Journal of Research in Personality 70 (2017) 107–121

Contents lists available at ScienceDirect

Journal of Research in Personality

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

Full Length Article

Evaluating the predictive validity of personality trait judgments using a naturalistic behavioral criterion: A preliminary test of the self-other knowledge asymmetry model

Andrew Beer [a,*], Simine Vazire [b]

[a] University of South Carolina Upstate, United States
[b] University of California Davis, United States

ARTICLE INFO

Article history:
Received 10 August 2016
Revised 20 June 2017
Accepted 21 June 2017
Available online 22 June 2017

Keywords:
Personality
Naturalistic observation
Validity
Self-knowledge
Zero acquaintance
Informant ratings

ABSTRACT

We tracked 87 participants over two days using the Electronically Activated Recorder (EAR). Coded variables included expressions of mood, amount of talking in various situations (e.g., with one other person, with a friend, etc.), locations, and behavioral markers of the Big Five. Collection of self-, informant-, and stranger-ratings on markers of the Big Five allowed for a unique test of the Self-Other Knowledge Asymmetry (SOKA) model. Although effect sizes were modest, there was evidence for the validity of both self- and informant-ratings across most trait dimensions. Stranger-ratings showed evidence of validity in the domain of Extraversion. Predictions derived from the SOKA model were partially supported, though more research with larger samples is needed to provide stronger tests of SOKA.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Personality traits are invisible – putative latent constructs that drive consistent individual differences in thought, feeling, and behavior. Thus, from a purely objective standpoint, they must generally be inferred rather than directly observed. Despite this theoretical thorniness, attempts to systematically catalog and effectively assess broad dimensions of individual differences have spanned 80 years—much of the time that psychology, as a formal discipline, has existed. The most common method of personality trait assessment has been self-report (Vazire, 2006), founded on the beliefs that individuals (a) have near complete access to their own behavior, and (b) have unique access to their mental states, including their motivations, intentions, and internal emotional states. Indeed, these advantages frequently produce circumstances under which self-report measures of personality traits predict meaningful life outcomes (Ozer & Benet-Martínez, 2006) and everyday behaviors (Mehl, Gosling, & Pennebaker, 2006).

However, the use of self-reports as a primary source of information about personality certainly leaves some things to be desired (for an extensive review of concerns with self-assessments of various sorts, see Dunning, Heath, & Suls, 2004; for a review of self-reports of personality in particular, see Back & Vazire, 2012; Paulhus & Vazire, 2007). Although individuals have access to most of their behavior, some more automatic or unconscious aspects of behavior may go unnoticed by the actor but be quite impactful on his or her environment and thus an important aspect of one's personality. For example, evading eye contact in personal interactions is an act about which the actor could easily be unaware but may influence interaction partners quite a bit. Moreover, when forming impressions of one's own behavior, people may be overly sensitive to the thoughts and feelings they were having, and place too little emphasis on their overt actions (Andersen & Ross, 1984). Indeed, much has been written about limitations of the actor's perspective for observing his/her own behavior (Jones & Nisbett, 1971; Malle, 2006; Robins, Spranca, & Mendelsohn, 1996; Watson, 1982).

People may also have biases that distort the accuracy of their self-views, even when they have perfect access to the information they would need to form accurate self-views. There is now a fairly large body of evidence suggesting that self-enhancement is quite common (Alicke, 1985; Kwan et al., 2011; Taylor & Brown, 1988), but that there are also important individual differences in self-bias (Paulhus, 1984; Paulhus & John, 1998). Importantly, the fact that people do not all share the same direction and level of bias about themselves is even more problematic for the accuracy of self-perceptions (Vazire, 2010). If everyone self-enhanced, and did so to more or less the same degree, this would inflate the

* Corresponding author.
E-mail address: abeer@uscupstate.edu (A. Beer).

http://dx.doi.org/10.1016/j.jrp.2017.06.004
0092-6566/© 2017 Elsevier Inc. All rights reserved.

absolute level of people's self-ratings, but the between-person rank-order accuracy of self-ratings would be intact. Instead, the fact that some people self-enhance, others self-deprecate, and still others are relatively unbiased (Bollich, Rogers, & Vazire, 2015), means that the between-person rank-ordering of people's self-views no longer matches up with the rank-ordering of their actual standing on a trait. Additionally, there are some circumstances in which individuals may have perfectly accurate self-views but may be motivated to willfully and knowingly misrepresent themselves if they believe something may be gained by doing so (e.g., assessments by current or prospective employers, dating website profiles, etc.).

For these and other reasons, psychologists will frequently turn to others for personality information regarding a given target individual. The rationale is that knowledgeable informants (e.g., friends, spouses, co-workers, roommates, etc.) have access to a wide variety of behaviors over time and across situations. Moreover, while close others certainly generate biased personality perceptions their biases tend to be more uniformly positive than self-biases and thus less disruptive of the between-person rank ordering on a given trait (Leising, Erbs, & Fritz, 2010; Leising, Gallrein, & Dufner, 2014). Indeed, informant judgments of personality do predict behavior (Connelly & Ones, 2010; Hofstee, 1994; Kolar, Funder, & Colvin, 1996). However, informants are not without their blind spots either. For example, Vazire (2010) found that friends' ratings of personality were less predictive of behavior than self-ratings for traits low in observability (e.g., neuroticism).

Since it appears that personality ratings made by the self or knowledgeable informants cannot be taken as completely valid on their own, the general temptation has been to simply aggregate across sources of data to achieve the most valid estimate of an individual's standing on a given trait domain (see Letzring, Wells, & Funder, 2006, for an extensive discussion). This logic is generally sound, and in a perfect world, we would simply collect massive amounts of data from multiple sources. However, this is a laborious process for those interested in practically applying the research findings, and in terms of theory, it seems clear that aggregation only buys predictive validity in some circumstances—in others, one source or another predicts a relevant outcome just as well on its own (Vazire & Mehl, 2008). In addition, to the extent that some of the variation across judges is non-random (e.g., a positive evaluative bias), these errors would be compounded or sustained—not eliminated—in an aggregate.

Considerations such as these require a more careful analysis of sources of personality data. Recently, Vazire (2010) put forth a general model to aid in determining which data source might be the most valid for a given trait assessment, based on previous work by John and Robins (1993) and Luft and Ingham (1955). The Self-Other Knowledge Asymmetry (SOKA) model employs two primary dimensions—which can be considered as properties of the traits themselves—to explicate the issue of source validity. The first consideration is the observability of the trait. Traits with clear, frequent, and publicly available behavioral manifestations should be judged quite accurately by knowledgeable informants. In extreme cases, such as extraversion, it is reasonable to assume that aggregating a few independent judgments from someone almost entirely unacquainted with the target individual may yield a fairly accurate estimate. On the other hand, traits defined more by internal affective or cognitive aspects, such as neuroticism, should be judged more accurately by the target herself than by others. However, these predictions must be qualified based on the second factor in the SOKA model, evaluativeness, or the extent to which the trait in question has a clearly socially desirable pole. For example, high agreeableness (warmth, compassion) is a quality generally admired by others and sought after in social relationships (Cuddy, Fiske, & Glick, 2007; Graziano & Eisenberg, 1997) and thus would be considered highly evaluative. Traits such as this should be especially susceptible to both positive and negative self-biases (i.e., by individual differences in self-enhancement/self-depreca tion) and thus accuracy for self-judgments, in particular, should be impaired. Traits for which there is no clear polar preference in the population should be less susceptible to such biases, leaving self-judgments largely unaffected. Thus, knowledgeable informants should be more accurate than the self for highly evaluative traits. Overall, self-reports should be more accurate than other-reports for traits low in observability (especially if they are also low in evaluativeness) and other-reports should be more accurate than self-reports for traits high in evaluativeness (especially if they are also high in observability).

How, then, should one test these predictions? The first step is to determine where traits lie on the observability and evaluativeness continua. Happily, some data exist about this for the Big Five traits (John & Robins, 1993). These data suggest that Extraversion and Neuroticism are generally lower in evaluativeness (relative to Agreeableness, Conscientiousness, and Openness), and that Extraversion is considerably more observable than the other domains. If one were to plot the major dimensions of the Big Five along the observability and evaluativeness continua, it might resemble something like Fig. 1 (the placement of each dimension is based on Figs. 4 and 5 in John & Robins, 1993), which yields predictions for accuracy across sources as seen in Table 1. Readers may question the description of Neuroticism as low in evaluativeness, but John and Robins's data show that the two poles of this dimension are not especially far apart in social desirability. This suggests that people should not be especially afraid of being judged harshly for being high (or low) on Neuroticism, and thus motivated reasoning should not be a threat to self-reports. Moreover, Neuroticism's low level of observability is a threat to the validity of informant reports.

In order to test these predictions, Vazire (2010) presented results based on a series of laboratory tasks relevant to three primary domains (extraversion, intellect, and neuroticism) and found general support for the SOKA model in that (a) self-ratings of Neuroticism (low observability, low evaluativeness) predicted neuroticism-relevant behaviors (e.g., nervous hand movements during a speech) better than did ratings of Neuroticism by knowledgeable informants or strangers, (b) informant-rated intellect (high observability, high evaluativeness) predicted intellect/creativity-related behaviors (e.g., performance on a creativity test) better than did ratings of intellect by the self or strangers, and (c) self-, informant-, and stranger-rated Extraversion (high observability, low evaluativeness) predicted extraversion-relevant behaviors (e.g., talking) equally well.

These initial findings are certainly interesting, but as is the case with any new theory, further tests of its generalizability are warranted. Specifically, there has been a recent push to take personality and social psychology back outside the laboratory (Baumeister, Vohs, & Funder, 2007; Furr, 2009; Wilson & Vazire, 2015). Personality psychology, in particular, is a science concerned with consistent, everyday thoughts, feelings, and behavior. The move into the laboratory and away from the field has been driven primarily by practicality: observing behavior as it naturally unfolds is difficult and time-consuming. However, recent innovations in technology and experience sampling have facilitated naturalistic observation. One particularly useful method for personality research has been the Electronically Activated Recorder (EAR; Mehl & Holleran, 2007; Mehl, Pennebaker, Crow, Dabbs, & Price, 2001; Mehl & Robbins, 2012), which systematically samples ambient sounds in a given individual's natural environment. This method has already helped to shed light on several topics, including narcissism (Holtzman, Vazire, & Mehl, 2010), gender (Mehl, Vazire, Ramírez-Esparza, Slatcher, & Pennebaker, 2007) and ethnic

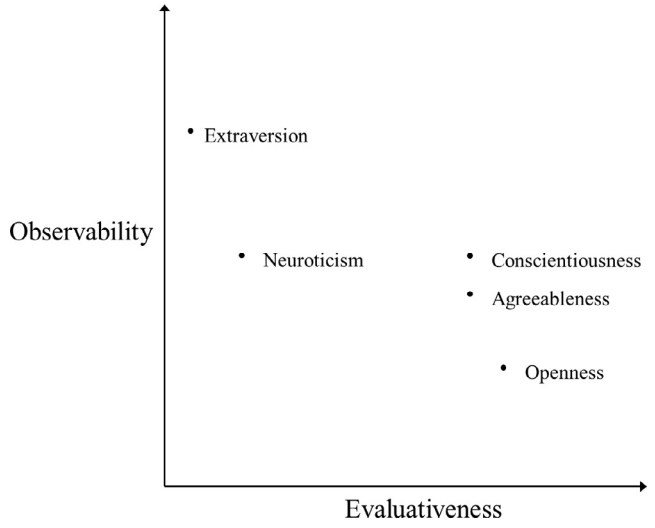

Fig. 1. Conceptual map of big five dimensions.

(Ramírez-Esparza, Mehl, Álvarez-Bermúdez, & Pennebaker, 2009) personality stereotypes, mood fluctuations (Hasler, Mehl, Bootzin, & Vazire, 2008), well-being (Mehl, Vazire, Holleran, & Clark, 2010) and depression (Mehl, 2006; Slatcher & Trentacosta, 2011; Robbins et al., 2011), and even determinants of problem behavior in children (Slatcher & Robles, 2012). Especially noteworthy are two studies which employed this methodology and have bearing on the current question: which source of personality information is most valid for a given trait? Mehl et al. (2006) investigated—among other things—the predictive validity of self-reported standings on Big Five variables, finding significant trait-behavior relations for Extraversion (e.g., talking), Agreeableness (e.g., time spent in public), Conscientiousness (e.g., time spent in class), and Neuroticism (e.g., time spent arguing). Vazire and Mehl (2008) took a slightly different approach, gathering self and informant predictions of specific classes of behavior and comparing these estimates to EAR-derived measures of these same behaviors. Specifically, they asked participants to estimate the relative frequency (compared to the average person) with which they engaged in a behavior (e.g., watching TV), obtained parallel estimates from knowledgeable informants, and then examined the predictive power of these estimates both jointly and separately. Individually, each perspective predicted relevant behavioral criteria equally well (mean *r*s across 20 behaviors for self and single informants were 0.26 and 0.23, respectively), and both perspectives provided unique predictive validity in regression analyses.

1.1. The current study

The current study aims to build on this previous work and provide another preliminary test of Vazire's (2010) SOKA model. To this end, we gathered self-, knowledgeable informant-, and stranger-ratings of Big Five variables and compared these to relevant EAR-derived behavioral codings from observation over two-day periods. To our knowledge, this is the first study which includes each of these three rater sources in conjunction with naturalistic observation, allowing for a unique test of the SOKA model. However, due to the intensive nature of collecting and coding EAR data, we were only able to collect a small amount of evidence. Thus, the results presented here should be considered preliminary – a "proof of concept" for how methods like the EAR can be used to test SOKA and other theories about the accuracy of self- and other-reports. We hope future research will use this approach as a model and provide more data in order to reach more definitive conclusions about the validity of self- and other-reports of personality using actual, naturalistic behavior as a criterion.

1.2. Predictions

Given the extant data (cf. Mehl et al., 2006; Vazire & Mehl, 2008) and the theoretical predictions derived from the SOKA model (Vazire, 2010), we would expect (1) that self-ratings of personality will predict relevant acoustically-encoded behavior in low evaluativeness domains, namely, Extraversion and Neuroticism, and to a lesser degree Agreeableness and Conscientiousness, and (2) that informant-ratings of personality will predict relevant acoustically-encoded behavior in high observability domains, namely, Extraversion, and to a lesser degree Agreeableness and Conscientiousness. The nature of the criterion variables will of course limit predictive power for low observability traits in general. Given previous findings regarding the validity of judgments of Extraversion under conditions of near zero acquaintance (Beer & Watson, 2008a; Beer & Watson, 2008b; Borkenau & Liebler, 1992; Naumann, Vazire, Rentfrow, & Gosling, 2009; Norman & Goldberg, 1966; Vazire, 2010) and Extraversion's status as a highly observable trait, we would expect that (3) strangers' ratings of Extraversion will correspond to relevant target behavior as well as do self and informant ratings. All of this said, given the dissimilarity in bandwidth (cf. Ashton, Jackson, Paunonen, Helmes, & Rothstein, 1995) between the predictors and criteria in this study (general trait measures versus specific behavioral manifestations over a relatively short observation period), we would expect that the effect sizes in this study will be unlikely to exceed $r = 0.30$ in most cases (for a similar case, see Mehl et al., 2006).

2. Method

2.1. Participants

The initial participant pool consisted of 114 college students (84 female, 66 Caucasian, 44 African American) mostly recruited from lower-level psychology courses. Individuals received partial fulfillment of a course requirement in exchange for their participation. One participant withdrew from the study, and missing or incomplete data also generate some instability in the sample size for various analyses. Additionally, sporadic equipment failure impacted

Table 1
Trait-specific hypotheses.

Variable	Observability	Evaluativeness	Prediction
Neuroticism	Low	Low	Self > Informant > Stranger
Extraversion	High	Low	Self = Informant = Stranger
Openness	Low	High	Informant > Self > Stranger
Agreeableness	Moderate	Moderate	Informant > Self > Stranger
Conscientiousness	Moderate	Moderate	Informant > Self > Stranger

Note. Categorization of traits as high or low in observability and evaluativeness is based on the results in Robins and John (1993).

approximately 20% of cases, thus the effective maximum sample size for key analyses was 87. These 87 participants did not differ significantly in terms of personality or demographics (62 female, 49 Caucasian, 35 African American) from the initial pool of 114 participants.

Sample size was determined in part by available time and resources, but we chose a round number of 100 initially for a couple of reasons. First and perhaps weakest, previous work of this type (e.g., Vazire & Mehl, 2008) involved samples of similar size. Second, the desired sample size of 250 (where correlations are quite stable) seemed practically impossible given available resources, so we opted for a goal of approximately 100, as a population effect size of 0.20 with a corridor of stability of 0.15 produces a point of stability at approximately N = 100 with 80% confidence (Schönbrodt & Perugini, 2013). Data loss of various sorts (primarily mechanical equipment failure) brought the usable dataset down to 84 for most analyses.

2.1.1. Phase 1: Initial laboratory session

Participants arrived at the laboratory individually and were asked to complete a demographic survey, some brief personality measures, and compile some lists of personal facts and preferences in open-ended form. After this, the experimenter introduced the EAR, explaining how the device worked and the parameters of the remainder of the study. Participants were asked to wear the EAR as often as possible but to keep the device out of harm's way (e.g., water, contact sports). Participants were also informed of the confidentiality policies, which included the freedom to delete any or all recorded data during or subsequent to the return of the device 48 h later (for a description of similar confidentiality policies, see Mehl & Pennebaker, 2003).

2.1.2. Phase 2: Assessment of naturalistic behavior

Participants wore the EAR for two consecutive days, always starting on a Monday. Given that most of our days are indeed weekdays, we thought important variation in behavior could be observed during this time frame. The EAR was scheduled to record for thirty-second intervals once every 12 min. In addition, the EAR was set to hibernate between the hours of 11:30 pm and 7:30 am. Thus, the maximum number of possible 30-s observations for each participant was 160. The average number of total files received per person was 137 (SD = 44). As mentioned previously, the devices occasionally experienced sudden, unexplained failure resulting in massive data loss due to the fact that these failures typically occurred relatively early in the observation process. As a result, we decided only to use cases in which at least 150 recordings were present upon return of the device. This effectively reduced the pool from 114 to 87 participants. However, in this subset of participants, the average number of total files received per participant was 158. Most importantly, the subset of participants with at least 150 available files did not differ significantly from the full sample in terms of self- or informant-reported personality variables.

To assess obtrusiveness of the methodology and general compliance, we asked participants a series of questions upon conclusion of the study, to which they responded using a 5-point scale (1 = not at all, 5 = a great deal). These self-reports indicate that participants were generally only somewhat aware of the EAR (M = 2.96, SD = 0.96) and did not feel uncomfortable wearing it (M = 1.84, SD = 1.15). They also reported that they were not impeded (M = 1.45, SD = 0.82) by the EAR, nor did it alter daily behavior (M = 1.55, SD = 0.75). Though participants did report others being aware of (M = 3.58, SD = 1.14) and talking about (M = 3.69, SD = 1.01) the EAR, they did not feel that others altered their behavior due to the presence of the EAR (M = 2.20, SD = 1.20). On average, participants estimated that they spent 85% (SD = 12%) of their waking hours wearing the EAR, and on a scale of 1 (not at all typical) to 5 (very typical) they rated the two-day observation as fairly typical (M = 4.01, SD = 1). Though these estimates are for the full sample, the parallel estimates from participants with 150 or more available files were basically identical.

A team of 5 coders evaluated the resulting files using a revised and expanded version of the Social and Environmental Coding of Sound Inventory (SECSI; Mehl et al., 2006). The system assesses four major categories: interactions (e.g., talking with a group, talking with one other, on the phone, etc.), activities (e.g., eating, sleeping, watching television), moods (e.g., laughing, singing, crying), and locations (e.g., in apartment, in transit, in public). In addition to these categories, coders evaluated each segment in terms of the six basic emotions (happiness, sadness, anger, fear, surprise, disgust) and, in the event of conversation, the type of conversation (e.g., deep, gossip, practical). Finally, we constructed an auditory behavioral measurement model for the Big Five, covering Extraversion (behavioral classes included dominance, affiliation, and energy), Neuroticism (worrying, complaining, sadness), Openness to Experience (creativity, unconventionality, spirituality), Agreeableness (sympathy, concern for others, affection), and Conscientiousness (cleaning, hygiene, responsibility). Due to generally low within-domain correspondence (perhaps due to low base rate in some cases), these behavioral classes were considered separately for analysis.

Coders were asked to listen to each file in a given case and evaluate it with respect to the 58 distinct categories (for a complete list and description of all categories, see the Appendix). For example, if the participant was talking angrily about a friend to another friend over lunch, the coder would be expected to mark a "1" in the spreadsheet under "talking with one other", "talking with friend", "gossip", "anger", "in public", and "eating", perhaps among other things, depending on the circumstances. To establish reliability, the five coders all evaluated four semi-randomly selected (to reflect the breadth of the target demographics) complete cases, which consisted of 632 total files. ICC (2, 1) exceeded 0.30 for most categories with a reasonable base rate of occurrence (see Table 2 for specific information). Particularly unreliable or incredibly low frequency behaviors were generally omitted from further analysis.

These raw codings were then converted to time-use estimates by dividing the number of coded instances in a given category by the number of total usable files.[1] The latter was defined by files in which some ambient sound of any kind could be detected. In the case of complete silence (which should be rare, given that the EAR successfully captured things as quiet as the turning of pages), coders were instructed to flag the file, and long periods of silence were compared to participant diaries[2] to confirm non-compliance (e.g., left EAR in the car). These flagged files were subtracted from the total number of files to produce the denominator in the time-use estimates. For the 87 participants who met the base file requirement, an average of 128 files per person were deemed usable (SD = 18). Conversely, the omitted cases only averaged 48 usable files per participant (SD = 29), further justifying the omission of these cases from further analysis. These proportions would serve as the central criterion for all subsequent analyses.

2.1.3. Phase 3: Second laboratory session

Exactly 48 h after having received the EAR, participants returned to the laboratory, at which point they completed the

[1] One exception to this practice was sleep. Because sleep instances were excluded from the total usable file count, the denominator for the time-use percentage in this case was the total number of files (as opposed to usable files).

[2] Paricipants were also asked to keep a very brief standardized diary (spreadsheets with one-hour blocks as rows), in which they accounted for their general activities during hour-long blocks (e.g., at work, in class). An additional column of the diary allowed participants to indicate whether they were wearing the EAR during a given time period.

Table 2
EAR descriptive statistics.

Behavior	Inter-coder reliability	At least once (N)	Base rate time usage %			
			Minimum	Maximum	Mean	SD
Talking						
Overall	.89 (.97)	87	6.2	59.1	30.0	12.1
With 1 person	.63 (.90)	87	0.8	45.2	16.7	9.7
With group	.61 (.89)	81	0.0	21.8	6.7	5.6
On phone	.89 (.98)	78	0.0	32.2	4.3	4.6
With romantic partner	.06 (.24)	31	0.0	36.5	3.6	7.6
With child	.66 (.91)	17	0.0	26.5	1.3	4.1
With other family	.02 (.10)	39	0.0	22.7	2.3	4.5
With friend	.71 (.92)	86	0.0	42.6	17.0	11.0
With stranger	.64 (.90)	37	0.0	29.4	1.2	3.7
To pet	–	12	0.0	4.0	0.2	0.7
Activities						
Socializing	.72 (.93)	63	0.0	54.6	12.2	14.4
In class	.87 (.97)	86	0.0	31.9	15.5	5.9
On computer	.50 (.84)	54	0.0	24.5	2.6	4.6
Working	.82 (.96)	29	0.0	49.7	6.3	11.1
Eating	.12 (.42)	53	0.0	8.6	1.5	1.8
TV on	.79 (.95)	87	0.7	72.7	28.6	15.5
Sleeping	.39 (.77)	69	0.0	33.3	8.4	8.1
Studying	.19 (.54)	53	0.0	35.9	4.5	6.9
Music on	.86 (.97)	86	0.0	41.8	14.8	9.2
Arguing	.33 (.71)	17	0.0	5.6	0.3	0.1
Swearing	.57 (.80)	26	0.0	5.7	1.1	1.5
Substance use	.25 (.62)	14	0.0	7.9	0.3	1.2
Laughing	.57 (.87)	81	0.0	19.9	5.2	3.8
Singing	.73 (.93)	62	0.0	10.7	1.7	2.0
Crying	–	4	0.0	1.6	0.0	0.2
Sighing	.14 (.44)	49	0.0	3.6	0.7	0.9
Coughing	.45 (.80)	45	0.0	9.4	0.9	1.4
Location						
Apartment	.66 (.91)	87	0.7	77.1	46.7	15.8
Outdoors	.46 (.81)	62	0.0	26.5	3.2	4.5
In public	.73 (.93)	87	13.2	67.2	38.7	13.8
In transit	.77 (.94)	87	3.5	43.5	14.3	7.1
Type of conversation						
Small talk	.26 (.63)	83	0.0	25.5	6.4	6.2
Deep conversation	–	40	0.0	19.6	1.0	2.4
Personal	.44 (.80)	60	0.0	13.5	2.1	2.8
Gossip	.24 (.62)	72	0.0	13.8	3.1	3.0
Practical	.62 (.89)	86	0.0	43.5	15.6	9.7
Emotions						
Happiness	.22 (.59)	49	0.0	13.5	2.6	3.7
Anger	.22 (.59)	31	0.0	11.6	0.7	1.6
Fear	–	1	0.0	1.4	0.0	0.1
Surprise	.00	17	0.0	2.3	0.2	0.4
Disgust	–	4	0.0	1.5	0.0	0.2
Big five behaviors						
Dominance	.23 (.60)	30	0.0	6.1	0.6	1.1
Affiliation	–	3	0.0	6.6	0.1	0.7
Energy	.04 (.16)	22	0.0	14.0	0.6	1.8
Sympathy	.83 (.96)	12	0.0	1.6	0.1	0.3
Concern for others	.35 (.73)	30	0.0	3.1	0.4	0.7
Affection	–	17	0.0	6.5	0.3	1.0
Complaining	.14 (.45)	49	0.0	5.2	1.0	1.0
Worrying	.10 (.36)	14	0.0	2.7	0.2	0.5
Sadness	.58 (.88)	8	0.0	1.0	0.1	0.2
Cleaning	–	39	0.0	20.6	1.3	2.9
Hygiene	.57 (.87)	55	0.0	5.0	1.0	1.2
Responsibility	–	26	0.0	4.8	0.4	0.9
Spirituality	–	10	0.0	2.9	0.1	0.4
Creativity	–	5	0.0	3.6	0.1	0.6
Unconventionality	–	5	0.0	1.4	0.1	0.2

Note. N = 83. Reliability = ICC (2, 1); ICC (2,k) in parentheses. If at least three raters had zero occurrences logged in the training files, we did not calculate reliability (–). At Least Once = number of participants for whom this behavior was coded at least one time during the observation period.

EAR compliance questionnaire. They then took part in a brief videotaped interview and were asked to provide the email addresses of two people who knew them well. We then contacted these individuals and asked them to complete a personality inventory (administered online) about the target individual (see Vazire, 2006 for a description of the general procedure). Of the possible 228 informants, we received 138 responses, with two informant reports for 50 participants, one for 88, and zero for 26.

Participants were also given the opportunity to review their EAR files at this point. Most declined the opportunity, though approximately 10% requested a copy of their files for personal use. Only two participants asked that we delete specific files (a total of three files), and, as mentioned previously, one participant asked that all files be deleted.

2.1.4. Phase 4: Stranger ratings

Several months after the conclusion of the laboratory sessions and as part of a subsequent study, we obtained personality ratings about the target individuals from a sample of individuals who were entirely unacquainted with these targets. In this study, judges made three separate personality evaluations of each target after having received three different kinds of information about the target: a three-item list of individuating facts written by the target participant about him/herself, a three-item list of personal values also written by the target participant (see Beer & Brooks, 2011, for a description of these types of information), and a muted video segment (approximately one-minute in length on average) from the Phase 1 laboratory session. The videos were muted for reasons related to the original goals of this second study. These three estimates were averaged within-rater, and each target was judged by three raters. The within-rater (across trial) coefficient alphas ranged from 0.71 (Extraversion) to 0.82 (Openness), and the within-trial (across rater) coefficient alphas ranged from 0.15 (Neuroticism) to 0.52 (Extraversion). Given the low number of items (3 in each case), these estimates were not so low as to preclude overall aggregation. Thus, the stranger-generated personality estimates for subsequent analyses represents an aggregate of nine ratings: three from each of three judges per target.

3. Measures

We assessed personality using the Big Five Inventory (BFI; John, Naumann, & Soto, 2008). The 44-item version of the BFI contains 8-item scales assessing Neuroticism and Extraversion, a 10-item Openness scale, and 9-item measures of Agreeableness and Conscientiousness. The participants rated themselves using a five-point scale (1 = disagree strongly, 5 = agree strongly) in response to a series of adjectives and phrases following a stem statement ("I see myself as someone who…"). This instrument was used for all ratings, and the informant- and stranger-rating forms featured a modified stem ("I see this person as someone who…"). Coefficient alphas for the scales ranged from 0.75 to 0.91, from 0.83 to 0.90, and from 0.87 to 0.94 for self-, informant-, and stranger-ratings, respectively. Behavior was acoustically captured via the EAR, which in this case was a Dell Axim v50, housed in a protective case.

4. Results

4.1. Descriptive statistics: behavior

As described above, coders evaluated each audio file for the 87 eligible participants along 56 dimensions. Table 2 provides the frequency estimates and inter-coder reliabilities for these behavioral variables. The reliability estimates are based on the five coders' ratings of four participants' complete files, whereas all other target participants' files were only coded by one coder. Vazire and Mehl (2008) reported ICC (2,k) estimates, as the likely intended use for EAR coding was in the form of aggregated ratings. However, the present application is to single-rater EAR estimates, thus ICC (2,1) estimates are more appropriate. For comparative purposes, Table 2 contains the ICC (2,1) estimates first, followed by the ICC (2,k) estimates. As the second column of the table indicates, the intraclass (ICC [2, k]) correlations that captured inter-coder reliability exceeded 0.70 in most (30) cases. There were 12 behaviors that did not occur in any of the 4 cases used for establishing reliability estimates and 16 behaviors for which the intraclass correlations were lower than 0.70. Several of the latter instances occurred for variables that were generally low base rate behaviors in this population (e.g., energy, worrying, substance use). Obviously, the ICC (2,1) values are considerably lower, potentially limiting the possibility of detecting relations with external criteria. However, comparable reliability estimates did not preclude the observation of substantial correlations with external measures in previous research (Mehl et al., 2006; Vazire & Mehl, 2008). Due to the sampling method for obtaining reliability estimates, we thought it prudent to retain all variables for subsequent analyses, with the understanding that the low reliability for some of these variables would certainly impact the likelihood of observing convergent relations with the trait predictors.

4.2. Descriptive statistics: personality measures

Table 3 provides the means and standard deviations for judgments of the Big Five by source. Since there were no significant differences for any source between trait estimates for participants with eligible EAR files (greater than 150 files obtained) and those falling short of that requirement, we have provided the full-sample estimates. These descriptive statistics may give some clues to the validity of each source. For example, some recent work has indicated that informants recruited using methods similar to those employed in this study may suffer from a "pal-serving bias" (Leising et al., 2010) which may appear in the form of means closer to the extreme ends of the scale and smaller standard deviations (restricted variance) on evaluative traits (cf. Wood & Wortman, 2012). Comparisons of standard deviations across methods was complicated by differential aggregation, so we will focus primarily on mean comparisons. For Agreeableness, Neuroticism, and Conscientiousness, we observed no substantial mean differences (*d*s range from −0.19 to 0.08) among the three judgment sources. Informant-ratings for Openness to Experience were higher than strangers' ratings of targets ($d = 0.85$) and the targets' self ratings ($d = 0.49$). A similar pattern emerged for Extraversion, with informants rating targets higher than did strangers ($d = 0.80$) and slightly higher than targets rated themselves ($d = 0.31$). However, Extraversion is the least evaluative of the Big Five traits so these mean differences do not necessarily reflect a strong positivity bias among informants.

4.3. Bivariate relations: preliminary analysis

Another method for examining potential issues with scale validity involves examining the inter-trait associations. Table 4 provides the multitrait-multimethod matrix (MTMM) for all trait judgments in this study. For all correlations, we first calculated scale scores, then aggregated across raters as necessary (i.e., for informant and stranger ratings). First, there is evidence of an acquaintanceship effect (Funder & Colvin, 1988; Watson, Hubbard, & Wiese, 2000) as evidenced by the convergent relations across the three sources. Self-informant agreement was quite strong for Neuroticism, Extraversion, and Openness to Experience, and somewhat weaker for Agreeableness and Conscientiousness. The self-stranger convergent correlations exceeded 0.20 for all traits except Conscientiousness ($r = 0.07$), though all effects were small (*r*s < 0.30). Finally, the informant-stranger relations tended to be weak, with only Extraversion ($r = 0.34$) showing agreement greater than 0.20. The discriminant relations, however, are probably of more interest in the current study. The heterotrait-heteromethod relations tended to be quite small in most cases, with only 5 of the 60 correlations exceeding |0.20| and all being

Table 3
Descriptive statistics for target (self-reported), informant, and stranger ratings.

	Self		$d_{self\text{-}informant}$	Informant		$d_{informant\text{-}stranger}$	Stranger		$d_{stranger\text{-}self}$
	Mean	SD		Mean	SD		Mean	SD	
Neuroticism	21.32	6.90	*−.10*	22.04	7.04	*.17*	21.11	2.98	*−.04*
Extraversion	27.99	7.30	*−.36*	30.37	5.99	*.80*	26.10	4.57	*−.31*
Openness	36.31	6.36	*−.27*	37.94	5.83	*.85*	33.75	3.81	*−.49*
Agreeableness	36.27	5.11	*.08*	35.82	5.66	*−.02*	35.92	3.40	*−.08*
Conscientiousness	34.56	5.08	*−.19*	35.53	5.65	*.23*	34.45	3.32	*−.03*

Note. N = 110 (Self), 86 (Informant), 109 (Stranger). Italicized numbers are Cohen's d, with the direction of the difference indicated by the order of the source (e.g., $d_{self\text{-}informant}$ denotes the effect size when subtracting the informant estimate from the self estimate for the given trait).

Table 4
Multitrait-multimethod matrix of personality judgments.

	Self					Informant					Stranger				
	N	E	O	A	C	N	E	O	A	C	N	E	O	A	C
Self															
N	(.87)														
E	**−.30**	(.91)													
O	−.17	**.25**	(.81)												
A	**−.49**	**.21**	.00	(.77)											
C	**−.27**	.14	−.01	**.27**	(.75)										
Informant															
N	**.50**	−.04	−.05	**−.33**	−.16	(.90)									
E	−.17	**.67**	.03	.15	.09	−.02	(.86)								
O	−.05	**.20**	**.51**	−.08	−.09	**−.23**	**.23**	(.83)							
A	**−.30**	−.03	**−.23**	**.34**	−.02	**−.59**	.11	.16	(.83)						
C	−.11	−.15	**−.23**	.03	**.32**	**−.48**	−.02	**.27**	**.41**	(.84)					
Stranger															
N	**.22**	−.11	−.13	−.17	−.03	.19	−.11	−.09	−.07	.06	(.87)				
E	−.17	**.29**	.11	**.24**	.10	−.06	**.34**	.17	.00	−.02	**−.55**	(.94)			
O	.08	−.02	**.29**	−.02	−.09	.08	−.03	.15	−.15	−.20	**−.31**	.19	(.87)		
A	−.06	−.02	.01	**.21**	−.15	.04	.19	.00	0.13	−.13	**−.49**	**.31**	**.30**	(.91)	
C	−.06	−.13	.05	.06	.07	−.01	.00	−.05	.02	.02	**−.30**	.11	**.24**	**.36**	(.90)

Note. *Ns* vary depending upon sources. For Self, Informant, and Stranger intercorrelations, *Ns* are 110, 86, and 109, respectively. Self-Informant N was 85; Self-Stranger N was 107; Informant-Stranger N was 88. Correlations greater than .20 are in bold; convergent relations are underlined. Coefficient Alpha appears in parentheses. N = Neuroticism, E = Extraversion, O = Openness to Experience, A = Agreeableness, C = Conscientiousness.

fairly small in magnitude (rs < |0.34|). Four of these correlations could be found off the diagonal in the self-informant heteromethod block. Still, the convergent-discriminant pattern for self and informant judgments was largely solid in that convergent relations were generally stronger than the discriminant relations. The stranger judgments, however, showed little convergence (monotrait-heteromethod relations) and all but one heterotrait-monomethod relation (Openness-Conscientiousness) exceeded |0.20|, with a few being quite strong (e.g., Neuroticism-Extraversion, r = −0.55). Though shared method variance was evident in all three sources, there was a general trend toward more and stronger intercorrelations among trait judgments with decreasing acquaintanceship, parallel with previous findings (Beer & Watson, 2008b). This stronger halo effect may serve to depress each scale's relations with external criteria for informant- and especially stranger-ratings. Thus, one might expect that informant- and stranger-ratings may fare worse in terms of behavior prediction, in general, than self-ratings of personality, in part due to the lower differentiation among traits for those raters.

4.4. Bivariate relations: criterion validity

The central focus of this study is behavior prediction from personality judgments. What follows is a trait-by-trait analysis of the criterion validity of trait judgments by each source. In addition to the aggregates described previously, we calculated single-informant and single-stranger estimates for each personality variable. The single-informant estimates were obtained by simply excluding the second informant from each aggregated informant estimate. We did this because the first informant would most likely reflect the response we would have gotten had we only asked for one informant. Furthermore, given that target-chosen informant ratings tend to be uniformly high in liking and positivity (Leising et al., 2010), we did not suspect that randomly choosing an informant would significantly alter the results. The single-stranger ratings still include some aggregation, as each estimate represents the average of three successive judgments made by one perceiver. However, the single-stranger estimates result from calculating the trait-behavior correlation for each of three perceivers across targets, Fisher-transforming these correlations, averaging them, and then back-transforming these estimates. We will primarily discuss the aggregated estimates, but the single-rater estimates are available for comparison purposes. These help address the question of whether any differences across the three sources (self, informant, stranger) could be due to the fact that there was greater aggregation for some sources than others (stranger > informants > self), which would have led to greater reliabilities and therefore more opportunity for predictive validity.

Given the mismatch in bandwidth of the predictors (generalized trait measures as assessed by questionnaire) and criteria (EAR-coded behavior over a two-day period as assessed by frequency counts) in this study, we expected small effects. However, as described above, the intensive nature of data collection and coding for this study precluded us from achieving the sample size we

would need to have adequate statistical power to reject the null hypothesis when effects are small. Thus, we decided not to focus on p-values and instead chose to examine these relations in terms of effect sizes and confidence intervals only, interpreting effects of the typical size (*r*s in the 0.20s) observed in previous studies connecting trait ratings and EAR-coded behaviors (cf. Mehl et al., 2006) as evidence for an association. Specifically, we will be noting correlations at or exceeding |0.20| in magnitude throughout the following sections. We want to emphasize that our design does not permit precise effect size estimation, and should be followed up with more data to refine the effect size estimates. We know of at least one larger EAR study (N = 380) that has been in progress for four years and is expected to yield data that can attempt to replicate these analyses within the next few years (Vazire et al., in progress). Thus, the purpose of the results presented here is to provide preliminary evidence regarding the validity of self- and other-reports of personality against a naturalistic, behavioral criterion, and to provide a framework for linking Big Five traits to EAR-coded behaviors in future studies.

For each section, we first identified (a priori) a subset of trait-relevant behaviors and then examined the correlations between trait judgments from each source and these criterion behaviors. We begin with Extraversion, the trait with the most theoretically relevant acoustically coded behavioral manifestations, and progress to Openness to Experience, the trait with the fewest theoretically relevant acoustically coded behavioral manifestations.

4.4.1. Extraversion

Given the conceptual definition of extraversion, we expected that trait judgments should predict behaviors related to sociability (e.g. various talking behavior, socializing, affiliation, being in public places), positive emotionality (e.g., laughing, singing, happiness), and general activity or arousal seeking (e.g., energy, listening to music, being in transit). Furthermore, in accordance with the SOKA model, we expected that self-, informant-, and stranger-ratings would show relatively equivalent criterion validity in this domain. Table 5 provides the relevant relations between self-, informant-, and stranger-rated Extraversion and theoretically relevant behavioral manifestations of extraversion.

In terms of sociability, all three sources predicted talking with friends, while only informant judgments predicted talking with one other person. Interestingly, all three correlations with "talking with a romantic partner" were negative ($r_{self} = -0.19$; $r_{informant} = -0.15$; $r_{stranger} = -0.23$). Although it is not surprising that this relation is not positive, the fact that spending more of one's time spent talking with a romantic partner is associated with introversion was unanticipated. In line with expectations, both self- and informant-ratings of Extraversion predicted engaging in gossip ($r_{self} = 0.20$; $r_{informant} = 0.21$) and practical conversations ($r_{self} = 0.21$; $r_{informant} = 0.23$), and stranger-ratings of Extraversion predicted affiliative behavior ($r = 0.24$). However, none of the three sources predicted socializing at or above $r = 0.20$.

In terms of general activity or arousal seeking, informant- and stranger-rated Extraversion predicted time spent in transit ($r_{informant} = 0.30$; $r_{stranger} = 0.21$), stranger-ratings predicted time spent listening to music ($r = 0.24$), and all three sources predicted energetic behavior ($r_{self} = 0.21$; $r_{informant} = 0.23$; $r_{stranger} = 0.22$).

Finally, in terms of positive emotionality, self- and stranger-reports both predicted laughing ($r_{self} = 0.22$; $r_{stranger} = 0.23$), self-reports predicted singing ($r = 0.21$) and stranger-rated Extraversion predicted coded happiness ($r = 0.39$).

In summary, each perspective on Extraversion showed some degree of criterion validity, with aggregated stranger-ratings showing the strongest relations overall. This is not entirely surprising, given previous findings with respect to the validity of stranger-rated extraversion in low acquaintance situations (Beer & Watson, 2008a, 2010; Naumann et al., 2009; Norman & Goldberg, 1966; Vazire, 2010; Watson, 1989) and predictions derived from the SOKA model. The benefit of aggregation is somewhat pronounced, however. When using single informants and single strangers (results listed in Table 5), the average absolute magnitude of the predictive validity of self-ratings (mean absolute $r = 0.14$), informant-ratings (mean absolute $r = 0.16$) and stranger ratings (mean absolute $r = 0.14$) show that these three sources predicted Extraversion-related behavior at similar levels.

4.4.2. Neuroticism

Given the conceptual definition of neuroticism, we expected that trait measures of the construct would predict behaviors related to negative mood states[3] (e.g., sadness, anger, crying, complaining), social withdrawal (e.g., time at home or indoors), and markers of mental or physical illness (e.g., sleeping, coughing, less working). Given that Neuroticism could be considered relatively low in observability and in evaluativeness, we expected self-ratings to more strongly predict relevant behaviors than informant- or stranger-ratings and informant-ratings to outperform stranger-ratings.

Table 6 provides the correlations between self-, informant-, and stranger-ratings of Neuroticism and neuroticism-relevant behavior, as captured by the EAR. Contrary to predictions, informant-ratings seemed to best predict behavior, with at least one of the aggregated- or single-informant estimates of Neuroticism predicting 6 of the 11 relevant behaviors at r = |0.20| and yielding estimates approaching this value for 2 of the remaining 5 (the three exceptions—sighing, complaining, worrying—were low reliability and/or low base rate behaviors in this sample). Conversely, self- and stranger-rated Neuroticism failed to predict any relevant behavior. More specifically, effect sizes were strongest for informants (mean absolute $r = 0.18$), followed by self-ratings (mean absolute $r = 0.09$) and stranger-ratings (mean absolute $r = 0.06$).

4.4.3. Conscientiousness

Given the conceptual definition of conscientiousness, we expected trait ratings to predict goal-oriented behavior (e.g., attending class, studying, time on computer and less television viewing), organization and preparation (e.g., cleaning and hygienic behavior), norm adherence (e.g., low substance use, low swearing, refraining from gossip), and responsibility. Given that conscientiousness could be described as moderately observable and evaluative, the SOKA model does not make a clear prediction regarding which perspective should be most valid. However, since it is not *highly* evaluative and conscientiousness is a trait that typically generates stronger convergence even in lower acquaintance situations (e.g., Borkenau & Liebler, 1992), we expected that all three sources would have some validity, with self- and informant-ratings more strongly predicting relevant behavior due to greater access to information. We would also perhaps expect a slight advantage for informant-judgments (relative to self-judgments) due to the impact of its moderate evaluativeness.

Table 7 provides the correlations between self-, informant-, and stranger-ratings of Conscientiousness and Conscientiousness-relevant behavior, as captured by the EAR. These data are less interpretable than those for Extraversion and Neuroticism. Stranger-rated Conscientiousness shows some strong evidence of criterion validity, predicting time spent in class ($r = 0.31$), sleeping ($r = -0.27$), and swearing ($r = -0.41$). Indeed, one would expect highly conscientious individuals to be in class, not swear, and be awake more often between the hours of 7:30 am and 11:30 pm (this effect is likely driven by differences in the early part of the day).

[3] Fear and disgust occurred so infrequently in this sample (present in 1 and 4 files, respectively) that we excluded them from analysis.

Table 5
Correlations between Ratings of Extraversion and Relevant Acoustically-Captured Behavior.

Predictor	Self	Informant		Stranger	
		Aggregate	Single	Aggregate	Single
Talking					
1 Other	.06 [−.16, .27]	**.29** [.06, .49]	.27 [.04, .48]	−.10 [−.31, .12]	−.08 [−.29, .04]
Group	.01 [−.21, .22]	−.05 [−.28, .19]	−.06 [−.28, .19]	**.24** [.03, .43]	.18 [−.04, .38]
Friend	**.24** [.02, .43]	**.30** [.07, .50]	**.28** [.05, .49]	**.27** [.06, .46]	**.20** [−.02, .40]
Partner	−.19 [−.39, .03]	−.15 [−.37, .09]	−.12 [−.35, .11]	**−.23** [−.42, −.01]	−.17 [−.37, .05]
Location					
In Transit	.14 [−.08, .34]	**.30** [.07, .50]	**.28** [.04, .48]	**.21** [−.01, .40]	.15 [−.07, .35]
In Public	−.03 [−.24, .19]	−.06 [−.29, .18]	−.09 [−.32, .15]	.14 [−.07, .35]	.11 [−.11, .32]
Activities					
Socializing	−.04 [−.26, .18]	.04 [−.20, .27]	.04 [−.20, .27]	.17 [−.05, .37]	.12 [−.10, .33]
Music	.09 [−.13, .30]	.18 [−.06, .40]	.12 [−.12, .35]	**.28** [.07, .46]	**.20** [−.02, .40]
Laughing	**.22** [.00, .41]	.16 [−.08, 38]	.14 [−.10, .37]	**.23** [.01, .42]	.17 [−.05, .37]
Singing	**.21** [−.01, .41]	.13 [−.11, .36]	.08 [−.16, .32]	.11 [−.11, .31]	.08 [−.29, .04]
Type of conversation					
Gossip	**.20** [−.01, .40]	**.21** [−.02, .43]	.19 [−.04, .42]	.05 [−.16, .26]	.04 [−.18, .25]
Practical	**.21** [.00, .41]	**.23** [.00, .45]	**.22** [−.03, .43]	−.02 [−.24, .19]	−.02 [−.23, .20]
Emotion					
Happiness	.17 [−.05, .37]	.08 [−.16, .31]	.08 [−.16, .31]	**.39** [.19, .56]	**.28** [.07, .47]
Big five aspects					
Affiliation	.15 [−.06, .36]	.19 [−.04, .41]	**.21** [−.04, .42]	**.24** [.03, .43]	.17 [−.05, .37]
Energy	**.21** [−.01, .40]	**.23** [−.01, .44]	**.20** [−.05, .41]	**.22** [.01, .41]	.16 [−.06, .36]

Note. N = 83 (Self), 69 (Informant), 83 (Stranger). Values in brackets represent lower and upper bounds of 95% confidence intervals. Correlations stronger than |.20| appear in bold.

Table 6
Correlations between ratings of neuroticism and relevant acoustically-captured behavior.

Predictor	Self	Informant		Stranger	
		Aggregate	Single	Aggregate	Single
Location					
Apartment	.06 [−.15, .28]	**.23** [−.01, .44]	.19 [−.05, .41]	.07 [−.14, .28]	.05 [−.17, .26]
In Public	−.03 [−.25, .19]	**−.22** [−.43, .02]	**−.22** [−.45, .00]	−.07 [−.28, .15]	−.05 [−.26, .17]
Activities					
Working	−.14 [−.34, .08]	**−.27** [−.48, −.04]	**−.21** [−.45, .00]	.00 [−.22, .21]	.00 [−.22, .22]
Sleeping	.13 [−.09, .33]	**.22** [−.02, .43]	**.22** [−.04, .41]	−.02 [−.24, .19]	−.01 [−.23, .21]
Crying	.14 [−.08, .35]	.19 [−.05, .41]	**.20** [−.04, .41]	−.05 [−.26, .16]	−.05 [−.26, .17]
Coughing	.15 [−.07, .35]	**.20** [−.04, .42]	.15 [−.09, .37]	.05 [−.15, .28]	.05 [−.17, .26]
Sighing	.09 [−.13, .30]	.11 [−.13, .33]	.19 [−.06, .40]	.02 [−.19, .23]	.02 [−.20, .23]
Emotion					
Sadness	.06 [−.15, .28]	.18 [−.06, .40]	**.22** [−.02, .43]	−.13 [−.33, .09]	−.09 [−.30, .13]
Anger	−.06 [−.27, .15]	.19 [−.05, .41]	.13 [−.11, .36]	.17 [−.05, .37]	.12 [−.10, .33]
Big five aspects					
Worrying	.10 [−.23, .20]	.15 [−.09, .37]	.17 [−.08, .38]	−.09 [−.30, .13]	−.07 [−.28, .15]
Complaining	−.02 [−.12, .31]	.18 [−.06, .40]	.12 [−.13, .33]	.00 [−.22, .21]	−.01 [−.23, .21]

Note. N = 83 (Self), 69 (Informant), 84 (Stranger). Values in brackets represent lower and upper bounds of 95% confidence intervals. Correlations stronger than |.20| appear in bold.

Self-ratings of Conscientiousness predicted less time spent on the computer (r = −0.26), less time engaged in gossip (r = −0.20), and swearing less often (r = −0.25). At first glance, the inverse relation with time spent on the computer would seem puzzling, but in our sample, coders reported that computers were typically used as a distraction (e.g., video games, viral videos) rather than for the ostensible completion of work-related tasks.

Informants generally fared relatively worse (compared with self- and stranger-ratings) in terms of criterion validity in this category, predicting only sleeping (r = −0.23) in the hypothesized direction (single informants predicted swearing in a manner contrary to hypothesis, r = 0.23). This is not entirely surprising, given the relatively weak convergent/discriminant pattern displayed by informant-rated Conscientiousness (see Table 4) in this study. Thus, the hypotheses for this trait were largely unsupported.

4.4.4. Agreeableness

Given the conceptual definition of Agreeableness, we expected trait ratings in this domain to predict behaviors that indicated investment in relationships and concern for others (e.g., talking to romantic partners or family, engaging in personal conversations, sympathy, affection, concern for others) and the avoidance of confrontation (e.g., less anger). Given that Agreeableness is moderately evaluative and not easily visible at low acquaintance (Ames & Bianchi, 2008), we expected that informant-ratings would better predict relevant behaviors than would self-ratings, and that stranger-ratings would show very little criterion validity.

Table 8 provides the correlations between self-, informant-, and stranger-ratings of Agreeableness and Agreeableness-relevant behavior, as captured by the EAR. The most obvious feature of this table is the lack of substantial relations between ratings of

Table 7
Correlations between ratings of conscientiousness and relevant acoustically-captured behavior.

Predictor	Self	Informant		Stranger	
		Aggregate	Single	Aggregate	Single
Activities					
In Class	.10 [−.12, .30]	−.02 [−.26, .22]	−.07 [−.30, .17]	**.31** [.10, .49]	**.21** [−.01, .41]
Computer	−**.26** [−.45, −.04]	−.11 [−.34, .13]	−.08 [−.31, .16]	−.09 [−.30, .12]	−.07 [−.28, .15]
Sleeping	−.02 [−.23, .20]	−**.23** [−.45, .01]	−**.22** [−.43, .02]	−**.27** [−.46, −.06]	−.18 [−.38, .04]
Studying	**.21** [.00, .41]	.09 [−.15, .32]	.10 [−.14, .33]	.04 [−.17, .25]	.03 [−.19, .24]
Swearing	−**.25** [−.49, .03]	.13 [−.18, .41]	**.23** [−.08, .50]	−**.41** [−.62, −.15]	−**.31** [−.49, −.10]
Substance Use	.02 [−.20, .23]	.07 [−.17, .31]	.07 [−.17, .31]	.01 [−.21, .22]	.01 [−.21, .23]
TV on	−.09 [−.30, .12]	−.15 [−.38, .09]	−.10 [−.33, .14]	−.12 [−.32, .10]	−.08 [−.29, .14]
Working	−.01 [−.22, .21]	.05 [−.19, .29]	.10 [−.14, .33]	.07 [−.15, .28]	.04 [−.18, .25]
Type of conversation					
Gossip	−**.20** [−.40, .02]	−.08 [−.31, .16]	−.06 [−.29, .18]	−.07 [−.29, .15]	−.04 [−.25, .18]
Big five aspects					
Cleaning	−.06 [−.27, .16]	.01 [−.23, .25]	.04 [−.20, .28]	−.16 [−.36, .06]	−.11 [−.32, .11]
Hygiene	.12 [−.10, .33]	.14 [−.10, .37]	.09 [−.15, .32]	.03 [−.19, .24]	.01 [−.21, .23]
Responsibility	.15 [−.07, .35]	.02 [−.22, .26]	.02 [−.22, .25]	.03 [−.19, .24]	.02 [−.20, .23]

Note. $N = 83$ (Self), 68 (Informant), 84 (Stranger). *N*s for Swearing are only 50 (for self- and stranger-ratings) and 42 (for informant-ratings) because two coders failed to register the category. Values in brackets represent lower and upper bounds of 95% confidence intervals. Correlations stronger than |.20| appear in bold.

Table 8
Correlations between ratings of agreeableness and relevant acoustically-captured behavior.

Predictor	Self	Informant		Stranger	
		Aggregate	Single	Aggregate	Single
Talking					
Partner	−.09 [−.30, .12]	−.14 [−.37, .10]	.02 [−.22, .26]	−.06 [−.27, .16]	−.03 [−.24, .19]
Family	.07 [−.15, .28]	.17 [−.07, .39]	.15 [−.08, .38]	.16 [−.05, .37]	.12 [−.10, .33]
Emotion					
Anger	.11 [−.11, .32]	.02 [−.21, .26]	.04 [−.20, .27]	−.06 [−.27, .15]	−.04 [−.25, .18]
Big five aspects					
Sympathy	.03 [−.19, .24]	−.03 [−.27, .20]	−.05 [−.28, .19]	.02 [−.19, .23]	.02 [−.20, .23]
Concern	.17 [−.04, .37]	.11 [−.13, .34]	.16 [−.07, .39]	.17 [−.04, .37]	.13 [−.09, .34]
Affection	−.07 [−.28, .14]	.02 [−.22, .26]	.05 [−.19, .29]	−.03 [−.25, .18]	−.01 [−.23, .21]

Note. $N = 84$ (Self), 69 (Informant), 84 (Stranger). Values in brackets represent lower and upper bounds of 95% confidence intervals. Correlations stronger than |.20| appear in bold.

Agreeableness and relevant behavior. The lack of predicted correlations suggest the possibility that the EAR (or the coding system employed) did not adequately capture Agreeableness-related behavior, so with respect to the hypotheses tested, we would consider the results inconclusive, and encourage future EAR research to explore other possible acoustically-detectable behavioral manifestations of Agreeableness.

4.4.5. Openness to Experience

Given the conceptual definition of Openness to Experience, we expected trait ratings to predict creativity, unconventionality, engagement in philosophical discussion (deep conversations) and exploratory behavior (substance use). Given the high evaluativeness of Openness (as assessed by the BFI) and its low observability, particularly to strangers, we expected that self- and informant-ratings would better predict Openness-relevant behavior than stranger-ratings, with a slight advantage to informant-ratings.

Table 9 provides the correlations between self-, informant-, and stranger-ratings of Openness and Openness-relevant behavior, as captured by the EAR. With one exception (substance use), stranger-ratings of Openness generally failed to predict relevant behavior. Informant-ratings of Openness tended to more strongly (though still modestly) predict relevant behavior. Specifically, aggregated informant-ratings predicted substance use ($r = 0.22$), and the single informant estimate predicted unconventionality ($r = 0.26$). Self-ratings predicted only EAR-coded creativity ($r = 0.22$). The relations between self- and informant-rated Openness and creativity and unconventionality, in particular, are noteworthy due to the extremely low base rates for each of these behavioral categories.

Overall, it would seem that the EAR (or, again, the coding system employed in the current study) generally fails to capture Openness-relevant behavior. However, to the extent that it does, informant-ratings seemed to be the most valid of the three sources.

5. Discussion

We obtained personality measures from three sources (self, knowledgeable informants, and strangers) in an attempt to provide a preliminary test of the predictive validity of each with respect to a naturalistic behavioral criterion. Due to the intensive nature of data collection and coding, we were not able to collect enough data to provide precise estimates of the effects we examined, and so our results should be interpreted with some degree of caution. Rather than drawing firm conclusions from these results, we encourage readers to consider this study as a framework for future research to provide more evidence regarding this important research question. The summary of our findings that follows should be read with the understanding that there is still a great deal of uncertainty around all of these results.

At the outset, we made three general predictions. First, we predicted that self-ratings of personality would predict relevant acoustically-encoded behavior in the domains of Extraversion, Agreeableness, Conscientiousness, and Neuroticism. In fact,

Table 9
Correlations between ratings of openness to experience and relevant acoustically-captured behavior.

Predictor	Self	Informant		Stranger	
		Aggregate	Single	Aggregate	Single
Activities					
Substance Use	.06 [−.16, .27]	**.22** [−.02, .44]	**.21** [−.04, .43]	**.21** [.00, .41]	.15 [−.07, .35]
Type of conversation					
Deep	.11 [−.11, .31]	−.12 [−.36, .12]	−.14 [−.37, .10]	.12 [−.10, .32]	.08 [−.14, .29]
Big five aspects					
Creativity	**.22** [.01, .41]	.17 [−.07, .40]	.16 [−.09, .39]	−.04 [−.25, .18]	−.03 [−.24, .19]
Unconventionality	.03 [−.19, .24]	.18 [−.07, .41]	**.24** [.00, .46]	.05 [−.17, .26]	.03 [−.19, .24]

Note. N = 83 (Self), 65 (Informant), 84 (Stranger). Values in brackets represent lower and upper bounds of 95% confidence intervals. Correlations stronger than |.20| appear in bold.

self-reports only predicted behavior to an acceptable degree in the domain of Extraversion. One possible explanation for these results is that the evaluativeness of Agreeableness and Conscientiousness impaired the accuracy of self-reports, and that Neuroticism may in fact be more evaluative (and therefore harder for the self to judge accurately) than previous findings suggest. Looming large over this general predictive failure, however, is the fact that informant and/or stranger perspectives are likely more similar to those of the coders than those of the self. In other words, it is likely that informant and stranger personality ratings are based heavily on behavioral observation, whereas self-ratings are likely based on a combination of behavior and mental states (thoughts, feelings, intentions, etc.).

We also predicted that informant-ratings would predict behavior for Extraversion, Agreeableness, and Conscientiousness, whereas in actuality, these ratings predicted relevant behaviors for Extraversion, Neuroticism and Openness. These findings are inconsistent with previous research that suggests that Neuroticism and Openness are hard for others to judge because of their low observability (but see Vazire & Solomon, 2015, for a discussion of contextual variation in the observability of traits). If future research corroborates these findings, that would suggest that informants may be a better source than previously thought for assessing Neuroticism and Openness. However, our final general prediction—that stranger-ratings of Extraversion would predict relevant target behavior as well as self or informant-ratings—was strongly supported.

5.1. Limitations

To our knowledge, this is the only study to date that has examined the validity of trait judgments from the self, informants, and strangers with respect to a naturalistic behavioral criterion. This key advantage, however, is offset by a few limitations. First and foremost, in a current scientific culture keenly (and appropriately) interested in power and replicability, the central analyses included in this paper are based on a fairly small sample size. In our opinion, the incorporation of real-life behavior—and the methodological and analytic difficulties that such incorporation poses—justifies the value of this study despite its small sample size, but this does not change the fact that our results are inconclusive. Future endeavors should aim to (a) gather similar data from more people, (b) employ longer (than two days) observation periods, (c) employ multiple (non-contiguous) observation periods, and (d) include measures for increasing the reliability of the behavioral coding (i.e., multiple coders for all files, refinement of coding system). Although accomplishing these goals simultaneously is a somewhat daunting task, we believe that this would provide a clearer picture of the validity of these judgments and allow for some more sophisticated data analysis aimed at more detailed research questions. In the interim, researchers without access to the resources to conduct such a study should continue to collect behavioral data, and publish it without selection for significance (i.e., publish all results including null and small effects), in order to provide unbiased material for later synthesis via meta-analytic review.

Secondly, although a chief advantage of this particular study, the prediction of behavior in natural settings is only one method of criterion validation. In the end, the general concept of a personality trait includes consistency in affective experience, cognition, and motivation in addition to consistent patterns in behavior. Self reports, for example, did not grade out well in our analyses, but this may not be an entirely fair test. Due to the self's unique perspective, perhaps one would expect that self-judgments might fail to predict some behavior. On the other hand, for outside observers, the adage "behavior engulfs the field" takes on particular importance in this instance. Principally in the case of the strangers, these trait judgments are formed solely on small snippets of behavior. Thus, it would follow that predicting other behavior would be simpler for them than, say, predicting emotional responses to events or understanding the reasons for these actions. That said, we do not feel that this simple fact makes the validity of stranger-ratings tautological. The strangers in this study only had access to a few statements made by each target and about a minute's worth of nonverbal behavior in a laboratory setting. The fact that this information could be utilized to make summary personality judgments that in turn predicted behavior in real-world settings is still fairly impressive.

Nevertheless, our criterion measure was restricted to observable (i.e., behavioral) manifestations of traits, and thus likely favored observer- (i.e., informant- and stranger-) reports over self-reports. To the extent that readers agree that behavioral manifestations of traits are most central to personality, this is not necessarily a weakness of our criterion measure. However, future research should try to incorporate criterion measures that also capture less overt manifestations of the Big Five (e.g., subtle linguistic cues, behavioral residue, life outcomes).

5.2. Future directions

The work discussed here represents a small step toward understanding the utility and validity of trait judgments from various sources. At its heart, this is among the most basic of research programs in personality. However, seventy-plus years into the investigation of personality traits, our sense of how actual behaviors tend to cluster together in individuals is still somewhat limited. The SOKA model can be used as an orienting principle for elaborating on this network.

More practically, at the very least, these data, in conjunction with recent findings (Vazire, 2010; Vazire & Mehl, 2008) indicate that there may be key differences in the validity of trait ratings by source, meaning that in practical application, we should rely on different predictor variables depending upon the circumstance and focal dimension. For example, the potential of a prospective

sales representative might best be evaluated by a panel of strangers, whereas clinicians might choose to seek informant reports of personality when considering a client's propensity for melancholy in the future (but see Carlson, Vazire, & Oltmanns, 2013). For researchers investigating accurate personality perception, perhaps the same accuracy criterion is not appropriate across traits. For example, self-ratings may serve as an appropriate accuracy criterion for peer ratings of say, Neuroticism, but for Openness, knowledgeable informant-ratings may better index accuracy.

Future research should also continue to improve upon our current methods of naturalistically measuring behavior. Specifically, the rather coarse coding categories for acoustically-detectable behavior probably underestimate the predictive validity of trait judgments. Aside from drastically altering the method of data acquisition (e.g., using a video analog to the EAR – which has important ethical constraints), there may be some ways to adjust coding methods that could prove illuminating. For example, rather than employing pure counting methods, coders might be instructed to listen to small segments of recordings and then rate the personality state (Fleeson, 2007) of the individual at various time points. Furthermore, methods such as this could be jointly employed with attempts to measure context or situation type (Rauthmann et al., 2014), allowing for a slightly more objective evaluation of person-situation interactive effects (Sherman, Nave, & Funder, 2010).

6. Conclusion

Our study provides a framework for examining the fundamental question of whether different perspectives – the self versus others – have complementary strengths and weaknesses when it comes to accurately judging a person's personality. We also provided the test of the validity of these different perspectives against a naturalistic, objective behavioral criterion measure. One lesson from this study is that the EAR may be an ideal method for collecting behavioral measures of personality for some domains (e.g., Extraversion, Neuroticism) but not others (e.g., emotions, Openness). For those domains for which the EAR provides reliable data (e.g., Extraversion and Neuroticism), one intriguing possibility is that the EAR could potentially pick up on some trait-relevant behavior that neither the self nor close others can judge very accurately (e.g., some aspects of Conscientiousness). Behavioral measures, when reliable, could also help adjudicate differences between the self and close others. These applications could be extended to other domains where self- and other-reports may be untrustworthy or may disagree (e.g., relationship behaviors, workplace behaviors, parent-child interactions). We hope this approach will be used to address questions about self- and other-knowledge in a wide range of contexts (Vazire & Wilson, 2012), questions that may have been intractable without the ability to unobtrusively and objectively observe people's naturalistic behavior.

Appendix A

	Upstate EAR Coding Categories
General	
EAR	Subject is talking about the EAR or the study
Interesting	Sound file seems interesting to you
Noise	High noise or interference
Problem	No entry = everything ok; 1 = insufficient acoustic information or silence; 2 = bad recording quality
Talking	Subject is talking; includes non-fluencies. Talking to herself would be 'with others' = 0; when subject is on the phone, mark this category
With 1 other	Subject is with only one other person: dyadic interaction
With Group	Subject is with a group of people – more than one *other* person
On the Phone	Subject is on the phone
Friend/Acquaintance	Subject is talking to platonic friend or acquaintance
Partner	Subject is talking to romantic partner (spouse, boyfriend or girlfriend)
Other Family/Relative	Subject is with other family members. When in doubt mark "friend/acquaintance."
Laughing	Subject laughing
Singing	Subject is singing or whistling
Crying	Subject crying
Mad/Arguing	Subject is arguing with, yelling at, screaming at, shouting at, mad at another person. You detect any anger in voice (can be frustration)
Sighing	Subject is sighing: an exaggerated exhalation of breath.
Coughing/Sneezing	Any sign of illness; coughing, sneezing, sniffing, or any acoustic sign of health problems This does not include complaints about or references to physical symptoms
Socializing	Subject is socializing or hanging out with others (e.g. watching a movie with other people)
In class	Subject is in class
Computer	Subject is working on the computer (may hear typing or clicking)
Working	Subject is working (e.g. in an office, retail business).
Church	Subject is attending a service, at a bible study group, choir practice, etc.
Eat/Drink	Subject is eating or drinking
TV	TV is on, irrespective of whether it is just playing in the background or subject is engaged in watching a movie
Sleeping	Subject is sleeping – can infer if you hear nothing else, and it is at night or very early in the morning
Studying	Subject reports studying in the diary or is overheard studying audibly with someone else. You might also hear pages turning
Radio/Music	There is some music or radio in the background. It is not important whether the subject is *only* listening to music. Live music at a concert or party counts. Music from a movie or TV does not

Appendix A (*continued*)

	Upstate EAR Coding Categories
Substance Use/Abuse	Either discussion of or suspicion of substance use or abuse; includes any present party referring to drugs or alcohol or the participant behaving in a manner which suggests intoxication
Apartment/Home	Subject is at home or at someone else's house
Outdoors	Subject is outdoors, defined as subject is able to see the sky. Driving in a car or bus would NOT be considered as outdoors
In Public	Subject is in any other public place (wherein they may be observed by or observe others)
In Transit	Subject is in transit (e.g., in a car, bus, or is walking)
Unknown	Cannot determine location.
Conversation Type	
Practical	Participant is talking about practical everyday things. The information exchanged serves a pragmatic purpose in the participant's everyday life. Can include making plans, discussing what is for dinner, picking up kids, travel arrangements e.g. "We need parmesan cheese for our dinner tonight"; "I will pick up Sally on Monday and Wednesdays at 5"
Small Talk	The purpose of this interaction is completely non-instrumental. No (or very trivial) information is exchanged; everything would be the same if the conversation never happened. e.g. "how's the weather?"; "I stepped on something"; "What are you up to?"
Deep/Substantive Conversation	Any conversation that has the purpose to exchange thoughts, information, values, ideas about a (non-emotional) topic; it could be about news of the day, about political issues, philosophical topics, theoretical ideas; information only The conversation does not really have to be "deep" but must have "substance." e.g., "Aren't Muslims not supposed to drink alcohol?"; "you heard that the WTC was attacked?"; "I found this book interesting."; "Guns n' Roses has a real rock n' roll sound to them."
Personal/Emotional Disclosure	Participant sharing of own personal feelings or emotions. Can include talking about their or a parent's divorce and their hopes and dreams for the future. The conversation passes a threshold of being trivial for the participant e.g., "I feel so terrible"; "I am scared about my grades in class"; "I have a crush on X" There are not accusatory statements, such as "It pisses me off when you talk with other women at parties." This constitutes complaining
Gossip	Participant is talking (e.g., divulging personal information) about another person while they are not there. Spreading rumor/reputational information about another person in their absence. This does not have to be negative e.g. "X talked back and that's why he was fired"; "Did you hear about their break-up?"; "Frank is so silly sometimes"; "Did you hear the lead singer of Gun n' Roses is dating X?"
FFM-Relevant Behaviors (mark 1 if any occur in a file)	
Dominance	Any command given to another person; any attempt to control another person's behavior; any statement about influencing others
Affiliation	Any statement of group membership or mention of a group to which the subject belongs
Energy	Excited or anticipatory speech (e.g., "let's go!" or "I'm excited/ready"); any verbal or nonverbal behavior that indicates active engagement
Sympathy	Verbal or nonverbal behavior indicating that the subject feels with or for someone else – could be a knowing sigh or "awwhh, that's tough" type statement
Concern for Others	A sincere inquiry about the well-being of another person (e.g., "how's your grandmother doing?" or a non-routine "how are you?"). Also, any helping behavior goes here
Affection	Participant expresses admiration or love for living beings (not only for romantic partner). Can be a compliment or using pet names. Not a backhanded compliment or sarcastic display of affection. Participant is being sincere beyond a social script. e.g. "You are beautiful"; "I love you"; "Felix is a fun cat"; "I love your cooking."
Complaining/Whining	Participant blames someone or something, complains, whines (not constructive criticism). e.g. "I hate the wind"; "Why do I have to be here?"; "That's a dumb idea."
Worry/Anxiety	Participant states concern of upcoming events (e.g., "I'm nervous about my Biology exam") or sounds tense or nervous when discussing an event
Sadness/Melancholy	Will often coincide with the Sadness emotion category, this category may also involve discussion of past sadness
Cleaning/Organizing	Subject is sweeping, picking up objects, making the bed, folding clothes, cooking, doing the dishes, house maintenance issues
Hygiene	Subject is brushing teeth, taking a shower, or washing hands. Any personal maintenance behavior
Tending to Responsibilities	Subject is paying bills or making systematic plans to accomplish either proximal or distal goals
Spirituality	Participant is talking about a spiritual or religious topic – not "Karma's a B****" or "Oh my god." Important: could be any religion or notion of god. And not church discussion (coded elsewhere). e.g. soul, universe, afterlife, karma, God, or lack thereof

(*continued on next page*)

Appendix A (*continued*)

	Upstate EAR Coding Categories
Creativity	Subject is participating in creative pursuits (e.g., making music) or discussing creative pursuits (e.g., art, literature)
Unconventionality	Participant discusses his or her uniqueness in some domain "I only use my left hand to eat") or desires to do something "out of the norm." Examples of the latter could range from expressing a desire to teach overseas or travel aimlessly to skydiving. The point is that the person is setting him or herself and his or her actions apart from others
Emotions (mark "1" if subject displays any of these emotions in a given audio file)	
Happiness	Any verbal or nonverbal display of happiness. You may mark this category if you have already marked laughing, but statements of happiness also apply
Sadness	Any verbal or nonverbal display of sadness. Mark this category if you have already marked crying, but statements of sadness also apply. ∗*merged with sadness/melancholy*∗
Anger/Frustration	Any verbal or nonverbal display of anger or frustration, e.g., angry exclamations, growl-like utterances
Fear	Any verbal or nonverbal display of fear (e.g., "I'm scared" or a shriek)
Surprise	Any verbal or nonverbal display of surprise (e.g., "wow" or a sudden excited noise)
Disgust	Any verbal or nonverbal display of disgust (e.g., "that's gross" or "yuck!" or "eww")

References

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*, 1621–1630. http://dx.doi.org/10.1037/0022-3514.49.6.1621.

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

Andersen, S. M., & Ross, L. (1984). Self-knowledge and social inference: I. The impact of cognitive/affective and behavioral data. *Journal of Personality and Social Psychology, 46*, 280–293. http://dx.doi.org/10.1037/0022-3514.46.2.280.

Ashton, M. C., Jackson, D. N., Paunonen, S. V., Helmes, E., & Rothstein, M. G. (1995). The criterion validity of broad factor scales versus specific facet scales. *Journal of Research in Personality, 29*, 432–442. http://dx.doi.org/10.1006/jrpe.1995.1025.

Back, M. D., & Vazire, S. (2012). Knowing our personality. In S. Vazire & T. D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 131–156). New York, NY: Guilford Press.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*, 396–403. http://dx.doi.org/10.1111/j.1745-6916.2007.00051.x.

Beer, A., & Brooks, C. (2011). Information quality in personality judgment: The value of personal disclosure. *Journal of Research in Personality, 45*, 175–185. http://dx.doi.org/10.1016/j.jrp.2011.01.001.

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

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

Bollich, K. L., Rogers, K. H., & Vazire, S. (2015). Knowing more than we can tell: People are aware of their biased self-perceptions. *Personality and Social Psychology Bulletin, 41*, 918–929. http://dx.doi.org/10.1177/0146167215583993.

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

Carlson, E. N., Vazire, S., & Oltmanns, T. F. (2013). Self-other knowledge asymmetries in personality pathology. *Journal of Personality, 81*, 155–170.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092–1122. http://dx.doi.org/10.1037/a0021212.

Cuddy, A. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*, 631–648. http://dx.doi.org/10.1037/0022-3514.92.4.631.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69–106. http://dx.doi.org/10.1111/j.1529-1006.2004.00018.x.

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality, 75*, 825–862. http://dx.doi.org/10.1111/j.1467-6494.2007.00458.x.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55*, 149–158. http://dx.doi.org/10.1037/0022-3514.55.1.149.

Furr, R. M. (2009). The study of behaviour in personality psychology: Meaning, importance and measurement. *European Journal of Personality, 23*, 437–453. http://dx.doi.org/10.1002/per.726.

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. In R. Hogan, J. A. Johnson, S. R. Briggs, R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 795–824). San Diego, CA, US: Academic Press. http://dx.doi.org/10.1016/B978-012134645-4/50031-7.

Hasler, B. P., Mehl, M. R., Bootzin, R. R., & Vazire, S. (2008). Preliminary evidence of diurnal rhythms in everyday behaviors associated with positive affect. *Journal of Research in Personality, 42*, 1537–1546. http://dx.doi.org/10.1016/j.jrp.2008.07.012.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality, 8*, 149–162.

Holtzman, N. S., Vazire, S., & Mehl, M. R. (2010). Sounds like a narcissist: Behavioral manifestations of narcissism in everyday life. *Journal of Research in Personality, 44*, 478–484. http://dx.doi.org/10.1016/j.jrp.2010.06.001.

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

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. http://dx.doi.org/10.1111/j.1467-6494.1993.tb00781.x.

Jones, E. E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior*. Morristown, NJ: General Learning Press.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311–337. http://dx.doi.org/10.1111/j.1467-6494.1996.tb00513.x.

Kwan, V. Y., Diaz, P., Wojcik, S. P., Kim, S. Y., Matula, K. A., & Rodriguez, K. (2011). Self as the target and the perceiver: A componential approach to self-enhancement. *Psychological Studies, 56*, 151–158. http://dx.doi.org/10.1007/s12646-011-0072-3.

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology, 98*, 668–682. http://dx.doi.org/10.1037/a0018771.

Leising, D., Gallrein, A. B., & Dufner, M. (2014). Judging the behavior of people we know: Objective assessment, confirmation of preexisting views, or both? *Personality and Social Psychology Bulletin, 40*, 153–163. http://dx.doi.org/10.1177/0146167213507287.

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

Luft, J., & Ingham, H. (1955). The Johari window, a graphic model of interpersonal awareness. In *Proceedings of the Western Training Laboratoryin Group Development*. Los Angeles: University of California, Los Angeles.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin, 132*, 895–919. http://dx.doi.org/10.1037/0033-2909.132.6.895.

Mehl, M. R. (2006). The lay assessment of subclinical depression in daily life. *Psychological Assessment, 18*, 340–345. http://dx.doi.org/10.1037/1040-3590.18.3.340.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*, 862–877. http://dx.doi.org/10.1037/0022-3514.90.5.862.

Mehl, M. R., & Holleran, S. E. (2007). An empirical analysis of the obtrusiveness of and participants' compliance with the electronically activated recorder (EAR).

European Journal of Psychological Assessment, 23, 248–257. http://dx.doi.org/10.1027/1015-5759.23.4.248.
Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*, 857–870. http://dx.doi.org/10.1037/0022-3514.84.4.857.
Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorded (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments & Computers, 33*, 517–523. http://dx.doi.org/10.3758/BF03195410.
Mehl, M. R., & Robbins, M. L. (2012). Naturalistic observation sampling: The Electronically Activated Recorder (EAR). In M. R. Mehl, T. S. Conner, M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 176–192). New York, NY, US: Guilford Press.
Mehl, M. R., Vazire, S., Holleran, S. E., & Clark, C. S. (2010). Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological Science, 21*, 539–541. http://dx.doi.org/10.1177/0956797610362675.
Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science, 317*(5834), 82. http://dx.doi.org/10.1126/science.1139940.
Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin, 35*, 1661–1671. http://dx.doi.org/10.1177/0146167209346309.
Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4*, 681–691. http://dx.doi.org/10.1037/h0024002.
Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*, 401–421. http://dx.doi.org/10.1146/annurev.psych.57.102904.190127.
Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609. http://dx.doi.org/10.1037/0022-3514.46.3.598.
Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality, 66*, 1025–1060. http://dx.doi.org/10.1111/1467-6494.00041.
Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, R. F. Krueger, R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York, NY, US: Guilford Press.
Ramírez-Esparza, N., Mehl, M. R., Álvarez-Bermúdez, J., & Pennebaker, J. W. (2009). Are Mexicans more or less sociable than Americans? Insights from a naturalistic observation study. *Journal of Research in Personality, 43*, 1–7. http://dx.doi.org/10.1016/j.jrp.2008.09.002.
Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., & Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology, 107*, 677–718. http://dx.doi.org/10.1037/a0037250.
Robbins, M. L., Focella, E. S., Kasle, S., López, A. M., Weihs, K. L., & Mehl, M. R. (2011). Naturalistically observed swearing, emotional support, and depressive symptoms in women coping with illness. *Health Psychology, 30*, 789–792. http://dx.doi.org/10.1037/a0023431.
Robins, R. W., Spranca, M. D., & Mendelsohn, G. A. (1996). The actor-observer effect revisited: Effects of individual differences and repeated social interactions on actor and observer attributions. *Journal of Personality and Social Psychology, 71*, 375–389. http://dx.doi.org/10.1037/0022-3514.71.2.375.
Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609–612. http://dx.doi.org/10.1016/j.jrp.2013.05.009.
Sherman, R. A., Nave, C. S., & Funder, D. C. (2010). Situational similarity and personality predict behavioral consistency. *Journal of Personality and Social Psychology, 99*, 330–343. http://dx.doi.org/10.1037/a0019796.
Slatcher, R. B., & Robles, T. F. (2012). Preschoolers' everyday conflict at home and diurnal cortisol patterns. *Health Psychology, 31*, 834–838. http://dx.doi.org/10.1037/a0026774.
Slatcher, R. B., & Trentacosta, C. J. (2011). A naturalistic observation study of the links between parental depressive symptoms and preschoolers' behaviors in everyday life. *Journal of Family Psychology, 25*, 444–448. http://dx.doi.org/10.1037/a0023728.
Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193–210. http://dx.doi.org/10.1037/0033-2909.103.2.193.
Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality, 40*, 472–481. http://dx.doi.org/10.1016/j.jrp.2005.03.003.
Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*, 281–300. http://dx.doi.org/10.1037/a0017908.
Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology, 95*, 1202–1216. http://dx.doi.org/10.1037/a0013314.
Vazire, S., & Solomon, B. C. (2015). Self- and other-knowledge of personality. In M. Mikulincer, P. R. Shaver, M. L. Cooper, & R. J. Larsen (Eds.), *Handbook of personality processes and individual differences* (pp. 261–281). Washington, DC: American Psychological Association.
Vazire, S., & Wilson, T. D. (2012). *Handbook of self-knowledge*. New York, NY: Guilford Press.
Watson, D. (1982). The actor and the observer: How are their perceptions of causality divergent? *Psychological Bulletin, 92*, 682–700. http://dx.doi.org/10.1037/0033-2909.92.3.682.
Watson, D. (1989). Strangers' ratings of the five robust personality factors: Evidence of a surprising convergence with self-report. *Journal of Personality and Social Psychology, 57*, 120–128. http://dx.doi.org/10.1037/0022-3514.57.1.120.
Watson, D., Hubbard, B., & Wiese, D. (2000). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology, 78*, 546–558. http://dx.doi.org/10.1037/0022-3514.78.3.546.
Wilson, R. E., & Vazire, S. (2015). Taking personality to the next level: What does it mean to know a person? In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* (pp. 1–13). http://dx.doi.org/10.1002/9781118900772.etrds0327.
Wood, D., & Wortman, J. (2012). Trait means and desirabilities as artifactual and real sources of differential stability of personality traits. *Journal of Personality, 80*, 665–701. http://dx.doi.org/10.1111/j.1467-6494.2011.00740.x.