



Eavesdropping on social life: The accuracy of stranger ratings of daily behavior from thin slices of natural conversations

Shannon E. Holleran *, Matthias R. Mehl, Stephanie Levitt

University of Arizona, Psychology, Tucson, AZ 85721, United States

ARTICLE INFO

Article history:
Available online 8 April 2009

Keywords:
Zero-acquaintance
Thin slice
Personality judgment
Gender stereotypes
First impression

ABSTRACT

In two studies the authors examined the accuracy of stranger ratings of daily behavior based on thin slices of natural conversations. Methodologically, the studies extend past research by using a behavioral accuracy criterion, benchmarking zero-acquaintance accuracy against target and informant accuracy, and employing a representative design that sampled contexts from targets' daily situations. Theoretically, the studies investigate how stereotypes influence the accuracy of first impressions depending on their sample-based validity. Across both studies, after listening to five conversational snippets (2.5 min total), the ratings of strangers were as accurate as the targets' and informants' ratings. Further, ratings for gender-stereotypic behaviors with a kernel of truth resulted in greater initial accuracy than ratings for gender-stereotypic behaviors with no kernel of truth.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Every day people witness bits and pieces of conversations from people they do not know. Whether waiting in line at the grocery store or overhearing a chat at the next table during lunch, people routinely and intuitively form impressions about other people. Recently, research on the accuracy of first impressions has flourished with a surge of studies showing that they can be surprisingly accurate (Ambady & Rosenthal, 1992; Back, Schmukle, & Egloff, 2008; Blackman & Funder, 1998; Borkenau & Liebler, 1992; Chaplin, Phillips, Brown, Clanton, & Stein, 2000; Gosling, Ko, Mannarelli, & Morris, 2002; Letzring, Wells, & Funder, 2006; Marcus, Machilek, & Schütz, 2006; Paulhus & Bruce, 1992; Rentfrow & Gosling, 2006; Vazire & Gosling, 2004).

The research we report here had two major purposes: first, it sought to address three methodological challenges in zero-acquaintance research revolving around the assessment of accuracy and provide a novel solution for obtaining an accuracy criterion that is ecological and behavioral in nature and free of targets' self-reports. Second, it sought to test how stereotypes affect the accuracy of first impressions depending on their validity, that is the existence of a kernel of truth in the stereotype. We report data from two studies that used naturalistically observed act-frequencies of daily behaviors as accuracy criterion. Study 1 compared the accuracy of stranger ratings of targets' daily behavior

from five snippets (or 2½ min) of their natural conversations against a theoretically and practically important "benchmark", the accuracy that the targets themselves and their close acquaintances achieved. Study 2, then, experimentally varied the amount of information – from 1 to 10 conversational snippets (i.e. 30 second to 5 minute) – and tested how accuracy changed as a function of the information quantity.

With respect to the second aim, our study design allowed us to test how stereotypes affect judgmental accuracy with different amounts of available information. Perceivers with little or no individuating information about a person tend to rely on characterizations associated with categorical information they know about that person (i.e. stereotypes). It is a classic finding in the field that relying on such stereotypes can have negative consequences, namely lead to biased or erroneous judgments. However, it has also been theorized that stereotypes with some validity, can facilitate accurate judgments (Judd & Park, 1993). Existing person perception models (e.g. Kenny, 1994, 2004) make testable predictions about how the presence or absence of a kernel of truth in stereotypes affects accuracy, but – due to methodological challenges in operationalizing stereotype accuracy – these predictions are difficult to submit to empirical tests.

Here, we operationalized stereotype accuracy in a novel way. We directly compared perceived gender differences in daily behavior (e.g. talking on the phone, spending time at the computer) to actual, sample-based gender differences in these behaviors. That way, we empirically determined which stereotypes did and did not have a kernel of truth in our data. In Study 1 we examine how the use of gender stereotypes affects the accuracy of

* Corresponding author.
E-mail addresses: shollera@email.arizona.edu (S.E. Holleran), mehl@email.arizona.edu (M.R. Mehl).

zero-acquaintance ratings of daily behavior. In Study 2 we examine how the effect of gender stereotypes on accuracy varies as a function of the amount of information that is available about a person.

1.1. Methodological challenges in zero-acquaintance research

Before we describe the theoretical background of the studies in more detail, we briefly highlight three methodological issues in zero-acquaintance research: the lack of an objective accuracy criterion, the lack of meaningful empirical benchmarks for interpreting accuracy levels, and the representativity of the context in which judgmental accuracy is studied.

1.1.1. The lack of an objective accuracy criterion

The question of how to assess accuracy has long preoccupied the field (Kruglanski, 1989). Consensus, or the level of agreement that two or more people achieve about the judgment of a target person, is often used as an indicator of accuracy (Kenny & West, 2008). Consensus, though, is a necessary but not a sufficient condition for accuracy because perceivers can agree in their perception but still be wrong (Blackman & Funder, 1998). Self-other agreement is the most common way to assess accuracy (e.g. Gill, Oberlander, & Austin, 2006; Holleran & Mehl, 2008; Kolar, Funder, & Colvin, 1996). Conceptually, self-other agreement assumes that perceptions are accurate if they correspond to targets' self-perceptions. Self-other agreement represents true accuracy to the extent that the self is accurate about a judgment. Self-perceptions, however, are often far from perfect reflections of reality as they tend to be subject to biases and based on insufficient or invalid information (Paulhus & Vazire, 2007; Wilson & Dunn, 2004).

Realistic accuracy is a third possible criterion for accuracy (Funder, 1995). Realistic accuracy is a hypothetical construct that represents judgments which use multiple criteria as indicators of accuracy. Conceptually, realistic accuracy assumes that accuracy cannot be achieved by using one type of measurement or criterion, but instead, suggests that accuracy be estimated through the use of multiple methods such as self-, peer- and clinical judgments, behavioral assessments, and physiological measurements (Letzring et al., 2006).

In the real world, however, people rarely have the opportunity to gather information from several methodologically different sources such as a person's therapist, standardized observation, and physiological measurement. Instead, most people will ask a few of the person's closest friends or look at a few behaviors the person engages in to gain information about the person. Research on informant reports has recently received increasing attention (Vazire, 2006), but the direct observation of "actual behavior" is underrepresented in the field (Baumeister, Vohs, & Funder, 2007; Furr, in press). This is unfortunate given that "behavioral ratings are often the best possible way to measure the person's actual standing on a trait" (Kenny, 1994; p. 135).

1.1.2. Benchmarks for interpreting levels of accuracy

Is an accuracy correlation of .20 or .30 high or low? In the absence of standards for what constitutes high accuracy, similar estimates can result in different conclusions. Thus, benchmarks are needed to evaluate the magnitude of zero-acquaintance effects. What assessment perspectives could serve as such benchmarks?

The self is often implicitly considered the most accurate assessment perspective, but, as pointed out above, the self also routinely falls short of being a gold-standard because people are often unaware of their internal states and self-presentation concerns can be in the way of valid assessments. To the extent that the self's privileged role in person perception is questioned, other people who know the target well (e.g. spouses, friends) are often considered the fall back option. Research comparing the accuracy of self and

informant reports against behavioral criteria suggests that informant reports are usually right on par with self-reports for predicting psychologically-relevant behavior (Kolar et al., 1996; Spain, Eaton, & Funder, 2000).

Regardless of the relative accuracy of self- or informant reports, though, using the two assessment perspectives as benchmarks in zero-acquaintance research would be theoretically and practically desirable because of the eminent role that they play in person perception. Doing so, however, requires that accuracy is assessed independent of self- and informant reports. One way to achieve this is to use directly observed behavior as a criterion (Vazire & Mehl, 2008).

1.1.3. The representativity of contexts in accuracy research

In zero-acquaintance research, the generalizability of the accuracy estimate is of great importance. Often, though, it is ultimately limited by the specificity of the contexts being studied. Funder and West (1993) noted that "research needs to be conducted in diverse, commonly experienced naturalistic settings or in the laboratory using the full array of stimuli available in the naturalistic setting" (p. 463). To the extent that the goal is to understand how first impressions "work" in daily life, it would be desirable for the studied contexts to represent the full range of potential real-world contexts. Brunswik's (1956) concept of representative designs postulates that for research to generalize to its intended real-world phenomenon, both participants and social contexts have to be considered random factors and need to be sampled representatively from their underlying populations of individuals and ecologies of situations. Thus, it would be desirable to not only sample participants from the population of potential targets but also behaviors from the ecology of targets' naturally-occurring daily behaviors.

1.2. A naturalistic observation approach to studying zero-acquaintance judgments

One solution to these methodological challenges lies in the use of behavioral observation in zero-acquaintance research (Funder & Sneed, 1993; Kenny & West, 2008). In the current studies we used the *Electronically Activated Recorder* to unobtrusively sample behavior in naturalistic settings (EAR; Mehl, Pennebaker, Crow, Dabbs, & Price, 2001). The EAR¹ is a modified digital voice recorder that periodically records brief snippets of ambient sounds. Participants wear the EAR attached to their belt or in a purse-like bag while going about their daily lives. The method is unobtrusive because the EAR operates imperceptibly. In recording moment-to-moment ambient sounds, it yields acoustic logs of people's days as they naturally unfold. These acoustic logs, then, can be used as stimulus material for person perception studies.

Recently, Mehl, Gosling, and Pennebaker (2006) used this approach for examining the accuracy of implicit folk theories of personality. Judges listened to 2 days worth of EAR sound files (>100 per target) and made judgments about the targets' personalities. Consistent with prior research, accuracy ratings were highest for Extraversion as a highly observable trait. Trait ratings of Neuroticism also achieved high accuracy presumably because over time the EAR captured participants "off stage" in situations where self-presentation concerns were low.

This study successfully addressed the concerns around the representativity of contexts by giving raters access to a systematic sample of targets' daily behavior. However, it failed to address the other two methodological limitations. The study used self-reports as accuracy criterion and the design did not allow bench-

¹ The original data set included 80 participants. The analyses here are based on 78 participants because two targets did not meet the criteria of having at least five conversations of five or more words during the four days of EAR monitoring.

marking the findings against other assessment perspectives. Further, judges pragmatically based their ratings on about an hour and a half of ambient sounds recorded from the targets' daily lives, rendering the design ultimately unrealistic for studying naturalistic first impressions (cf. Mehl, 2006). In the real world, people rarely have access to a representative sample of behavioral acts about a person.

The current studies sought to use the EAR method to extend prior zero-acquaintance research. Specifically, we had naïve judges listen to a limited number of EAR sound files and rate targets with respect to the frequency with which they engaged in a set of daily behaviors (e.g. talking on the phone, listening to music, laughing). We then compared these ratings against the same set of daily behaviors naturalistically observed over a period of 4 days using the EAR method. Through the use of an ecological and behavioral accuracy criterion that is independent of self- and peer reports, we could compare the judges' accuracy against the accuracy that the targets themselves and their peers achieved in rating the same behaviors—two theoretically and practically important benchmarks in person perception research.

The use of this behavioral accuracy criterion allowed us to look at an important person perception issue that has received considerable theoretical but only limited empirical attention—the role that stereotypes play in facilitating and/or undermining accurate personality judgments.

1.3. The effects of stereotypes on the accuracy of personality judgments

The information that judges have in zero-acquaintance studies can be divided into behavioral and categorical information (Kenny, 1994). With little or no behavioral information, first impressions are primarily based on categorical information. As behavioral information becomes available, judges incorporate such individuating information into their impression.

The complementary influence of behavioral and categorical information is modeled in Kenny's classic person perception models (WAM; PERSON; Kenny, 1994, 2004). In these models, the "kernel of truth" parameter refers to the extent to which a stereotype about a perceived group difference in a behavior corresponds to an actual group difference in that behavior. The theoretical importance of this parameter is clear but empirically it has proven notoriously difficult to measure the "actual" independent of the "perceived" part (Lee, Jussim, & McCauley, 1995; Levesque & Kenny, 1993). One notable exception is Swim's (1994) comparison of perceived gender stereotypes with meta-analytic findings on gender differences (cf. Hall & Carter, 1999). Theoretically, stereotypes that contain a kernel-of-truth should facilitate and those that have none should undermine accuracy when minimal information is available. Empirically, though, this prediction has hardly been tested in personality psychology.

Two features of our study design allowed for a direct test of this prediction. First, auditory person perception stimuli minimize the use of stereotypes (Borkenau & Liebler, 1992). The EAR sounds eliminate information about the targets' physical appearance leaving their voice as essentially the only cue to categorical information about them. Given the homogeneity of student samples with respect to age and – in our case also – ethnicity, the use of stereotypes is thereby effectively limited to targets' gender—which naturally tends to be one of the most salient cues shaping first impressions (Fiske, Haslam, & Fiske, 1991). Second, our behavioral accuracy criterion—observed act-frequencies of daily behavior—render it possible to empirically determine the degree of validity in a stereotype. Specifically, it allows us to compare perceived gender differences in various daily behaviors (e.g. talking on the phone, spending time at the computer) to actual, sample-based gender differences in these behaviors. That way, we can categorize

daily behaviors as (a) gender-stereotypic with kernel of truth, (b) gender-stereotypic with no kernel of truth, and (c) gender neutral and estimate accuracy separately for the three types of behaviors.

2. Study 1: gender stereotypes and the accuracy of first impressions

Study 1 modeled the naturalistic person perception scenario in which a person gets to eavesdrop on snippets of strangers' conversation. Naïve judges rated the daily behavior of targets (e.g. how much time relative to the average person the targets spend alone, talking, or watching TV) after having listened to five 30 second sound bites of their natural conversation. As benchmark for the judges' accuracy, self and peer ratings of the target's behavior—assessed using the same measure were taken from data originally reported in Vazire and Mehl (2008). All three sources, that is, the judge, self, and peer ratings were then compared to how the targets actually behaved as documented by raw behavior counts derived from a 4-days EAR monitoring.

Based on prior zero-acquaintance research, we predicted that five short conversational snippets would provide enough information for strangers to rate targets' daily behavior with a significant level of accuracy. We further predicted that judges' accuracy would be substantial but lower than the accuracy that the targets themselves and their highly acquainted informants achieved (Kenny, 2004). With respect to the influence of gender stereotypes, we predicted based on theoretical person perception models (Kenny, 1994, 2004) that accuracy would be higher for gender-stereotypic behaviors with a kernel of truth than for gender-stereotypic behaviors with no kernel of truth with accuracy for gender-neutral behaviors falling in-between.

2.1. Method

2.1.1. Participants

110 undergraduate students at the University of Arizona served as naïve judges for the study. 47% of the judges were female, 75% White, 5% African American, 5% Asian, 9% Hispanic, 1% Native American, and 5% of another ethnicity. Judges ranged in age from 18 to 27 years old ($M = 19.3$, $SD = 1.3$). None of the judges knew any of the targets they rated.

2.1.2. Stimulus materials for the thin-slice ratings

The judges rated 78 targets (36 males, 42 females; mean age $M = 18.7$, $SD = 1.4$) who were the primary participants in a large EAR project. Details on the targets as well as the larger project are reported in Vazire and Mehl (2008). The targets wore the EAR for approximately 4 days during their waking hours from Friday afternoon until Tuesday night. It was set to record 30 s intervals every 12.5 min (or 4.8 recordings per hour). On average, the EAR recorded 308 sound files ($SD = 192$) per participant. Of all recorded sound files, on average, 33.5% ($SD = 14.9$) were conversations the targets had with other people. For each of the 78 targets, five conversations were randomly selected from their full set of EAR sound files using a random number generator. A conversation was selected if it contained five or more words by the target.

2.1.3. Judges' ratings of targets' daily behavior

The judges were told the study was about forming first impressions of other people. They were instructed to listen to all five sound files for each target and to then complete a set of questionnaires about their first impression of the target. Judges were divided into blocks of 8–10 and each judge rated 6 targets. That way, 72 targets were coded by 8 judges, 5 targets were coded by 9 judges, and 2 targets were coded by 10 judges. To control for

Table 1
Descriptive statistics (reliabilities and base rates) for the EAR-derived ACT behaviors.

EAR-derived ACT behaviors	Intercoder reliability	Base rates		
		All targets <i>M</i> (<i>SD</i>)	Male targets <i>M</i> (<i>SD</i>)	Female targets <i>M</i> (<i>SD</i>)
<i>Gender-stereotypic behaviors with kernel of truth</i>				
Laughing (<i>f</i>)	.89	7.6 (5.4)	5.8 (3.7)	9.1 (6.2)
On the computer (<i>m</i>)	.87	7.2 (8.9)	10.6 (11.0)	4.3 (5.4)
Attending class (<i>f</i>)	.99	4.3 (3.5)	3.3 (2.8)	5.1 (3.8)
Average	.92	6.4 (6.3)	6.4 (6.8)	6.2 (6.5)
<i>Gender-stereotypic behaviors with no kernel of truth</i>				
Listening to music (<i>m</i>)	.95	15.0 (9.7)	14.0 (9.7)	15.8 (9.7)
Talking on the phone (<i>f</i>)	.97	3.6 (3.0)	3.1 (3.4)	3.9 (2.5)
At work (<i>m</i>)	– ^a	2.4 (7.8)	1.7 (4.1)	3.0 (10.0)
Talking to opposite sex (<i>m</i>)	.95	6.9 (7.7)	6.5 (7.5)	7.2 (7.8)
Watching TV (<i>m</i>)	.95	16.2 (15.5)	14.8 (11.8)	17.4 (18.1)
Talking to same sex (<i>f</i>)	.95	14.4 (10.0)	14.0 (10.5)	14.7 (9.7)
At a restaurant/coffeeshop (<i>f</i>)	.91	2.6 (3.3)	2.6 (3.8)	2.6 (2.8)
With other people (<i>f</i>)	.97	31.7 (14.3)	31.9 (16.0)	31.6 (12.9)
Socializing (<i>f</i>)	.91	15.6 (12.2)	15.4 (11.8)	15.7 (12.6)
Talking one-on-one (<i>f</i>)	.94	21.6 (11.5)	21.4 (12.8)	21.7 (10.4)
Outdoors (<i>m</i>)	.90	3.5 (2.5)	3.3 (2.8)	3.7 (2.2)
Crying (<i>f</i>)	– ^a	0.1 (0.3)	0.1 (0.4)	0.1 (0.2)
Arguing (<i>m</i>)	– ^a	0.2 (0.5)	0.1 (0.3)	0.3 (0.5)
Average	.94	10.1 (9.1)	10.1 (8.8)	11.1 (9.3)
<i>Gender-neutral behaviors</i>				
Singing	.74	2.8 (3.2)	2.9 (3.2)	2.7 (3.3)
Communing	.89	6.2 (6.2)	5.6 (6.9)	6.6 (5.5)
Talking in a group	.88	10.1 (9.4)	10.5 (8.5)	9.9 (10.3)
Indoors	.88	60.9 (16.4)	60.4 (17.6)	61.3 (15.4)
Average	.85	20.0 (10.1)	19.9 (10.5)	20.1 (9.8)

Note: Intercoder reliabilities are intraclass correlations (ICC[2, k]); intercoder agreement was computed from a set of training EAR sound files (see Vazire & Mehl, 2008). Base rates are expressed as percentages of EAR sound files in which the behavior was present. Because in each condition, the sound files that judges listened to were deleted from the criterion, the base rates differed slightly across conditions. The base rates reported here are for the 10 sound file condition (Study 2). *f* = rated as stereotypically female, *m* = rated as stereotypically male.

^a Reliability could not be determined due to a lack of variance in the codings in the training set.

the effect of exposure (i.e. rater experience) on accuracy, the order in which the targets were rated was counterbalanced. The judges rated each target on the ACT questionnaire using a 7 point scale ranging from 1 (strongly disagree) to 7 (strongly agree). The ACT questionnaire is a measure that was designed to obtain ratings of a set of behaviors that can be directly assessed with the EAR (i.e. are detectable from ambient sound).

In the current study, we focused on the 20 behaviors identified as reliable and non-redundant by Vazire and Mehl (2008). Descriptive statistics (i.e. reliabilities and base rates) for these behaviors are provided in Table 1.² The mean inter-judge agreement across all 20 ACT items was .55, calculated as intraclass correlations based on one-way random effect models, (ICC[1, k]), (ranging from .18 for commuting to .81 for talking on the phone).

2.1.4. Targets' self-ratings and informant reports of the targets' daily behavior

Information on how the targets rated themselves and on how three informants who knew them well rated them on the ACT questionnaire was available from Vazire and Mehl (2008). In this study, these self- and peer ratings of the targets' daily behaviors were used as accuracy benchmarks against which the judges' ACT ratings could be compared.

² Note that the base rates for the EAR-derived ACT behaviors differed minimally between Study 1 and Study 2 and from condition to condition in Study 2 because the sound files that judges listened to were deleted from the respective criterion to avoid information overlap between source and criterion. Table 1 shows the base rates for the 10-sound file condition in Study 2 as they reflect the information from those sound files that were included in both studies and all conditions.

2.1.5. Behavioral accuracy criterion: targets' EAR-assessed daily behavior

A team of 10 research assistants coded the EAR sound files for the 20 behaviors that were assessed with the ACT questionnaire (Vazire & Mehl, 2008). Each target participant was coded by one coder and inter-coder reliabilities were determined from a set of training EAR recordings (221 sound files) independently coded by all research assistants. Intraclass correlations (ICC [2, k]) exceeded .70 for all categories. The raw codings were then converted into time-use estimates by calculating the percentage of a person's valid (i.e. compliant and codable) waking EAR recordings in which an ACT behavior was present (e.g. percentage of sound files in which the target was with people, laughing, at home, or in class).

2.1.6. Data analytic strategy

2.1.6.1. Computation of judges' accuracy. To test our predictions, we computed accuracy correlations between judges' ACT ratings and the corresponding EAR-assessed behavior counts. To ensure that our accuracy criterion had no overlap with the stimulus material on which judges based their impressions, we removed the five sound files that judges had listened to from each target's set of EAR recordings. We then computed two "versions" of accuracy correlations. First, we computed the correlations between the average of the judges' ratings and the targets' actual behaviors (i.e. behavior counts derived from the full set of EAR recordings). This accuracy index provides information about the extent to which our composite measure aggregated over eight judges corresponded to the targets' EAR-assessed behavior. We further computed a single-judge accuracy index by correlating each judges' ratings separately with the targets' EAR-assessed behaviors and then averaging across all of these single-judge accuracy correlations. This index considers that accuracy is a function of "test-length", that is the

number of ratings that are then correlated with the criterion (cf. Epstein, 1979; Moskowitz, 1982). The magnitude of the single-judge accuracy index, thus, is directly comparable to the magnitude of the accuracy for the targets' self-ratings and the informant reports (which were also computed as single-informant accuracy correlations). All accuracy correlations were averaged using Fisher's r -to- z formula.

2.1.6.2. Assessment of gender stereotypes and kernel of truth. To assess gender stereotypes, we rated each of the 20 ACT behaviors on gender-stereotypicality. In two separate ratings, six new judges (none of them participated in the thin-slice ratings) rated the extent to which each ACT behavior was stereotypically male and stereotypically female using a scale from 1 (not at all) to 7 (extremely). The two sets of ratings were correlated $r = -.54$. Using a median split, the behaviors that were rated as stereotypically female were: talking on the phone ($M = 6.1, SD = .67$), crying ($M = 5.7, SD = .78$), spending time with others ($M = 4.9, SD = 1.7$), talking one-on-one ($M = 4.9, SD = 1.5$), talking to same sex ($M = 4.8, SD = 1.7$), laughing ($M = 4.0, SD = 1.6$), going to coffee shops ($M = 4.5, SD = 1.4$), attending class ($M = 4.4, SD = 1.4$), and socializing ($M = 4.4, SD = 1.4$). The behaviors that were rated as stereotypically male were: arguing ($M = 5.3, SD = .87$), spending time outside ($M = 4.6, SD = 1.4$), talking to opposite sex ($M = 4.6, SD = 1.4$), working ($M = 4.6, SD = 1.5$), watching television ($M = 4.5, SD = 1.5$), listening to music ($M = 4.4, SD = 1.5$), and on the computer ($M = 4.2, SD = 1.3$). Finally, we refer to the following ACT behaviors as gender-neutral because they were rated as below the median for 'stereotypically male' and 'stereotypically female': Singing, commuting, talking in a group, spending time indoors.

We then tested for actual gender differences in these behaviors by comparing the means for male and female targets on the EAR-assessed behaviors. Three behaviors yielded significant differences: spending time on the computer ($M_{\text{males}} = 10.5\%$ vs. $M_{\text{females}} = 4.2\%$, $t = 3.32, p = .001$), attending class ($M_{\text{males}} = 3.3\%$ vs. $M_{\text{females}} = 5.1\%$, $t = 2.38, p = .02$), laughing ($M_{\text{males}} = 6.4\%$, vs. $M_{\text{females}} = 9.5\%$, $t = 2.38, p = .02$). Finally, through combining the rated stereotypically of a behavior and the presence or absence of an actual gender difference in our sample, we categorized the ACT behaviors as (a) gender-stereotypic with a kernel of truth (laughing, spending time on the computer, attending class), (b) gender-stereotypic with no kernel of truth (crying, spending time with others, talking one-on-one, going to coffee shops, socializing, outside, talking with opposite sex, working, watching television, listening to music), or (c) gender neutral (singing, commuting, talking in a group, spending time indoors). To test our predictions regarding the effect of gender stereotypes on accuracy we averaged the correlations for the different ACT behaviors within each of the three categories.

2.2. Results

2.2.1. How accurate were the judges in rating the targets' daily behavior?

To address our first research question examining the judges' accuracy when rating the targets' daily behavior from five sound bites of their natural conversations, we compared the judges' ACT ratings to the corresponding act-frequencies determined from the targets' EAR monitoring. Across all ACT behaviors, the correspondence between our composite of 8–10 judges' ratings and the behavioral accuracy criterion was $r = .25, p = .01$ (see Fig. 1). The level of accuracy for the average of a single judge, again across all ACT behaviors, was $r = .12, p > .05$. Thus, the data supported our predictions for the aggregated measure but the accuracy that individual judges achieved failed to meet conventional standards for statistical significance.

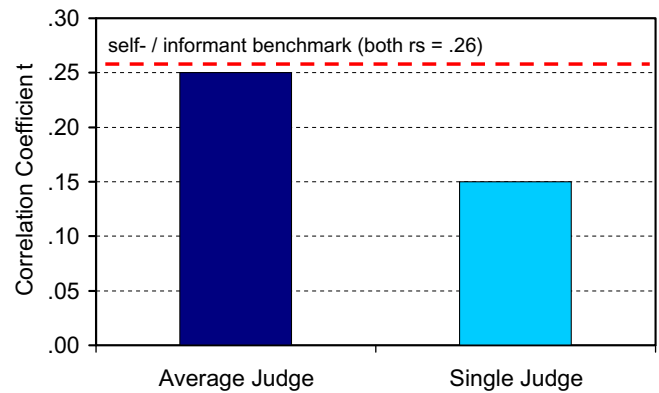


Fig. 1. Accuracy of judges' ratings of targets' daily behavior based on five 30-s snippets of natural conversations (aggregated across 20 ACT behaviors). *Note:* This figure displays the mean levels of judges' accuracy across all ACT behaviors; the level of accuracy for the targets' self-ratings and the informants' ratings are inserted as "benchmark line" (data from Vazire & Mehl, 2008); average judge = accuracy was computed based on a composite measure of eight judges; single judge = accuracy was computed for each judge individually and averaged across all judges.

Because interpretations of zero-acquaintance accuracy correlations of this or similar magnitudes can range between "trivially small" and "surprisingly large", we compared judges' accuracy against two benchmarks: the accuracy that targets' themselves achieved in predicting their own behavior and the accuracy that targets' informants achieved in predicting the targets' behavior. We predicted that judges' accuracy would be substantial but—given the limited available information—still lower than the targets' and the informants' level of accuracy.

The targets' accuracy averaged $r = .26, p = .01$ and the informants' accuracy (computed as single-informant accuracy) was $r = .26, p = .01$ (dashed line in Fig. 1). Interestingly, although—as predicted—this level of accuracy is indeed considerably higher than the accuracy that individual judges achieved ($r = .12$), it is highly comparable to the accuracy for the composite measure of eight judges' ratings ($r = .25$). In essence then, eight strangers that eavesdropped on five 30-s snippets of targets' natural conversations predicted the targets' daily behavior as well as the targets themselves and informants that knew the target well.

2.2.2. How were gender stereotypes related to the accuracy of judges' ratings of targets' daily behavior?

To address our second research question examining how the use of gender stereotypes affects accuracy, we compared the correspondence between the judges' ratings and targets' EAR-assessed behavior for the three ACT clusters, gender-stereotypic behaviors with a kernel of truth, gender-stereotypic behaviors with no kernel of truth, and gender-neutral behaviors. Table 2 presents the means and standard deviations of the average judges' ratings separately for male and female targets. Judges' ratings showed a significant gender difference for 9 of the 16 ACT behaviors that were classified as gender-stereotypic and for 0 of the 4 ACT behaviors that were classified as gender-neutral. This suggests that judges tended to use gender as a cue when rating targets on gender-stereotypic behaviors but not when rating them on gender-neutral behaviors.

As shown in Table 3, judges achieved greater accuracy for gender-stereotypic ACT behaviors with a kernel of truth ($r = .39$) than for those with no kernel of truth ($r = .23$). Hotelling's t -tests (with Williams Modification) indicated that the difference between the two correlations was marginally significant, $p = .06$. Judges' accuracy for gender-neutral ACT behaviors was $r = .31$ and not statistically different from the correlations for gender-stereotypic

Table 2

Average judges' ratings for male and female targets for gender-stereotypic behaviors with and without kernel of truth and gender-neutral behaviors (Study 1).

ACT item	Male targets M (SD)	Female targets M (SD)	t-test t
<i>Gender-stereotypic behaviors with kernel of truth</i>			
Laughing (f)	4.16 (0.55)	4.46 (0.99)	-2.89*
On the computer (m)	4.89 (0.64)	4.34 (0.71)	4.06*
Attending class (f)	4.16 (0.74)	3.96 (0.67)	0.75
Average	4.40 (0.65)	4.25 (0.80)	-
<i>Gender-stereotypic behaviors with no kernel of truth</i>			
Listening to music (m)	4.51 (0.62)	4.77 (0.69)	-1.79
Talking on the phone (f)	3.80 (0.78)	4.97 (0.89)	-6.57*
At work (m)	4.16 (0.75)	3.61 (0.73)	1.78
Talking to opposite sex (m)	4.00 (0.71)	4.61 (0.87)	-3.36*
Watching TV (m)	4.61 (0.44)	4.56 (0.59)	0.35
Talking to same sex (f)	4.73 (0.57)	5.03 (0.58)	-2.44*
At a restaurant/coffee shop (f)	4.01 (0.58)	4.50 (0.83)	-3.19*
With other people (f)	4.33 (0.59)	4.82 (0.79)	-2.65*
Socializing (f)	4.44 (0.63)	4.50 (0.81)	-0.21
Talking one-on-one (f)	4.24 (0.52)	4.73 (0.43)	-4.06*
Outdoors (m)	3.72 (0.47)	3.69 (0.53)	-0.59
Crying (f)	2.99 (0.70)	3.93 (0.52)	-6.96*
Arguing (m)	3.61 (0.71)	3.96 (0.67)	-0.13
Average	4.09 (0.63)	4.43 (0.70)	-
<i>Gender-neutral behaviors</i>			
Singing	3.42 (0.82)	3.79 (0.86)	-1.64
Commuting	3.98 (0.40)	4.04 (0.44)	-0.76
Talking in a group	4.29 (0.66)	4.69 (0.85)	-1.84
Indoors	4.62 (0.44)	4.55 (0.60)	0.39
Average	4.08 (0.60)	4.27 (0.71)	-

Note: N = 78; ratings are based on a 1–7 scale; f = rated as stereotypically female; m = rated as stereotypically male.

* $p < .05$, two-tailed, independent sample t-test, $df = 76$.

Table 3

Accuracy of judges' ratings of targets' daily behavior: accuracy correlations for gender-stereotypic behaviors with and without kernel of truth and gender-neutral behaviors.

ACT item	Average judge accuracy	Single judge accuracy
<i>Gender-stereotypic behaviors with kernel of truth</i>		
Laughing (f)	.56*	.32*
On the computer (m)	.51*	.27*
Attending class (f)	.05	.03
Average	.39	.21
<i>Gender-stereotypic behaviors with no kernel of truth</i>		
Listening to music (m)	.46*	.25*
Talking on the phone (f)	.45*	.23*
At work (m)	.42*	.22*
Talking to opposite sex (m)	.42*	.24*
Watching TV (m)	.38*	.20*
Talking to same sex (f)	.27*	.12
At a restaurant/coffee shop (f)	.22*	.13
With other people (f)	.21*	.10
Socializing (f)	.14	.08
Talking one-on-one (f)	.10	.05
Outdoors (m)	-.03	.08
Crying (f)	-.04	-.01
Arguing (m)	-.12	-.06
Average	.23	.13
<i>Gender-neutral behaviors</i>		
Singing	.44*	.24*
Commuting	.39*	.15
Talking in a group	.31*	.16
Indoors	.08	.03
Average	.31	.15

Note: N = 78; accuracy correlations represent the correspondence between judges' ratings of targets' daily behaviors assessed on the ACT questionnaire and targets' actual daily behaviors assessed from 4 days of EAR behavioral monitoring.

* $p < .05$; f = rated as stereotypically female, m = rated as stereotypically male; average judge = accuracy was computed based on a composite measure of eight judges; single judge = accuracy was computed for each judge individually and averaged across all judges.

behaviors with and without a kernel of truth. Consistent with our predictions, then, accuracy was descriptively highest for gender-stereotypic behaviors with a kernel of truth, followed by gender-neutral behaviors and by gender-stereotypic behaviors with no kernel of truth.

2.3. Discussion

Study 1 tested how accurate naïve judges are when rating the daily behavior of unknown targets based on five 30-s sound bites of their natural conversations. The accuracy that the judges achieved was statistically significant for the composite of eight judges' ratings but failed to meet the traditional significance threshold for the average of each individual judge. Our comparison with two benchmarks, the accuracy that the targets' themselves and the targets' informants achieved, revealed that the judges' accuracy was substantial: the aggregate of eight judges predicted the targets' behavior at the same level as the targets were able to predict their own behavior and informants – who were selected because they knew the target well – were able to predict the targets' behavior. Each individual judge by her-/himself, on the other hand, was on average about half as accurate as the targets and the informants (and non-significantly so).

Overall, these findings fit well into the picture that prior thin-slice (Ambady, Krabbenhoft, & Hogan, 2006; Carney, Colvin, & Hall, 2007) and zero-acquaintance research (Borkenau & Liebler, 1992; Levesque & Kenny, 1993) has painted. They extend this research, though, by demonstrating empirically—relatively to two meaningful benchmarks—exactly how accurate thin-slice judgments are and are not in the context of predicting the daily behavior of strangers.

Our second goal was to test the role of stereotypes in the accuracy of zero-acquaintance judgments. Features of our study design allowed us to submit this research question to a direct empirical test. First, the auditory EAR recordings of the targets' natural conversations limited the categorical information that judges had available about the targets effectively to their sex. That way, the only stereotype that judges could base their ratings on, were stereotypes about gender differences in behavior. Second, the use of a behavioral accuracy criterion that is independent of self- or other perceptions (i.e. self- or informant reports) allowed us to classify the assessed daily behaviors either as gender-stereotypic with or without a kernel of truth or as gender neutral.

We found that judges achieved higher accuracy for gender-stereotypic behaviors that contained a kernel of truth than for gender-stereotypic behaviors with no kernel of truth with their accuracy for gender-neutral behaviors falling in-between. This finding is consistent with theoretical person perception models (2004; Kenny 1994) that predict that with little or no individuating behavioral information being available, first impressions tend to be based on salient categorical information, or—in other words—stereotypes. The models further assume that accuracy is facilitated if these stereotypes contain a kernel of truth and undermined to the extent that they do not. Together, the results of our analyses provide strong correlational—though not direct causal—evidence that stereotypes influence zero-acquaintance accuracy.

Theoretically, though, person perception models make clear predictions that the influence of stereotypes on accuracy is only strong in the absence of individuating information and should fade out as such information becomes available (Judd & Park, 1993; Kenny, 1994, 2004). In Study 2, we experimentally manipulated the amount of information available to judges to test the role of gender stereotypes on judgmental accuracy at varying levels of information.

3. Study 2: gender stereotypes and the accuracy of zero-acquaintance judgments at increasing amounts of available information

One theoretical model that links information quantity to judgmental accuracy is the Weighted Average Model or WAM (Kenny, 1994). Recently, the WAM was reformulated as the PERSON model of interpersonal perception (Kenny, 2004). The WAM and the PERSON model make identical predictions, but Kenny (2004) prefers the PERSON model because it is better tied to psychological theory and allows easier hypothesis generation and interpretation of results.

Theoretically, the PERSON model predicts increased accuracy with increasing information (with an asymptotic function), but empirically, the evidence is somewhat mixed. Ambady and Rosenthal's (1992) meta-analysis found that judgments based off a 30-s behavioral stream (i.e. thin slices) were not significantly different from judgments based off a 5-min behavioral stream. Ambady, Bernieri, and Richeson (2000) concluded that more information did not increase accuracy arguing that once first impressions are formed they are not very malleable and stylistic variables that contribute to accurate judgments are contained even in very short segments of behavior. Similarly, Ballew and Todorov (2007) found that longer exposure to faces of candidates of governmental races did not increase the accuracy with which raters predicted the right election outcomes (i.e. the candidates who ultimately won).

Other evidence suggests that the amount of available information does influence accuracy. In a longitudinal study testing the effect of acquaintanceship on accuracy, students in a seminar rated each other's personality repeatedly over the course of the semester (Paulhus & Bruce, 1992). Across the Big Five personality traits, the level of accuracy, operationalized as self-other agreement, gradually increased over 7 weeks. Similarly, Borkenau and colleagues (2004) found evidence that information quantity and accuracy are positively related. Participants in their study engaged in a variety of "personality-revealing" tasks such as telling a joke, engaging in conversations with strangers, and solving complex problems. They found that the accuracy of ratings of personality traits increased as raters viewed larger numbers of behavioral acts – with accuracy reaching an asymptotic value after six behavioral acts. Theoretically, these findings converge with the PERSON model which predicts decreasing accuracy slopes with increasing acquaintance. In giving judges varying amounts of information about the targets (from 1 to 10 sound files or 30 s to 5 min) and assessing accuracy with the same behavioral criterion as in Study 1, we submitted the PERSON model prediction to another empirical test.

3.1. The Influence of stereotypes on judgmental accuracy at varying amounts of information

The PERSON model predicts that variance due to stereotypes will dominate initial impressions. With increasing information, however, variance due to behavior or individuating information (i.e. information about what the target is actually like) increases. One widely held assumption is that relying on stereotypes undermines accuracy. But, as previously discussed, stereotypic behaviors vary in the extent to which they do or do not contain kernels of truth. Thus, different predictions can be made about how increasing amounts of information influence accuracy for stereotypic behaviors with and without kernel of truth.

For stereotypic behaviors with a kernel of truth, we expect that accuracy will be substantial even with minimal information—resulting from stereotype accuracy—and increase only modestly

with additional information. For stereotypic behaviors with no kernel of truth, we expect that accuracy will initially be close to zero—resulting from the absence of stereotype accuracy and the absence of individuating, or in PERSON terms "personality", information—and increase rapidly with additional information. Finally, for behaviors that do not have a salient stereotype associated with them, we predict that accuracy will follow the same pattern of increasing accuracy with greater information as predicted theoretically by the person model and shown empirically by past research (Borkenau et al., 2004; Paulhus & Bruce, 1992). Specifically, we predict that the accuracy trajectory for non-stereotypic behaviors will fall in between the trajectory for stereotypic behaviors with and without kernel of truth. In other words, we predict that the use of stereotypes for stereotypic behaviors with no kernel of truth will—in relative terms—hamper accuracy early in the person perception process (Fiske & Neuberg, 1990).

3.2. Study design and hypotheses

To test these predictions, we experimentally varied the amount of information available to judges. Targets were randomly assigned to one of six conditions: listening to one, two, three, four, five, or ten EAR sound files. After listening to the assigned number of conversations, judges' rated the targets' behavior using the ACT questionnaire. We derived two sets of hypotheses for how increases in the available information affect (1) judges' overall accuracy and (2) judges' accuracy for gender-stereotypic and gender-neutral behaviors.

Specifically, based on the Study 1 findings, we hypothesized that judges' overall accuracy would (1a) increase—with decreasing slopes—from the one to the ten sound file condition (1b) essentially reach a plateau in the five sound file condition, and (1c) ultimately, that is in the 10 sound file condition, be comparable to the accuracy that the targets and their informants achieved. With respect to the effects of gender stereotypes on accuracy trajectories, we hypothesized that judges' accuracy (2a) when rating gender-stereotypic behaviors with a kernel of truth would already be substantial in the one-sound file condition, and show little further increase from the two to the ten sound file condition; (2b) when rating gender-stereotypic behaviors without a kernel of truth would be close to zero in the one-sound file condition and gradually increase from the two to the ten sound file condition; (2c) when rating gender-neutral behaviors would be minimal in the one-sound file condition and gradually increase from the two to the ten sound file condition. Because the use of stereotypes in the absence of a kernel-of-truth should undermine accuracy, we predicted that the trajectory for gender-neutral behaviors would start out above the trajectory for gender-stereotypic behaviors without a kernel of truth but quickly converge with it as individuating information becomes available.

3.3. Method

3.3.1. Participants

317. University of Arizona undergraduate students (different from those in Study 1; 55% female; 72% White, 2% African American, 8% Asian, 10% Hispanic, 1% Native American, and 5% of another ethnicity) served as naïve judges for the study. Their age ranged from 18 to 60 years ($M = 19.2$, $SD = 3.1$). None of the judges knew any of the targets they rated.

3.3.2. Stimulus materials for the thin-slice ratings

The same set of targets from Study 1 was used. This time, for each of the 78 targets, 10 conversations were randomly selected from their full set of EAR sound files using a random number generator. A conversation was chosen if it met the criteria of

containing at least five words uttered by the target. The sound files in each condition (e.g. the five or four sound file condition) were randomly selected from the sound files in the preceding condition (e.g. the ten or five sound file condition) so that with respect to the contained information lower order conditions were proper subsets of higher order conditions. To control for the effect of rater experience on accuracy, the order in which the targets were rated was again counterbalanced.

3.3.3. Judges' ratings of targets' daily behavior

Judges were told the study was about forming first impressions of other people. Judges were randomly assigned to a one, two, three, four, five, or ten sound file condition. They were instructed to listen to all of a target's sound files prior to completing a set of questionnaires about their first impression of the target. Judges rated each target on the ACT questionnaire using a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). Judges were divided into blocks of 4–5 so that each judge rated 6 targets. That way, 72 targets were rated by 4 judges and 6 targets by 5 judges. The mean inter-judge agreement across all ACT items for the one, two, three, four, five, and ten sound file condition was .34, .42, .35, .39, .24 and .39, respectively, calculated as intraclass correlations based on one-way random effect models, $(ICC[1, k])$.

3.4. Results

3.4.1. How did judges' overall accuracy change with increasing amounts of information?

Parallel to our Study 1 data analytic strategy, we computed accuracy correlations for the average judge (i.e. composite of 4–5 judges) and for a single judge to be able to compare the judges' accuracy to the accuracy that the targets' themselves and their informants achieved. Also, identical to Study 1, accuracy was again computed by correlating the judges' ratings of each ACT behavior with the EAR-assessed behavior frequencies. Table 4 presents the two sets of accuracy correlations separately for all ACT behavior and all experimental conditions.

As shown in Fig. 2, consistent with Hypothesis 1a, judges' overall accuracy increased with decreasing slopes from the least to the most information condition. Across all behaviors, the accuracy for the composite measure of the average judge was .08 in the one, .13 in the two, .20 in the three, .19 in the four, .23 in the five, and .25 in the ten sound file condition. Hotelling's *t*-tests (with Williams Modification) indicated that the accuracy correlation in the one-sound file condition was marginally different from the accuracy correlation in the 10 sound file condition, $p = .06$. Consistent with Hypothesis 1b, judges' accuracy essentially reached a plateau in the 5 sound file condition. Fig. 2 also shows the accuracy trajectory for the single judge measure. Across all ACT behaviors, the accuracy for the single judge was .04 in the one, .08 in the two, .12 in the three, .12 in the four, .14 in the five, and .16 in the 10 sound file condition. None of the correlations in this condition were statistically different from each other. Similar to Study 1 and as predicted by the Spearman–Brown prophecy formula, the single-judge correlations were weaker, but – as a mathematical necessity – the trajectory paralleled the average-judge pattern.

Fig. 2 again contains the self and informant benchmarks as a dashed line. Supporting our Hypothesis 1c, in the 10 sound file condition, the composite measure of the average-judge accuracy ($r = .25$) was virtually identical to the two accuracy benchmarks ($r_s = .26$). Again, the single-judge accuracy was lower ($r = .16$) and did not reach statistical significance. Together, this suggests although a single judge was not on par with the targets and the informants' accuracy after 5 min of conversational information, the composite of 4–5 judges was.

3.4.2. How were gender stereotypes related to changes in the judges' accuracy with increasing amounts of information?

To test our predictions regarding the influence of stereotypes on accuracy trajectories, we computed accuracy correlations in each condition for gender-stereotypic behaviors with a kernel of truth, gender-stereotypic behaviors with no kernel of truth, and gender-neutral behaviors. The classification of ACT behaviors into one of the three categories was identical to Study 1. Fig. 3 shows the accuracy trajectories for the three types of behaviors across the six experimental conditions. For the gender-stereotypic behaviors with a kernel of truth, judges' accuracy was .05 in the one, .25 in the two, .29 in the three, .27 in the four, .23 in the five, and .25 in the 10 sound file condition. Hotelling's *t*-tests (with Williams Modification) indicated that the accuracy correlation in the one-sound file condition was significantly different from the accuracy correlation in all the other conditions, $p < .05$. For the gender-stereotypic behaviors with no kernel of truth, judges' accuracy was .07 in the one, .10 in the two, .18 in the three, .18 in the four, .23 in the five, and .24 in the ten sound file condition. Hotelling's *t*-tests (with Williams Modification) indicated that the accuracy correlation in the one-sound file condition was marginally different from the accuracy correlation in the 10 sound file condition, $p = .06$. Finally, for the gender-neutral behaviors, it was .14 in the one, .15 in the two, .19 in the three, .16 in the four, .20 in the five, and .27 in the ten sound file condition. None of the accuracy correlations in any of the conditions were significantly different from each other, $p > .11$.

Our Hypothesis 2a received partial support. Although judges' accuracy when rating gender-stereotypic behaviors with a kernel of truth was substantial already in the two-sound file condition ($r = .25$) and did not reliably increase further with more information, accuracy in the one-sound file condition was unexpectedly minimal ($r = .05$). Hypothesis 2b, on the other hand, received good support. Judges' accuracy when rating gender-stereotypic behaviors with no kernel of truth was minimal in the one-sound file condition ($r = .07$) and gradually increased to a level comparable to the one for the gender-stereotypic behaviors with kernel of truth ($r = .24$). Finally, Hypothesis 2c also received good support with judges' accuracy when rating gender-neutral behaviors being initially low ($r = .10$) and gradually increasing to the level of the other two categories ($r = .27$). Also, as predicted, the trajectory for gender-neutral behaviors started out slightly higher than the trajectory for the gender-stereotypic behaviors with no kernel of truth and both had essentially converged in the three sound file condition.

3.5. Discussion

In Study 2, we manipulated the amount of available information and tested the effect of information quantity on judges' overall accuracy and their accuracy for rating gender-stereotypic behaviors with and without kernel of truth and gender-neutral behaviors. As predicted, we found that judges' overall accuracy increased with decreasing slopes as a function of information quantity. We further found that judges' overall accuracy had essentially reached a plateau with four EAR sound files or 2.5 min of conversational information about the targets and was equal to the accuracy that targets' themselves and the their informants achieved in their ratings.

With respect to the influence of gender-stereotypes on accuracy trajectories, we found that (a) one-sound file failed to yield accurate ratings for any of our three behavior categories, (b) two-sound files yielded substantially accurate ratings for gender-stereotypic behaviors with a kernel of truth but not for gender-stereotypic behaviors without kernel of truth and gender-neutral behaviors, (c) 10 sound files yielded substantially accurate ratings for all three

Table 4
Accuracy of the judges' ratings of daily behavior for increasing amounts of available information.

ACT item	One-sound file	Two-sound files	Three sound files	Four sound files	Five sound files	Ten sound files
<i>Gender-stereotypic behaviors with kernel of truth</i>						
Laughing (f)	.13 (.07)	.31* (.23*)	.33* (.22*)	.37* (.24*)	.37* (.22*)	.28* (.21*)
On the computer (m)	.18 (.10)	.40* (.24*)	.29* (.18)	.35* (.24*)	.39* (.22*)	.43* (.23*)
Attending class (f)	-.15 (-.07)	.05 (.02)	.25* (.17)	.08 (.05)	-.05 (-.03)	.04 (.03)
<i>Gender-stereotypic behaviors with no kernel of truth</i>						
Talking on the phone (f)	.20* (.11)	.38* (.25*)	.28* (.18)	.30* (.20*)	.40* (.25*)	.36* (.25*)
Listening to music (m)	.13 (.07)	.21* (.15)	.24* (.17)	.29* (.20*)	.28* (.20*)	.34* (.24*)
At work (m)	.10 (.06)	.25* (.15)	.31* (.20*)	.29* (.17)	.19* (.11)	.39* (.25*)
Talking to opposite sex (m)	.12 (.09)	.36* (.26*)	.13 (.10)	.30* (.20*)	.33* (.21*)	.44* (.29*)
Watching TV (m)	.05 (.03)	.10 (.06)	.28* (.18)	.30* (.18)	.38* (.27*)	.11 (.18)
Talking to same sex (f)	.02 (-.01)	.14 (.09)	.28* (.19*)	.24* (.17)	.32* (.18)	.48* (.34*)
At a restaurant/coffee shop (f)	.25* (.16)	-.12 (-.08)	.20* (.13)	.28* (.17)	.39* (.23*)	.23* (.13)
With other people (f)	.04 (-.01)	.01 (.03)	.09 (.02)	.21* (.10)	.33* (.21*)	.29 (.10)
Socializing (f)	.15 (.09)	.12 (.08)	.16 (.10)	.19* (.10)	.19* (.11)	.33* (.16)
Talking one-on-one (f)	-.11 (-.07)	-.09 (-.07)	-.11 (-.09)	.00 (.02)	.18* (.10)	.07 (.03)
Outdoors (m)	-.08 (-.04)	-.04 (-.04)	.16 (.11)	-.01 (.00)	.07 (.00)	.05 (.05)
Crying (f)	-.06 (-.07)	-.01 (-.03)	.09 (.01)	.02 (-.02)	-.03 (-.03)	.01 (-.06)
Arguing (m)	.13 (.11)	-.05 (.04)	.26* (.17)	-.09 (-.03)	-.05 (.01)	.00 (.03)
<i>Gender-neutral behaviors</i>						
Singing	.20 (.11)	-.11 (-.06)	.26* (.18)	.29* (.18)	.27* (.20*)	.40* (.28*)
Commuting	.00 (.00)	.20* (-.09)	.28* (-.05)	.14 (.10)	.09 (.04)	.31* (.14)
Talking in a group	.19* (.12)	.31* (.22*)	.20* (.11)	.19* (.11)	.35* (.18)	.24* (.15)
Indoors	.18* (.12)	.18* (.13)	.07 (.03)	.02 (.00)	.07 (.03)	.15* (.11)

Note: N = 78; accuracy is the degree of correspondence between the ACT behavior ratings and the targets' EAR-observed behavior; numbers indicate accuracy for a composite of 4–5 judges; numbers in parentheses indicate accuracy for a single judges.
* p < .05.

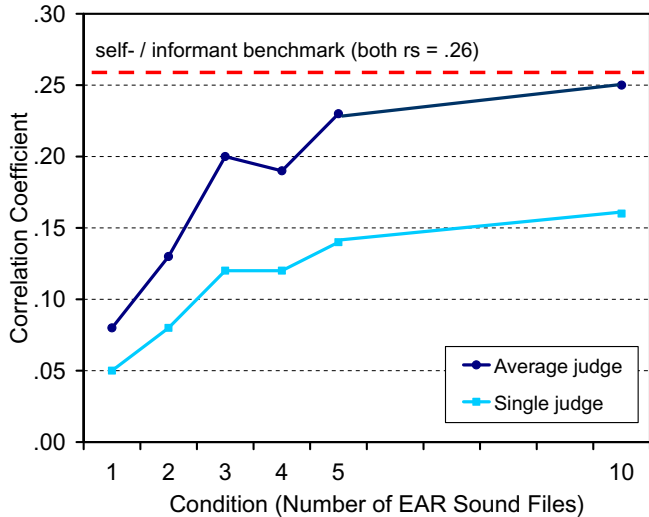


Fig. 2. Accuracy of judges' ratings of targets' daily behavior as a function of the amount of available information (aggregated across 20 ACT behaviors). Note: This figure displays the mean levels of judges' accuracy across all ACT behaviors for increasing levels of information; the level of accuracy for the targets' self-ratings and the informants' ratings are inserted as "benchmark line" (data from Vazire & Mehl, 2008); the dark blue line shows the accuracy for the composite of 4–5 judges, the light blue line for a single judge; accuracy is the degree of correspondence between the ACT behavior ratings and the targets' EAR-observed behavior. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

behavior categories, and (d) one and two-sound files yielded somewhat higher accuracy for gender-neutral behaviors than for gender-stereotypic behaviors without a kernel of truth – suggesting that the use of stereotypes in the absence of stereotype accuracy undermines accuracy.

The results regarding judges' overall accuracy are consistent with prior research showing a positive effect for information quantity on accuracy (Borkenau et al., 2004; Colvin & Funder,

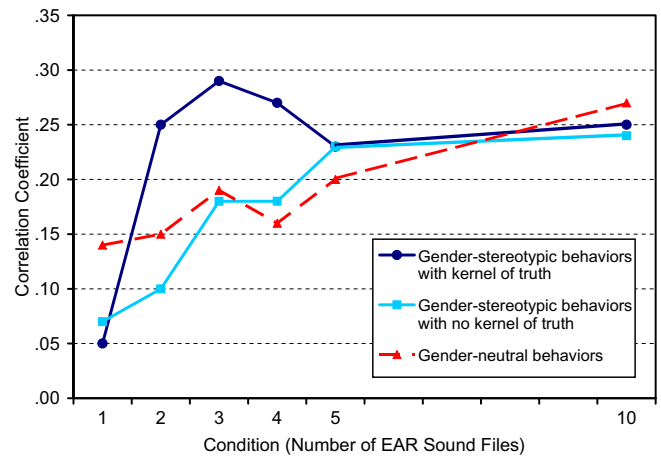


Fig. 3. Accuracy of judges' ratings of targets' daily behavior as a function of the amount of available information and the type of daily behavior. Note: This figure displays a comparison of the mean level of accuracy across ACT behaviors that are considered (a) gender-stereotypic (for either males or females) and have a kernel of truth, (b) gender-stereotypic and have no kernel of truth and (c) gender-neutral.

1991; Funder & Sneed, 1993). Borkenau and colleagues (2004) suggested that accuracy reaches its asymptote after six behavioral acts, but suggested that "this general rule should be clarified in future research" (p. 610). Although their findings look at the surface very similar to (and are certainly consistent with) our findings—we found asymptotic accuracy in the five sound file condition—it is important to consider that in their study, one act referred to one scenario (such as reading aloud or introducing oneself) and ranged in duration between 1 and 12 min. Our study, in contrast, operationalized one informational unit as a 30-s sound bite of a person's daily conversations—clearly less time on the stop-watch but maybe in some cases more information considering the variability in targets' natural behavior (Ickes, Snyder, & Garcia, 1997).

These are just two examples that illustrate how what is considered one “act” or one unit of information varies considerably across studies. Other studies have used such different act operationalizations as 20-min conversations (Paulhus & Bruce, 1992) or 5–30 min video recordings (Blackman & Funder, 1998). Interestingly, the PERSON model provides no clear specifications as to how much information constitutes one act. Instead, it states that “within PERSON, acts have a theoretical not an operational meaning, and PERSON does not specify the time that it takes to view an act. However, given most reasonable definitions of acts, the expectation is that if a perceiver were to observe a target for a few hours, 100 acts would be observed.” (Kenny, 2004, p. 274). Ultimately, the question of what defines a behavioral act is likely a philosophical one as much as it is an empirical one—though one that the field would benefit from if it could address it more systematically in future research.

Irrespective of how much behavior constitutes one psychological act, though, the finding that after only five 30-s clips, or 2½ min of information, four to five naïve judges rated strangers’ daily behavior with a degree of accuracy that is comparable to what the targets themselves and people who knew the targets well achieved, is in line with Kenny’s (2004) thought that it may be “time to rethink a fundamental assumption of person perception” and to replace the “peeling an onion” metaphor of how long it takes to know someone with a “scratching the surface of the onion” metaphor because “after all, the distinctive taste of an onion is as marked in its outer layer as it is in the innermost layer” (p. 277).

With regard to the effect of stereotypes on accuracy, we failed to find the differential pattern of accuracy for gender-stereotypic behaviors with and without a kernel of truth in the one-sound file condition (but found support for it in the two-sound file condition). Counter to our predictions, judges’ use of gender stereotypes did not lead to relatively accurate ratings with only one conversational snippet. Based on our experience conducting EAR research, we speculate that we failed to find the predicted pattern because judges may have had trouble identifying the target in the sound file. In all but the one-sound-file condition, judges could infer which of the recorded voices belonged to the target by looking for voice consistency across sound files (note that in all sound files the targets spoke at least five words). Even though the positioning of the microphone should in most cases ensure that the target’s voice is somewhat louder than the voices of the other captured parties, this was certainly not always the case (e.g. when soft-spoken targets talked to “loud” persons or when the microphone was put on the table instead of worn at the lapel). To the extent that judges could not unambiguously identify the target in the sound file, the use of categorical information was effectively rendered impossible. In the absence of categorical and individuating information to use, it is not surprising that the judges’ accuracy for both gender-stereotypic behaviors with and without a kernel of truth was practically zero.

4. General discussion

In everyday life, people routinely and intuitively form impressions of other on the basis of minimal information. Often such minimal information consists of randomly overhearing snippets of conversations. The purpose of this project was to (1) address three methodological challenges in zero-acquaintance research and (2) test the effect of stereotype use on the accuracy of person perceptions. Study 1 showed that the accuracy of a group of 8–10 judges achieve after listening to 2½ min sampled from the targets’ natural daily conversations was similar to the accuracy that the targets’ themselves and their informants achieved. Fur-

ther, judges’ accuracy was greater for gender-stereotypic behaviors with a kernel of truth than for those without a kernel of truth or for gender-neutral behaviors. In Study 2, we expanded the design and varied the amount of information judges had available. Consistent with predictions made by the PERSON model, overall accuracy increased as the judges received more information. We further found evidence for stereotypes facilitating accuracy only very early in the person perception process and only for gender-stereotypic behaviors that contain a kernel of truth.

4.1. Implications for research on the accuracy of zero-acquaintance judgments

This project expanded on prior zero-acquaintance research (Skowronski & Ambady, 2008) in several ways. First, instead of relying on self- and/or informant reported trait measures, we used unobtrusively observed daily behavior as accuracy criterion. In doing so, our studies responded to a call for using more behavioral measures in zero-acquaintance research (Kenny & West, 2008) and more generally in social and personality psychology (Baumeister et al., 2007; Furr, in press). Second, with researchers often interpreting accuracy coefficients vaguely as “relatively large”, “considerable”, or “substantial”, we measured judges’ accuracy against two theoretically and practically important benchmarks—the accuracy that targets achieved in rating their own behavior and the accuracy that peers who knew the targets well achieved in rating the targets’ behavior. Our comparisons with these reference groups make the results of the current studies compelling by showing that the judges’ joint accuracy was up to par with the knowledge that the targets had about their own daily behavior and friends had about the targets’ daily behavior. Third, compared to studies that base their findings on a single person perception context or behavioral sequence, the EAR method allowed us to take Brunswik’s (1956) notion of a representative design serious and sample both participants from an underlying population of targets and situations from an underlying ecology of contexts.

These studies further extend prior research by speaking to the bandwidth-fidelity dilemma (Cronbach & Gleser, 1965). Most zero-acquaintance studies have estimated the accuracy of personality judgments using relatively low-bandwidth (or high-fidelity, that is specific) person perception contexts (e.g. a picture or a short interaction) and a relatively broad bandwidth personality criterion (i.e. the Big Five domain). Here, in contrast, we used a relatively high-bandwidth person perception context (i.e. conversational snippets sampled representatively from the full spectrum of targets’ daily interactions) and a relatively high-fidelity personality assessment (i.e. a limited set of specific daily behaviors). In terms of the classic bandwidth-fidelity problem, both scenarios potentially “sacrifice” predictive validity through a “mismatch” between the level of assessment and the level of prediction. Interestingly, though, the two approaches seem to result in comparable outcomes, that is similar accuracy estimates. Future research should further investigate potential bandwidth-fidelity trade-offs in zero-acquaintance research by systematically varying levels of assessment and prediction.

4.2. Implications for research on the effect of stereotypes on zero-acquaintance accuracy

Our design allowed us to investigate an issue in zero-acquaintance research that has received considerable theoretical but limited empirical attention—the role of stereotypes in facilitating and undermining accurate person perceptions. Kenny and West (2008) identified this as an understudied topic and suggested that “the literature would benefit by a more controlled and focused

study of stereotypes" (p. 143). Our solution to assess accuracy using directly observed act frequencies (Buss & Craik, 1983) allowed us to empirically compare perceived to actual gender differences. That way, we could categorized behaviors as either gender-stereotypic with a kernel of truth, gender-stereotypic with no kernel of truth or gender neutral and compute accuracy separately for each type of behavior. Further, using the EAR sounds as person perception stimuli limited the use of stereotypes to gender as the only available categorical information about the targets (Borkenau & Liebler, 1992). Among other potential stereotypes, gender stereotypes are of particular theoretical interest because gender information is one of the most salient cues shaping first impressions (Fiske et al., 1991).

Our findings are broadly consistent with personality research and theorizing on the effects of stereotypes on person perception. Most directly they converge with predictions made by the PERSON model (Kenny, 2004) which suggests that categorical information dominates impressions when perceivers have very little information about a person but exert little or no influence when perceivers have considerable individuating ("personality") information to base their impressions on. Our findings were further consistent with social psychological research showing that stereotypes exert their influence predominantly when perceivers have little information about targets and that their use decreases as more behavioral information becomes available (e.g. Fiske & Neuberg, 1990; Krueger & Rothbart, 1988; Kunda & Thagard, 1996).

Research in the latter tradition generally focuses on the negative effects of stereotyping, that is, judgmental errors that result from applying stereotypes and our findings support this in cases where stereotypes do not contain a kernel of truth. Judges' accuracy for gender-stereotypic behaviors with no kernel of truth was about half as large in the one-sound file and two thirds as large in the two-sound file condition compared to their accuracy for gender-neutral behaviors. Consistent with prior research on stereotype accuracy (Lee et al., 1995), though, our findings also show that in cases where stereotypes do contain a kernel of truth, they can facilitate accurate first impressions—at least in the first 2 min of the process. Judges' accuracy for gender-stereotypic behaviors with kernel of truth was about two thirds larger in the two-sound file, almost twice as large in the three sound file, and 42% larger in the four sound file condition compared to their accuracy for gender-neutral behaviors.

Despite our finding that judges had this initial advantage relying on partially valid gender stereotypes, we obviously caution against arguing that using stereotypes to form first impressions is generally recommendable. In the real-world, the level of acquaintance that corresponds to our two to four sound file condition is quickly reached. According to our data, after that there is not much to gain from such an approach. A closer look at the three trajectories reveals that whereas judges' accuracy for gender-neutral behaviors was still on an upward trend even in the 10 sound file condition, the other two trajectories had already reached their asymptotes well before.

In this context it is also important that our classification of behavior into the three types yielded only three behaviors that had an empirical kernel of truth in comparison to 13 behaviors that did not. Therefore, practically, the potential overall accuracy gain from relying on partially valid stereotypes early in the person perception process may well be small compared to the potential overall accuracy loss that comes with relying on—the larger number—of invalid stereotypes. Interestingly also, the results of our kernel-of-truth analysis were not self-evident. The three behaviors that had a kernel of truth in our sample (laughing, spending time on the computer, attending class) were not those that had the strongest gender-stereotype associated with them; and the two behaviors that were perceived as most gender-stereotypic both emerged as hav-

ing no kernel of truth (talking on the phone, crying; cf. Mehl, Vazire, Ramirez-Esparza, Slatcher, & Pennebaker, 2007).

4.3. Limitations and future directions

The two studies have several potential limitations. One concern about the paradigm is that the behavioral information contained in the EAR sound files is very similar to the behavior we used as accuracy criterion. This could have led to artificially inflated accuracy estimates. We addressed this issue by removing the sound files that the judges listened to from the data from which we derived the accuracy criteria. However, there still remains a high degree of similarity between the behavior captured in the sound files and the behavior we asked the judges to rate. To this effect, prior research has noted this boundary effect on accuracy (Colvin & Funder, 1991). This research has found that stranger ratings of behavior can be as accurate as ratings by close acquaintances when the task involves predicting behavior in a context that is very similar to the one from which the ratings are derived. Importantly, though, the "context" in our study was a representative sample taken from the full spectrum of targets' daily conversations and therefore much less circumscribed than the contexts in usual, laboratory-based zero-acquaintance research.

Another limitation revolves around judges' use of the ACT questionnaire to rate targets' daily behavior. The ACT questionnaire assesses behavior at a highly specific, molecular level (e.g. spending time commuting or indoors). It is easy to conceive how in some cases behaviors at a more molar, psychological level would be a more natural choice for person perception studies (e.g. "easily falls in love", "is dedicated to social justice"; Funder, Furr, & Colvin, 2000; Vallacher & Wegner, 1987). Nevertheless, several of the ACT behaviors do refer to information that is molecular yet inherently important for getting to know a person (e.g. time spent alone vs. with others, watching TV, listening to music, on the phone). By design, the ACT behaviors are constrained by what can be detected (and therefore coded) from ambient sounds. Another concern is the degree to which the EAR-assessed behaviors are susceptible to "seasonal" effects, that is situational influences on people's behavior during the days of monitoring. Possibly, we oversampled behavior associated with specific events, such as studying during finals. Future studies should use progress in mobile computing technologies (Goodwin, Velicer, & Intille, 2008) to go beyond our combined EAR–ACT approach to study a broader spectrum of daily behaviors at different levels of psychological granularity and with a longer-term monitoring.

Along the same lines, the combined EAR–ACT approach also limited the scope of our gender-stereotype analyses. Our approach with respect to classifying behaviors as gender-stereotypic or gender-neutral was based on stereotypicality ratings of the ACT items. Juxtaposing ACT behaviors that were above the median in either male or female stereotypically with actual gender differences on the corresponding EAR-coded behaviors in our sample yielded our classification of gender-stereotypic behaviors with and without a kernel of truth. Although this is the first study to directly determine stereotype accuracy on the basis of primary study data (cf. Swim, 1994), our strategy did have clear limitations.

Specifically, the use of a simple median split to determine what counts as a stereotypic behavior and a simple significance tests to determine what counts as an actual gender difference neglects that both gender-stereotypically and kernel of truth are really continuous rather than dichotomous constructs. Also, with rating the existing ACT items for stereotypicality we adopted a bottom-up, empirical rather than a top-down, theoretical approach. An alternative strategy—which may be difficult to implement with the current EAR approach but which future research should explore further—would be to a priori select behaviors on the grounds of

being highly gender-stereotypic. It is likely that the restricted range of daily behaviors we could study constrained our estimates of the effects of stereotype use on accuracy and that a study using a wider range of gender-stereotypic behaviors would yield “sharpened” accuracy trajectories.

Including a broader range of gender-stereotypic behaviors would also increase the chances of being able to test how stereotypes with—what could be called—a negative kernel of truth affect judgmental accuracy. Kenny and West (2008) speculated “it would seem possible that some stereotypes actually reflect the opposite of reality. For instance, in judging how talkative someone would be in a group, people might use the stereotype that women are more talkative than men. Yet, the data show that when gender-neutral topics are discussed in groups men actually tend to talk much more than women” (p. 142). Although in our analysis talking in groups failed to emerge as gender-stereotypic (and failed to show an actual gender difference), our study design would be suited to test the consequences of such a special constellation.

Another potential limitation of our sample-based approach to determining stereotype accuracy comes from the fact that we only had 4 days worth of criterion EAR data. As a consequence, two low base rates behaviors, crying and arguing, showed relatively little variability (see Table 1) which may have led to them being artificially and invalidly classified as gender-stereotypic with no kernel of truth. Because restricted variability in the criterion can constrain the accuracy correlations, this effect could in part be responsible for judges’ lower accuracy for gender-stereotypic behaviors without relative to with kernel-of-truth. Interestingly, though, whereas the actual gender difference for crying was indeed in direction of the gender stereotype, the actual gender difference for arguing went in the opposite direction. Further, a comparison of the variability in the gender-stereotypic with and without kernel-of-truth ACT clusters showed no evidence of restricted variability in the latter one (as would be expected if range restriction shifted several behaviors into the no-kernel-of-truth cluster). Therefore, the methodological effect of restricted variance constraining judgmental accuracy is likely limited to crying and arguing—with crying being the only behavior for which it may have worked in direction of our predictions.

Finally, yet another limitation is that our studies focused exclusively on gender stereotypes. As mentioned before, although an important research area, our decision to focus on gender as compared to other stereotypes was largely methodologically motivated. To test the generalizability of our findings, it would be valuable if future research replicated our findings using for example existing aging stereotypes (Golub & Langer, 2007; Rodin & Langer, 1980). Given that in a more heterogeneous sample, age cues could likely be readily and accurately decoded from the targets’ voice, the current paradigm could potentially be adapted to such a test.

4.4. Conclusions

In two studies we have shown that on the basis of a hand full of overheard conversational snippets, a group of people can make ratings of the behaviors of unacquainted targets that are as accurate as the ratings that the targets themselves and their good friends make. Whereas initially the judges’ accuracy stems in part from relying on partially valid stereotypes, relying on stereotypes with no kernel of truth at the same time undermines accuracy. As people obtain more information about a person, the influence of stereotypes on accuracy fades out quickly.

Elsewhere, it has been argued that as a field social psychology has a tendency to focus on perceptual errors whereas personality psychology has a tendency to focus on perceptual accuracy (Jussim, 2005; Krueger & Funder, 2004). It has further been argued that the two lines of research subscribe to different paradigms that are difficult to merge. Our research attempted to combine both perspectives by

showing within the same study design how stereotypes can both hinder and help accuracy. It is our hope that way, this research has the potential of bringing the two sub-disciplines one step closer together in an area that has historically been at the heart of both fields.

Acknowledgments

We thank our research assistants for their help with collecting the data. We also thank Jeff Greenberg and Stephanie Fryberg for their suggestions with the study design and data analysis and Simone Vazire for her feedback on drafts of this paper.

References

- Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, 32, 201–271.
- Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-sec sale: Using thin slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology*, 16, 4–13.
- Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 256–274.
- Back, M. D., Schmukle, S. C., & Egloff, B. (2008). How extraverted is honey <honey.bunny77@hotmail.de?>. Inferring personality traits from email addresses. *Journal of Research in Personality*, 42, 1116–1122.
- Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, 104, 17948–17953.
- 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.
- Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology*, 34, 164–181.
- Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero-acquaintance. *Journal of Personality and Social Psychology*, 62, 645–657.
- Borkenau, P., Mauer, N., Rieman, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, 86, 599–614.
- Brunswik, E. (1956). *Perception and representative design of psychological experiments*. Berkeley: University of California Press.
- Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review*, 90, 105–126.
- Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41, 1054–1072.
- Chaplin, W. F., Phillips, J. B., Brown, J. D., Clanton, N. R., & Stein, J. L. (2000). Handshaking, gender, personality, and first impressions. *Journal of Personality and Social Psychology*, 79, 110–117.
- Colvin, R. C., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology*, 60, 884–894.
- Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.
- Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 37, 1097–1126.
- Fiske, A. P., Haslam, N., & Fiske, S. T. (1991). Confusing one person with another – What errors reveal about the elementary forms of social relations. *Journal of Personality and Social Psychology*, 60, 656–674.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum model of impression formation, from category based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in experimental social psychology*, 23, 1–74.
- Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102, 652–670.
- Funder, D. C., Furr, R. M., & Colvin, C. R. (2000). The Riverside behavioral q-sort: A tool for the description of social behavior. *Journal of Personality*, 68, 451–489.
- Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, 64, 479–490.
- Funder, D. C., & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality*, 61, 457–476.
- Furr, R. M. (in press). Personality psychology as a truly behavioral science. *European Journal of Personality*.
- Gill, A. J., Oberlander, J., & Austin, E. (2006). Rating e-mail personality at zero-acquaintance. *Personality and Individual Differences*, 40, 497–507.
- Golub, S. A., & Langer, E. J. (2007). Challenging assumptions about adult development: Implications for the health of older adults. In C. M. Aldwin, C. L. Park, & A. Spiro (Eds.), *Handbook of health psychology and aging*. New York: Guilford.

- Goodwin, M. S., Velicer, W. F., & Intille, S. S. (2008). Telemetric monitoring in the behavior sciences. *Behavior Research Methods*, 40, 328–341.
- Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, 82, 379–398.
- Hall, J. A., & Carter, J. D. (1999). Gender-stereotype accuracy as an individual difference. *Journal of Personality and Social Psychology*, 77, 350–359.
- Holleran, S. E., & Mehl, M. R. (2008). Let me read your mind: Personality judgments based on a person's natural stream of thought. *Journal of Research in Personality*, 42, 747–754.
- Ickes, W., Snyder, M., & Garcia, S. (1997). Personality influences on the choice of situations. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 165–195). San Diego, CA: Academic Press.
- Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, 100, 109–128.
- Jussim, L. (2005). Accuracy: Criticisms, controversies, criteria, components, and cognitive processes. *Advances in Experimental Social Psychology*, 37, 1–93.
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford Press.
- Kenny, D. A. (2004). Person: A general model of interpersonal perception. *Personality and Social Psychology Review*, 8, 265–280.
- Kenny, D. A., & West, T. (2008). Zero-acquaintance: Definitions, statistical models, findings, and process. In J. Skowronski & N. Ambady (Eds.), *First impressions* (pp. 129–146). New York: Guilford.
- 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.
- Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology. Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Brain and Behavioral Sciences*, 27, 313–327.
- Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, 55, 187–195.
- Kruglanski, A. W. (1989). The psychology of being “right”: The problem of accuracy in social perception and cognition. *Psychological Bulletin*, 106, 395–409.
- Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel constraint satisfaction measure. *Psychological Review*, 103, 284–308.
- Lee, Y. T., Jussim, L., & McCauley, C. R. (Eds.). (1995). *Stereotype accuracy: Toward appreciating group differences*. Washington, DC: American Psychological Association.
- Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quality and quantity affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology*, 91, 111–123.
- Levesque, M. J., & Kenny, D. A. (1993). Accuracy of behavioral predictions at zero-acquaintance: A social relations analysis. *Journal of Personality and Social Psychology*, 65, 1178–1187.
- Marcus, B., Machilek, F., & Schütz, A. (2006). Personality in cyberspace: Personal web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology*, 90, 1014–1031.
- Mehl, M. R. (2006). The lay assessment of subclinical depression in daily life. *Psychological Assessment*, 18, 340–345.
- 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.
- Mehl, M. R., Pennebaker, J. W., Crow, M., Dabbs, J., & Price, J. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, and Computers*, 33, 517–523.
- Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcler, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317, 82.
- Moskowitz, D. S. (1982). Coherence and cross-situational generality in personality: A new analysis of old problems. *Journal of Personality and Social Psychology*, 43, 754–768.
- Paulhus, D. L., & Bruce, M. N. (1992). The effect of acquaintanceship on the validity of personality impressions: A longitudinal study. *Journal of Personality and Social Psychology*, 63, 816–824.
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology*. New York: Guilford.
- Rentfrow, P. J., & Gosling, S. D. (2006). Message in a Ballad: The role of music preferences in interpersonal perception. *Psychological Science*, 17, 236–242.
- Rodin, J., & Langer, E. J. (1980). Aging labels: The decline of control and fall of self-esteem. *Journal of Social Issues*, 36, 12–29.
- Skowronski, K., & Ambady, N. (Eds.). (2008). *First impressions*. New York: Guilford.
- Spain, J. S., Eaton, L. G., & Funder, D. C. (2000). Perspectives on personality: The relative accuracy of the self versus others in the prediction of emotions and behavior. *Journal of Personality*, 68, 838–867.
- Swim, J. K. (1994). Perceived versus meta-analytic effect sizes: An assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology*, 66, 21–36.
- Vallacher, R. R., & Wegner, D. M. (1987). What do people think they're doing? Action identification and human behavior. *Psychological Review*, 94, 3–15.
- Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, 40, 472–481.
- Vazire, S., & Gosling, S. D. (2004). e-Perceptions: Personality impressions based on personal web-sites. *Journal of Personality and Social Psychology*, 87, 123–132.
- Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The relative accuracy and unique predictive validity of self- and other ratings of daily behavior. *Journal of Personality and Social Psychology*, 95, 1202–1216.
- Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, 55, 493–518.