan & Shelton, 1998) socioanalytic theory of personality measurement serves as another lens through which to view the relative advantage of other-ratings for predicting behaviors. This theory has distinguished between personality as internally held motives and identity and personality as externally expressed reputation. Other-ratings conceptually align closely with a target's reputation, but Hogan and colleagues have traditionally argued that self-report measures represent a form of self-presentation that also assesses a person's reputation. That is, when individuals complete a self-report measure, they consciously convey an impression that mimics the reputation they aim to convey with other behaviors. From a socioanalytic perspective, differences in validity may indicate that others' trait ratings more purely assess the reputation component of personality than do self-ratings because of a self-perception "sieve." This self-perception sieve may relate to the internally held motives and identity aspects of personality, but measuring these nonreputation components may actually contaminate and distort self-ratings when it comes to predicting behavior.

The precise cause for the relative strength of other-ratings for predicting behaviors demands further exploration across criteria and information sources. In this vein, clinical psychologists have recently highlighted similar validity advantages for predicting mental health and adjustment problems from informant descriptions of psychological disorders, particularly when multiple informants are used (Klonsky et al., 2002; Oltmanns, Melley, & Turkheimer, 2002). Though research across these domains points to relative advantages for other-ratings in predicting criteria related to behavior, this may not be the case for predicting all criteria. For example, two studies (Abe, 2004; Spain, Eaton, & Funder, 2000) have shown self-ratings to be more predictive of daily logs of emotional experiences. Precisely where and why other-ratings and self-ratings differentially relate to constructs across domains of psychology is among the most fascinating yet understudied questions in personality research today. Building such a body of knowledge will yield comprehensive understanding of what information self- and other-raters can access and how it is accessed.

General Discussion

Studies 1, 2, and 3 represent a wide span of analyses across many criteria. Here, we integrate our findings by addressing four overarching research questions: (a) how accurate are others' ratings of personality, (b) what are good traits for rating accuracy, (c) what is good information for rating accuracy, and (d) how accurate are other-ratings compared to self-ratings?

Accuracy of Other-Ratings

The accuracy of personality ratings from self-raters and from other-raters is an intriguing and long-debated topic. The results from Studies 1, 2, and 3 provide general support for the accuracy of other-ratings, with ratings from well-acquainted observers clearly surpassing Mischel's (1968) .30 validity barrier. Moreover, other-ratings predict targets' behaviors: behaviors observed by strangers, behaviors determining academic achievement, and behaviors related to job performance. These meta-analyses show that other-ratings are clearly linked to targets' personality traits and that targets behave consistently enough for other-raters to rate their personality accurately.

At the same time, these results do not indicate a complete redundancy of information across raters. Just as personality measures include multiple items to enhance reliability and construct coverage, personality ratings from multiple raters must be assessed to improve research reliability and validity. Based on estimates from Study 1, the interrater reliability for the best traits rated by the best information sources reaches .80 only when five raters are combined. For more typical traits or more typical information sources, nine or 10 raters should be combined to reach the same .80 level of interrater reliability. Although what observer-ratings measure is clearly part of the target's personality, one observer's rating is only one angle from which to assess that target's personality. Personality traits are hard to measure, and measuring them requires an assembly of multiple ratings. Researchers generally find single-item measures of personality inadequate due to their modest reliability, but this reasoning has not been extended to evaluating study designs using single raters, though the logic is identical. The more that researchers realize the psychometric payback and necessity of using multiple others, multiple scales, and/or multiple administrations to measure these personality traits, the stronger the field's predictions, explanatory power, and usefulness will be. These results afford researchers confidence, but not relaxation, in measuring personality traits with other-ratings.

What Makes for Good Traits?

The idea that some traits are easier for other-raters to perceive accurately has been a central tenet of person perception. Highly visible traits and nonevaluative traits should be rated more accurately by others. Extraversion and Conscientiousness—two behaviorally centered traits—typically had the greatest interrater reliabilities and self–other correlations, whereas accuracy was weaker for Emotional Stability and Agreeableness. These differences across traits were more pronounced when other-raters were less well acquainted. The findings indicate that Emotional Stability cues are more difficult to detect and utilize and that these cues are not particularly likely to be relevant and available unless other-raters are intimately acquainted with targets. This is not surprising: As acquaintance and intimacy increase, individuals gain greater access to internal thoughts and emotions of targets and the moderating effects of trait visibility decrease, especially when corrected for difficulty in detecting and utilizing cues.

The lower accuracy for perceiving Agreeableness potentially offers some support for the effect of trait evaluativeness on accuracy, though the high evaluativeness of Agreeableness may operate through different mechanisms than previously hypothesized.

Funder (1995) argued that high trait evaluativeness would weaken accuracy because targets would suppress evaluative trait information (i.e., trait evaluativeness affects the RA component of the accuracy process). Findings from Study 2 showed that the lower accuracy for Agreeableness was actually due to weaker DU of the raters, perhaps because forming impressions of the target's Agreeableness is more susceptible to idiosyncrasies in how well raters like the target. Future research using a broader set of traits (e.g., facets of the Big Five) may better illuminate any effects of trait evaluativeness on accuracy, once taxonomic refinements of Big Five facets become available (cf. DeYoung, Quilty, & Peterson, 2007; Hough & Ones, 2001; Roberts, Chernyshenko, Stark, & Goldberg, 2005).

Note, however, that less accurate traits identified in Studies 1 and 2 do not indicate inaccurate traits. In Study 3, even "bad" traits such as Emotional Stability, Openness, and Agreeableness were strongly predictive of behaviors and behavioral outcomes. Thus, it is not the case that other-ratings of these traits are inaccurate, given that they still predict relevant behaviors (perhaps the ultimate accuracy criterion). However, finding weak predictive validities for some traits in Study 3 does not indicate inaccuracy, either: The criterion simply may not be related to the trait construct. Thus, results from Studies 1 and 2 are most useful for directly comparing which traits can be rated more or less accurately than others, but only the inability to predict relevant criteria indicates true inaccuracy itself.

What Makes for Good Information?

These meta-analyses have also focused on the effects that having good information has on producing accurate ratings of another's personality. Our results show clear advantages for having increased acquaintance with the target. Although interrater reliability may be enhanced by having observers watch targets only in limited but common situations, having broader, cross-situational opportunities for observing the target produced considerable gains in self–other accuracy for rating all traits. Thus, ratings of strangers with this limited observation period were clearly less accurate than those of more closely acquainted other-raters.

In addition, interpersonal intimacy with the target produced further gains in interrater and self–other accuracy. Accuracy was greatest when other-raters were spouses or dating partners. Other family members and friends had slightly lower accuracy and were followed by work colleagues and incidental acquaintances. These findings for work colleagues are quite telling: Given the typical workweek, these individuals likely have the greatest opportunity to observe targets. The results suggest that, after a certain point in observing the target, the self-disclosure associated with interpersonally intimate relationships is necessary for improving other-rater accuracy for all traits. Thus, among acquainted individuals, quality of observation appears to count more than does quantity. Nonetheless, work colleagues' ratings in Study 3 still were strongly predictive of targets' job performance (considerably more strongly predictive than were self-ratings). Thus, the inaccuracy of work colleagues described above is purely relative to the accuracy of other types of other-raters and is by no means indicative of work colleagues' ratings of personality being wholly inaccurate. Explicit comparisons of the predictive validities of other-ratings from work colleagues versus ratings from different information sources are needed for more fully understanding work colleagues' accuracy. However, Studies 1 and 2 serve as useful tools for generating hypotheses about the relative validity of information sources for predicting behaviors.

Finally, these studies also examined what cues may represent good types of information in studies where strangers rated targets' personality based on particular instances of targets' behavior (or particular objects). Study 1 suggested some distinctions among information type effects on accuracy. Interrater reliability was higher for Emotional Stability and Extraversion when rating stimuli involving audio information and for Conscientiousness and Openness when personal objects served as stimuli. In addition, raters agreed more when behaviors were naturally occurring rather than constrained. However, these differences in interrater reliability did not translate into differences in self–other accuracy. Across traits, little variability remained in self–stranger correlations to allow for differences in information type. These findings held even for stranger ratings of Extraversion, where self–other accuracy was stronger. These results suggest that though differences in information type may affect consensus between raters, the information available to strangers is already so limited that these differences do not substantially impact self–other correlations. This pattern of findings may hold intriguing social ramifications as well. Even though observers may quickly reach agreement in judging a stranger, that agreement may breed overconfidence in observers' judgments. Developing these judgments directly mirrors the agreement, overconfidence, and fascination with gossip and speculation that permeate most social circles.

Are Other-Ratings as Accurate as Self-Ratings or Even More Accurate?

Study 3 examined the relative accuracy of self- and other-ratings through comparison of self- and other-reports' validity for predicting behavior. We found considerably stronger validities for other-ratings in predicting academic achievement and job performance but not first impressions. Thus, these results suggest that at least some other-ratings may be more accurate than are self-ratings. If these results hold across additional types of others and additional criteria, these results have major implications for personality theory and applications across many domains of psychology. Such findings would point to a clear need to understand differences in self-perception processes and other-perception processes in rating personality. Moreover, revisiting self-report findings aligning traits with criteria throughout psychology would be necessary.

Understanding Other-Rating Accuracy: Pragmatic Implications and Future Research

Daily life brings us into contact with more people across greater expanses now than ever before in history. We no longer just call, e-mail, and chat but friend, text, tweet, blog, poke, and wink as a basis for maintaining social networks. Regardless of the medium used, people involved in this mass of social interactions make decisions about with whom and how to initiate, maintain, and extend personal and working relationships based in part on perceptions of others' personality characteristics. Findings from these studies show that, although some accuracy is possible from a distance, the strongest accuracy comes only from building a close

relationship with the person. Perhaps even more concerning is that the two traits generally most relevant for interpersonal relationships—Emotional Stability and Agreeableness—tend to be the hardest to perceive. As interactions become more removed, the increased difficulty in perceiving others' traits may inhibit our ability to choose relationships wisely, anticipate, and adjust our own behaviors accordingly.

For the community of researchers studying person perception processes, these three studies provide an integrative knowledge base about the accuracy of other-ratings and also point toward several future directions. First, we organized our results for trait and information source moderators according to particular relationship categories and traits measured rather than according to underlying dimensions potentially differentiating between information sources and traits (i.e., work colleagues rather than high opportunity to observe other-raters and Extraversion rather than high-visibility trait). In part, the decision not to code moderators according to these underlying dimensions was based on studies providing only limited descriptions of traits and other-raters. More important, though, guiding principles about other-rating accuracy emerging from this research are likely more useful to personality researchers in many disciplines when phrased concretely as "other-rating accuracy for Emotional Stability is high only when raters are family members, friends, or roommates" than as "intimacy enhances accuracy for low-visibility traits." One potential downside of this approach is that it may be somewhat unclear which dimensions may drive moderator findings. Researchers focusing purely on perception processes in rating others may benefit from attending directly to dimensions underlying trait and information moderators. For a more general audience of psychologists interested in personality research, however, coding information source and trait categories rather than underlying dimensions yields information that is both more accessible and more closely linked to original empirical sources.

In addition, the number of independent samples contributing data across studies indicates a clear need for additional research on the validity of other-ratings for predicting behaviors and behavioral outcomes. The capacity of other-ratings to predict relevant external criteria is even more telling than the alignment of trait perceptions across raters as an indicator of accuracy. Research addressing other-rating's behavioral prediction has only begun to explore the potential strengths and limitations of other-ratings. Moreover, comparisons of other-ratings to self-ratings promise insight into the nature of self-perception and refinement of existing personality theory. Knowledge in many domains of personality research is likely to benefit from development of a broader research base with multisource ratings of personality.

The somewhat surprisingly lower accuracy for work colleagues and classmates suggests an important area for further exploration. This lower accuracy was attributed to a relative lack of interpersonal intimacy with the target. Nonetheless, Study 3 showed that personality ratings even from peers at work and school do predict performance in these spheres quite well. On one hand, it may be that individuals with greater intimacy might produce even stronger predictions than do work colleagues of behaviors such as job performance. On the other hand, it is possible that this interpersonal intimacy could be a contaminating factor that falsely aligns other-ratings toward self-misperceptions. These findings merit close scrutiny, particularly in studying observers' ratings predicting cross-context behaviors. An ideal design for disentangling these effects would collect ratings of targets from information sources in fully crossed Interpersonal Intimacy (high vs. low) × Context-Specific Knowledge (high vs. low) interactions. Comparing predictive validities within these cells would indicate any advantages of intimacy, context, or their interaction. Indeed, amid such general need for studies in which other-ratings are used to predict behaviors, these cross-context predictions could be particularly informative.

Research will benefit from further qualitative studies on the sources of discrepancies between self- and other-ratings. McCrae et al. (1998) conducted one such study examining self–spouse disagreements and found most disagreements on items were idiosyncrasies in interpreting items or in considering specific instances of behaviors. However, no qualitative research has yet examined sources of disagreement using nonspouse information sources, and more substantive disagreements might be expected if less intimately acquainted raters are used. Capturing the reasons for these disagreements affords a basis for building theory about the relative accuracy of self- and other-ratings.

Potential Applications of Other-Ratings of Personality Across Psychology Research Fields

In the opening of this paper, we noted the litany of fields of psychology in which personality research has been applied, with traits generally being measured via self-reports. The meta-analyses presented illustrate two concrete ways that using other-ratings can enhance personality research in these fields. First, these results indicate that single raters have pronounced idiosyncrasies in how they view targets, whether raters are observers (Study 1) or self-raters (Study 2). These idiosyncrasies weaken the relationships observed between personality traits and other constructs (just as other sources of measurement error weaken correlations), but researchers can mitigate these effects by soliciting ratings from multiple others (just as increasing the number of items in a measure improves its reliability). However, studies that use only self-ratings will inherently be limited, because a single rating will be idiosyncratic. This is a lesson frequently glossed over, but these changes in predictive power are not trivial. For example, when predicting job performance from Conscientiousness, estimates of the overlap in these constructs changed from $\rho_{o=1} = .29$ to $\rho_{o=\infty} = .55$ depending on whether one or a large number of other-raters were used. Some research in behavioral genetics, where multisource ratings are common, represents a notable exception. This field has found considerably stronger genetic effects on personality traits because the error associated with rater idiosyncrasies is reduced (Bouchard & Loehlin, 2001). Following suit in other research fields is likely to advance theory development and avoid undervaluing the relevance of personality traits.

Second, Study 3 presents preliminary evidence from two fields (educational and industrial and organizational psychology) that some behavioral constructs relate more strongly to other-ratings than to self-ratings. These findings may not necessarily hold across all domains of psychology. For example, finding the opposite advantage when studying how personality traits relate to daily reports of emotions (Spain et al., 2000) suggests that self- and other-ratings may assess valid but different realms of personality. Such points at which self- and other-ratings differentially relate to

constructs provide major stepping-stones for theory in studying personality both within specific fields of psychology and broadly across fields. Indeed, research on psychopathological traits by Oltmanns and colleagues (Klonsky et al., 2002; Oltmanns, Friedman, Fiedler, & Turkheimer, 2004; Oltmanns et al., 1998) has illustrated the value to be gained in clinical psychology by systematically comparing self- and other-ratings. In other fields where personality measurement via other-raters is scant, continued reliance on only self-reports will leave researchers blind to such fascinating and potentially useful findings.

Beyond highlighting potential contributions, these meta-analyses provide an empirically based theoretical and methodological framework for how researchers across these fields can best incorporate other-ratings. These results indicate that family members, friends, and roommates are the best choices for achieving accurate other-ratings. It is especially important to use such other-raters when rating low-visibility traits like Emotional Stability or Openness. Even when more intimate other-raters are used, researchers would likely need to collect other-ratings from five individuals to achieve generally accepted minimum standards of interrater reliability. For instance, suppose researchers were correlating friends' ratings of Extraversion (the trait and information source with the highest interrater reliability in Table 3) with a criterion for which the true correlation is $\rho_{o=\infty} = .50$ and the criterion is measured without error. With only one friend's ratings of Extraversion, the observed correlation would drop to $r = .34$ ($r = .40$ for two raters, .42 for three raters, .44 for four raters, and .45 for five raters). These represent substantively different conclusions about the importance of Extraversion, and the decreases with fewer raters are even more pronounced for other traits and information sources. Of course, researchers may be interested in collecting other-raters with context-specific knowledge of targets (e.g., work colleagues, support-group comembers). We encourage such research as well, but because such sources tend to be idiosyncratic, researchers should be even more cognizant of collecting ratings from multiple sources when drawing from these less accurate observers (at least seven or eight raters would be recommended for most traits).

Conclusion

Even as research on other-ratings was beginning 80 years ago, Shen (1925) prophetically commented that "it is always an interesting question as to whether an individual can know himself better than he knows his associates" (p. 104). Ensuing research on individuals' perceptions of another's personality has spanned decades, agendas, and types of relationships. Despite its breadth, this research finds a common framework in studying other-rating accuracy. Our purpose in these three meta-analyses was to provide a large-scale quantitative evaluation of the accuracy of other-ratings and of necessary and enhancing conditions for their accuracy. These meta-analyses provide an integrated view of other-rating accuracy research with three important take-home messages for researchers studying personality. First, traits appear to be readily expressed (high RA) to those intimately acquainted with targets, but considerably less trait expression is afforded to those less intimately acquainted with targets (even when interactions with a target are frequent). Second, despite this strong trait expression, other-raters are considerably idiosyncratic in how they view the target (modest DU), especially in rating traits low in visibility and high in evaluativeness. These findings necessitate that research solicit ratings from multiple other-raters. Third, other-ratings assess traits more validly than do self-ratings for predicting at least some important criteria (e.g., academic and job performance). The root causes and breadth of such differential predictions represent fascinating directions for personality theory and application throughout psychology. On the whole, there is extraordinary value in collecting other-reports to measure personality. Our study provides a starting ground for the future contribution of other-ratings of personality traits across psychology's many domains.

References

For a list of references contributing data to the meta-analyses and for lists of references included in Tables 1 and 2, go to http://dx.doi.org/10.1037/a0021212.supp

Abe, J. A. (2004). Shame, guilt, and personality judgment. *Journal of Research in Personality, 38,* 85–104. doi:10.1016/S0092-6566(03)00055-2

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9,* 9–30. doi:10.1111/1468-2389.00160

Bernieri, F. J., Zuckerman, M., Koestner, R., & Rosenthal, R. (1994). Measuring person perception accuracy: Another look at self–other agreement. *Personality and Social Psychology Bulletin, 20,* 367–378. doi:10.1177/0146167294204004

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117,* 187–215. doi:10.1037/0033-2909.117.2.187

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin, 130,* 887–919. doi:10.1037/0033-2909.130.6.887

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

Bouchard, T. J., & Loehlin, J. C. (2001). Genes, evolution, and personality. *Behavioral Genetics, 31,* 243–273. doi:10.1023/A:1012294324713

Cassin, S. E., & von Ranson, K. M. (2005). Personality and eating disorders: A decade in review. *Clinical Psychology Review, 25,* 895–916. doi:10.1016/j.cpr.2005.04.012

Chaplin, W. F., & Goldberg, L. R. (1984). A failure to replicate the Bem and Allen study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology, 47,* 1074–1090. doi:10.1037/0022-3514.47.5.1074

Connelly, B. S., & Ones, D. S. (2008, April). *Interrater unreliability in assessment center ratings: A meta-analysis.* Paper presented at the meeting for the Society for Industrial and Organizational Psychologists, San Francisco, CA.

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment, 15,* 110–117. doi:10.1111/j.1468-2389.2007.00371.x

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579. doi:10.1037/0021-9010.80.5.565

Costa, P. T., & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology, 54,* 853–863. doi:10.1037/0022-3514.54.5.853

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13,* 653–665. doi:10.1016/0191-8869(92)90236-I

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64,* 21–50. doi:10.1207/s15327752jpa6401_2

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology, 91,* 1138–1151. doi:10.1037/0022-3514.91.6.1138

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93,* 880–896. doi:10.1037/0022-3514.93.5.880

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417–440. doi:10.1146/annurev.ps.41.020190.002221

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73,* 1246–1256. doi:10.1037/0022-3514.73.6.1246

Dilchert, S., Ones, D. S., Van Rooy, D. L., & Viswesvaran, C. (2006). Big Five factors of personality. In J. H. Greenhaus & G. A. Callahan (Eds.), *Encyclopedia of career development* (pp. 36–42). Thousand Oaks, CA: Sage.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91,* 40–57. doi:10.1037/0021-9010.91.1.40

Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., . . . Belmaker, R. H. (1996). Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of novelty seeking. *Nature Genetics, 12,* 78–80. doi:10.1038/ng0196-78

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84,* 155–166. doi:10.1037/0021-9010.84.2.155

Eysenck, H. J. (1993). "The structure of phenotypic personality traits": Comment. *American Psychologist, 48,* 1299–1300. doi:10.1037/0003-066X.48.12.1299.b

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. *Personnel Psychology, 61,* 579–616. doi:10.1111/j.1744-6570.2008.00123.x

Friedman, H. S., & Booth-Kewley, S. (2003). The "disease-prone personality": A meta-analytic view of the construct. In P. Salovey & A. J. Rothman (Eds.), *Social psychology of health* (pp. 305–324). New York, NY: Psychology Press.

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. (1999). *Personality judgment: A realistic approach to person perception.* San Diego, CA: Academic Press.

Funder, D. C., & West, S. G. (1993). Consensus, self–other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality, 61,* 457–476. doi:10.1111/j.1467-6494.1993.tb00778.x

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34. doi:10.1037/0003-066X.48.1.26

Heller, D., Watson, D., & Hies, R. (2004). The role of person versus situation in life satisfaction: A critical examination. *Psychological Bulletin, 130,* 574–600. doi:10.1037/0033-2909.130.4.574

Hogan, R. (1996). A socioanalytic interpretation of the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality* (pp. 163–179). New York, NY: Guilford.

Hogan, R., & Shelton, D. (1998). A socioanalytic perspective on job performance. *Human Performance, 11,* 129–144. doi:10.1207/s15327043hup1102&3_2

Hough, L. M. (1992). The "Big Five" personality variables–construct confusion: Description versus prediction. *Human Performance, 5,* 139–155. doi:10.1207/s15327043hup0501&2_8

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595. doi:10.1037/0021-9010.75.5.581

Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology: Vol. 1. Personnel psychology* (pp. 233–277). Thousand Oaks, CA: Sage.

Hülsheger, U. R., & Connelly, B. S. (2010, April). *Validity of observer ratings with raters from outside the workplace.* Paper presented at the meeting of the Society for Industrial and Organizational Psychologists, Atlanta, GA.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8,* 275–292. doi:10.1111/1468-2389.00156

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61,* 521–551. doi:10.1111/j.1467-6494.1993.tb00781.x

Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist, 34,* 107–117. doi:10.1037/0003-066X.34.2.107

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87,* 530–541. doi:10.1037/0021-9010.87.3.530

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98,* 155–163. doi:10.1037/0033-295X.98.2.155

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

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin, 102,* 390–402. doi:10.1037/0033-2909.102.3.390

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the Big Five. *Psychological Bulletin, 116,* 245–258. doi:10.1037/0033-2909.116.2.245

Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 142–182). Orlando, FL: Academic Press.

Kenny, D. A., & Winquist, L. (2001). The measurement of interpersonal sensitivity: Consideration of design, components, and unit of analysis. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 265–302). Mahwah, NJ: Erlbaum.

Kenrick, D. T., & Funder, D. C. (1988). Lessons from the person–situation debate. *American Psychologist, 43,* 23–34. doi:10.1037/0003-066X.43.1.23

Klonsky, E., Oltmanns, T. F., & Turkheimer, E. (2002). Informant-reports of personality disorder: Relation to self-reports and future research directions. *Clinical Psychology: Science and Practice, 9,* 300–311. doi:10.1093/clipsy/9.3.300

Krueger, R. F., Caspi, A., Moffitt, T. E., Silva, P. A., & McGee, R. (1996). Personality traits are differentially linked to mental disorders: A multitrait–multidiagnosis study of an adolescent birth cohort. *Journal of Abnormal Psychology, 105,* 299–312. doi:10.1037/0021-843X.105.3.299

Krueger, R. F., & Tackett, J. L. (2003). Personality and psychopathology: Working toward the bigger picture. *Journal of Personality Disorders, 17,* 109–128. doi:10.1521/pedi.17.2.109.23986

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181. doi:10.1037/0033-2909.127.1.162

Kurtz, J. E., & Sherker, J. L. (2003). Relationship quality, trait similarity, and self–other agreement on personality ratings in college roommates. *Journal of Personality, 71,* 21–48. doi:10.1111/1467-6494.t01-1-00005

Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist, 64,* 241–256. doi:10.1037/a0015309

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93,* 268–279. doi:10.1037/0021-9010.93.2.268

Malloy, T. E., & Kenny, D. A. (1986). The social relations model: An integrative method for personality research. *Journal of Personality, 54,* 199–225. doi:10.1111/j.1467-6494.1986.tb00393.x

Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2005). The relationship between the five-factor model of personality and symptoms of clinical disorders: A meta-analysis. *Journal of Psychopathology and Behavioral Assessment, 27,* 101–114. doi:10.1007/s10862-005-5384-y

Marcus, B., Machilek, F., & Schütz, A. (2006). Personality in cyberspace: Personal websites as media for personality expressions and impressions. *Journal of Personality and Social Psychology, 90,* 1014–1031. doi:10.1037/0022-3514.90.6.1014

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51,* 882–888. doi:10.1037/0022-006X.51.6.882

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81–90. doi:10.1037/0022-3514.52.1.81

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52,* 509–516. doi:10.1037/0003-066X.52.5.509

McCrae, R. R., Costa, P. T., Jr., Martin, T. A., Oryol, V. E., Rukavishnikov, A. A., Senin, I. G., . . . Urbánek, T. (2004). Consensual validation of personality traits across cultures. *Journal of Research in Personality, 38,* 179–201. doi:10.1016/S0092-6566(03)00056-4

McCrae, R. R., Stone, S. V., Fagan, P. J., & Costa, P. T., Jr. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality, 66,* 285–313. doi:10.1111/1467-6494.00013

Mischel, W. (1968). *Personality and assessment.* New York, NY: Wiley.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683–729. doi:10.1111/j.1744-6570.2007.00089.x

Mount, M. K., Barrick, M. R., & Strauss, J. (1994). Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology, 79,* 272–280. doi:10.1037/0021-9010.79.2.272

Munafò, M. R., Clark, T. G., Moore, L. R., Payne, E., Walton, R., & Flint, J. (2003). Genetic polymorphisms and personality in healthy adults: A systematic review and meta-analysis. *Molecular Psychiatry, 8,* 471–484. doi:10.1038/sj.mp.4001326

Murphy, K. R., & Dzieweczynski, J. L. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance? *Human Performance, 18,* 343–357. doi:10.1207/s15327043hup1804_2

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice Hall.

Nunnally, J. C. (1978). *Psychometric theory.* New York, NY: McGraw-Hill.

Oltmanns, T. F., Friedman, J. N., Fiedler, E. R., & Turkheimer, E. (2004). Perceptions of people with personality disorders based on thin slices of behavior. *Journal of Research in Personality, 38,* 216–229. doi:10.1016/S0092-6566(03)00066-7

Oltmanns, T. F., Melley, A. H., & Turkheimer, E. (2002). Impaired social functioning and symptoms of personality disorders assessed by peer and self-report in a nonclinical population. *Journal of Personality Disorders, 16,* 437–452. doi:10.1521/pedi.16.5.437.22123

Oltmanns, T. F., Turkheimer, E., & Strauss, M. E. (1998). Peer assessment of personality traits and pathology in female college students. *Assessment, 5,* 53–65. doi:10.1177/107319119800500108

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60,* 995–1027. doi:10.1111/j.1744-6570.2007.00099.x

Ones, D. S., & Viswesvaran, C. (2001). Personality at work: Criterion-focused occupational personality scales used in personnel selection. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace: Decade of behavior* (pp. 63–92). Washington, DC: American Psychological Association. doi:10.1037/10434-003

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81,* 660–679. doi:10.1037/0021-9010.81.6.660

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46,* 598–609. doi:10.1037/0022-3514.46.3.598

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60,* 307–317. doi:10.1037/0022-3514.60.2.307

Paulhus, D. L., & Trapnell, P. D. (2008). Self-presentation of personality: An agency-communion framework. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 492–517). New York, NY: Guilford.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135,* 322–338. doi:10.1037/a0014996

Riemann, R., Angleitner, A., & Strelau, J. (1997). Genetic and environmental influences on personality: A study of twins reared together using the self- and peer-report NEO-FFI scales. *Journal of Personality, 65,* 449–475. doi:10.1111/j.1467-6494.1997.tb00324.x

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of Conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58,* 103–139. doi:10.1111/j.1744-6570.2005.00301.x

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2,* 313–345. doi:10.1111/j.1745-6916.2007.00047.x

Salgado, J. F., & Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in validity studies of personnel selection. *Perceptual and Motor Skills, 83,* 1195–1201.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2002). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson & H. K. Sinangil (Eds.), *Handbook of industrial, work and organizational psychology* (pp. 165–199). Thousand Oaks, CA: Sage. doi:10.2466/PMS.83.7.1195-1201

Saulsman, L. M., & Page, A. C. (2004). The five-factor model and personality disorder empirical literature: A meta-analytic review. *Clinical Psychology Review, 23,* 1055–1085. doi:10.1016/j.cpr.2002.09.001

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8,* 467–487. doi:10.1002/ejsp.2420080405

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection

methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274. doi:10.1037/0033-2909.124.2.262

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27,* 183–198. doi:10.1016/S0160-2896(99)00024-0

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8,* 206–224. doi:10.1037/1082-989X.8.2.206

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *The British Journal of Mathematical and Statistical Psychology, 62,* 97–128. doi:10.1348/000711007X255327

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame of reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80,* 607–620. doi:10.1037/0021-9010.80.5.607

Shen, E. (1925). The validity of self-estimate. *Journal of Educational Psychology, 16,* 104–107. doi:10.1037/h0073590

Shrauger, J. S., & Schoeneman, T. J. (1979). Symbolic interactionist view of self-concept: Through the looking glass darkly. *Psychological Bulletin, 86,* 549–573. doi:10.1037/0033-2909.86.3.549

Small, E. E., & Diefendorff, J. M. (2006). The impact of contextual self-ratings and observer ratings of personality on the personality–performance relationship. *Journal of Applied Social Psychology, 36,* 297–320. doi:10.1111/j.0021-9029.2006.00009.x

Spain, J. S., Eaton, L. G., & Funder, D. C. (2000). Perspectives on personality: The relative accuracy of self versus others for the prediction of emotion and behavior. *Journal of Personality, 68,* 837–867. doi:10.1111/1467-6494.00118

Starzyk, K. B., Holden, R. R., Fabrigar, L. R., & MacDonald, T. K. (2006). The Personal Acquaintance Measure: A tool for appraising one's acquaintance with any person. *Journal of Personality and Social Psychology, 90,* 833–847. doi:10.1037/0022-3514.90.5.833

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin, 133,* 65–94. doi:10.1037/0033-2909.133.1.65

Sullivan, P. S. (1995). *Interpersonal perception in dyadic and group settings: A social relations analysis* (Unpublished doctoral dissertation). University of Connecticut, Storrs.

Trapmann, S., Hell, B., Hirn, J. O. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Zeitschrift für Psychologie/Journal of Psychology, 215,* 132–151. doi:10.1027/0044-3409.215.2.132

Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality, 40,* 472–481. doi:10.1016/j.jrp.2005.03.003

Vernon, P. (1933). Some characteristics of the good judge of personality. *Journal of Social Psychology, 4,* 42–58.

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60,* 224–235. doi:10.1177/00131640021970475

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574. doi:10.1037/0021-9010.81.5.557

Zillig, L. M. P., Hemenover, S. H., & Dienstbier, R. A. (2002). What do we assess when we assess a Big Five trait? A content analysis of the affective, behavioral, and cognitive processes represented in the Big Five personality inventories. *Personality and Social Psychology Bulletin, 28,* 847–858. doi:10.1177/0146167202289013

Received July 30, 2009
Revision received June 21, 2010
Accepted June 26, 2010 ■