Journal of Research in Personality 47 (2013) 385–389

Contents lists available at SciVerse ScienceDirect

Journal of Research in Personality

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

Group personality judgments at zero acquaintance: Communication among judges versus aggregation of independent evaluations

Andrew Beer *

University of South Carolina Upstate, United States

ARTICLE INFO

Article history:
Available online 26 March 2013

Keywords:
Personality assessment
Reliability
Aggregation
Accuracy
Validity
Consensus
Zero acquaintance

ABSTRACT

The current study ($N = 264$) compared the validity of personality judgments made by groups of 2, 3, or 4 people to the validity of personality judgments from 2, 3, or 4 aggregated individual reports. I replicated the general increase in validity that accompanies the aggregation of independent judgments. However, group judgments did not follow this pattern. Small groups outperformed the average single rater, but increasing group size did not lead to similar increases in validity. In short, two heads are better than one across both judgment scenarios, but the point of diminishing returns on additional group members occurs more quickly when judgments are made interactively.

1. Introduction

When consequential personality judgments form in daily life, they frequently do so as a result of discussion. In fact, it seems that the more important the assessment, the more likely we are to seek outside opinions prior to solidifying our impressions of others' general, individualized tendencies—potential spouses must meet parents and hiring committees must confer in person prior to making decisions about applicants. Such practices have a clear implication: we believe that together, in groups, we make more accurate personality judgments than we do on our own.

Research in personality assessment suggests that aggregating multiple independent personality judgments of a single target will enhance accuracy relative to any single judgment (Watson, 1989). In short, many heads are better than one, and—though there are limits—more heads are better than fewer. However, there is a substantial literature in social psychology indicating that group decisions can frequently be inferior to those made by individuals in the absence of the social pressures created by certain group situations. Group members will sometimes expend less energy on a joint task (social loafing; Ingham, Levinger, Graves, & Peckham, 1974). Some conditions lead groups to become more extreme after deliberating together (group polarization; Moscovici & Zavalloni, 1969), and while they theoretically have more total information, groups do not always share unique information effectively (Stasser & Titus, 1985). Even some well-established groups are subject to various pitfalls in group decision making that can have disastrous implications (group think; Janis, 2007).

So, are groups better or worse in assessing personality? Clearly, it depends. The issue centers on the process of information synthesis. In cases such as Watson's (1989), the "group judgment" is simply an arithmetic aggregation of independent judgments. This process capitalizes on the tenets of reliability theory, with multiple indicators serving to reduce random measurement error and highlight true score variance. In addition to this, aggregating independent judgments may also reduce rater-specific bias and idiosyncrasies in impression formation (Kenny, 2004). The other process, which involves active discussion and thus non-independence in judgments (i.e., correlated error components), seems to be less effective in general (for several examples, see Surowiecki, 2004). To date, there has been little research examining the effectiveness of this particular type of judgment process applied specifically to personality assessment, but some existing data indicate that groups are indeed less efficient if they are allowed or encouraged to communicate prior to making a personality judgment. Borkenau and Liebler (1994) asked groups of five raters to first provide initial, idiosyncratic trait judgments for a given target, and then subsequently come to a group consensus rating for each of three traits (Extraversion, Conscientiousness, Intelligence). They found that the consensus ratings were less accurate than even single independent ratings made by a separate group of judges. More recently, however, Leising, Fritz, and Borkenau (submitted for publication) did not replicate this effect for judgments of intelligence, finding instead that small groups performed relatively equivalently to single individuals.

* Address: Department of Psychology, University of South Carolina Upstate, 800 University Way, Spartanburg, SC 29303, United States.
E-mail address: abeer@uscupstate.edu
URL: http://www.uscupstate.edu/academics/arts_sciences/psychology/psychologylab/

0092-6566/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jrp.2013.03.008

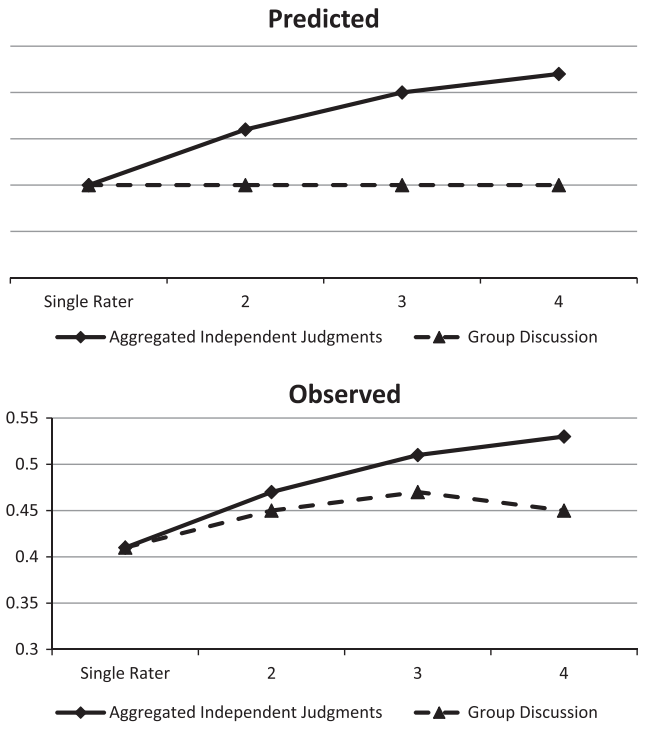

Fig. 1. Predicted and observed accuracy for independent judgments versus group discussion collapsed across traits.

The aim of this study is to replicate and extend these findings by systematically comparing aggregated independent judgments with group judgments made by the same number of individuals. Doing so affords an opportunity to determine whether (a) non-independent aggregation is truly harmful to validity in personality judgment and (b) at what point (in terms of number of opinions or judgments) we begin to observe gains and/or diminishing returns in both independent and interdependent group judgments of personality. This study has the advantages of (a) utilizing the same group of targets for each type of rating (group versus aggregated individual) and (b) systematically evaluating the effect of group size for each type of rating. In a review of group judgment accuracy, Gigone and Hastie (1997) conclude that "group judgments tend to be more accurate than the judgments of typical individuals, approximately equal in accuracy to the mean judgments of their members, and less accurate than the judgments of their most accurate member" (pp. 153). However, based on the most relevant extant research (Borkenau & Liebler, 1994; Leising et al., submitted for publication; Watson, 1989), I formulated three primary hypotheses with respect to accuracy in personality judgment (also summarized in the top portion of Fig. 1):

Hypothesis 1. Increasing the number of aggregated independent judgments would increase accuracy. General tenets of reliability theory would indicate that these estimates would be decreasingly influenced by random error, as evidenced in Watson (1989).

Hypothesis 2a. Increasing group size in interdependent judgment conditions would not increase accuracy. Previous studies have not systematically examined the influence of group size on the accuracy of personality judgment, but in general, groups tend to perform worse than their best individual members (Hastie, 1986) and benefits from aggregation of independent judgments tend to disappear when judges have repeated access to others' estimates or opinions (Lorenz, Rauhut, Schweitzer, & Helbing, 2011).

Hypothesis 2b. Group judgments—irrespective of group size—made after discussion would be no more accurate than those made by single raters. The two primary studies conducted on this topic to date have yielded mixed findings in this respect. Borkenau and Liebler (1994) observed lower self–other agreement correlations for Extraversion, Intellect, and Conscientiousness for judgments made by 5-member groups after discussion relative to single independent raters, but Leising et al. (submitted for publication) observed generally equivalent accuracy correlations for Intelligence in group discussion and independent rating situations. Given that (a) the current study employs a composite accuracy criterion (more similar to the latter study) and (b) involves a situation in which judges may have access to slightly more information than judges in the former study, I expected that groups would perform similarly to the average single independent rater.

2. Method

2.1. Participants

2.1.1. Targets

Watson and Humrichouse (2006) conducted a study of 300 married couples in which each spouse was photographed and rated both himself/herself and his or her spouse on a series of personality dimensions. A subset of these individuals served as targets in this study. Due to the nature of the experimental task, stimulus selection depended in part on the target's score on whichever trait dimension with which his or her photograph would be presented. I chose only targets for whom both self and spouse ratings were extreme (above 80th or below 20th percentile for the dimension) or normative (at or within .2 standard deviations of the median for the dimension). From this pool of 600 possible targets, 72 (36 female) were selected to be used in the current study: 12 each (four high, four median, and four low scorers) for the five dimensions comprising the Five-Factor Model and six each (two high, two median, two low scorers) for two additional dimensions (Positive and Negative Affectivity). Gender was split evenly among the dimensions, such that the same of number of males and females would be presented in each dimension.

2.1.2. Judges

Judges were 264 (206 female, 136 Caucasian, 106 African American) undergraduates from an introductory psychology course at a mid-sized southeastern university. Individuals participated in partial fulfillment of a course research requirement.

2.2. Measures

Both group and independent personality judgments of targets were made using the Ten-Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003). This instrument was designed as a short form version of the Big Five Inventory (BFI; John & Srivastava, 1999), and has shown strong convergent validity with the BFI scales (Ehrhart et al., 2009; Gosling et al., 2003). Participants rated targets using a 7-point scale (1 = disagree strongly, 7 = agree strongly) in response to a series of adjectives and phrases following a stem statement ("I see myself/this person as…").

2.3. Procedure

Participants were asked to make a series of personality judgments from photographs of individuals in one of two experimental conditions. In one condition, participants made judgments independently (six targets in a half-hour session), and in the other condition, small groups (2–4 people) were asked to come to a

consensus about each target and form a "group rating" of the target (three targets in a half-hour session). Thus, 48 judges participated individually, yielding four independent judgments for each of the 72 targets. The other 216 judges participated in groups of 2 (48 judges), 3 (72 judges), or 4 (96 judges), yielding group estimates for each size for each target.

Each target photograph was accompanied by a sentence that provided a clue as to the general nature of the person pictured. These clues provided extra information to help ameliorate the floor effects inherent in zero acquaintance personality judgment, allowing for the possibility to observe any potential deleterious influence of group discussion. Each group of raters would be exposed to trait information specific to only one domain, and at all three possible levels (high, median, low) within that domain. For example, if Group A's first target photo was selected as a high scorer in the domain of Conscientiousness, his or her photo would appear along with a sentence implying this standing ("This person is generally responsible and tends to follow through with plans"). Group A's next target would then be accompanied by a sentence indicating the opposite ("This person is generally irresponsible and tends not to follow through with plans"), and the group's third and final target photo would be accompanied by a sentence implying that this individual scored between the extremes on the dimension ("This person is sometimes responsible and follows through with plans, but can also be irresponsible and leave things unfinished"). The primacy of "high" and "low" targets was counterbalanced within dimensions (and across groups), but median scorers were presented last in all cases, given that the sentence implying "median" standing can best be interpreted only after seeing the high and low designations. In all, each of the 72 targets was judged by 4 independent judges, and groups composed of 2, 3, and 4 individuals. In group rating conditions, judges were instructed to work together on each rating, and a single judge was randomly selected to record the group's consensual ratings.

3. Results

To examine accuracy, I correlated peer judgments (made independently or after group discussion) with the average of the target's self-judgments and judgments made of the target by his or her spouse for a given trait domain. It is generally preferable to utilize multiple sources of information to estimate personality (Letzring, Wells, & Funder, 2006), and the target selection criteria ensured that these estimates were highly correlated in the sample, further justifying the aggregation. Thus, what follows is a variable-centered analysis of accuracy with the target serving as the primary unit of analysis and comparisons being made across judgment conditions. Given the difficulty in observing statistically significant differences between correlations (detecting a significant difference between independently obtained correlations of .30 and .50 would require that each sample contain more than 135 observations), my discussion of these results will rely on establishing patterns of effect sizes in the data, an accepted practice in personality accuracy research (e.g., Vazire, 2010). This is not to suggest that statistical significance is unimportant as a concept but rather to acknowledge that previous work in this area (e.g., Watson, 1989) has relied upon establishing consistent patterns of very small effect sizes (correlations often differing by .05 or less).

Table 1 provides the accuracy estimates across judgment conditions. The most obvious trend concerns the value of the additional personality-related information, as all single-rater accuracy correlations (save Extraversion) exceed levels typically observed when participants are provided only a photograph (Beer & Watson, 2010; Borkenau & Liebler, 1992; Naumann, Vazire, Rentfrow, & Gosling, 2009).

Table 1
Accuracy correlations across judgment conditions.

Variable	Single rater	Aggregation of independent judgments			Group discussion		
		2	3	4	2	3	4
Neuroticism	.44	.55	.61	.65	.42	.54	.40
Extraversion	.34	.40	.44	.46	.54	.38	.48
Openness	.29	.34	.36	.37	.35	.37	.47
Agreeableness	.50	.58	.61	.63	.52	.56	.44
Conscientiousness	.43	.46	.52	.53	.41	.48	.45
M	.40	.47	.51	.53	.45	.47	.45

Note. $N = 72$ (target photographs). Single-rater estimates represent the mean accuracy of each of the four independent peer ratings. The 2-person independent composite represents the mean accuracy of the six possible 2-rater combinations. The 3-person independent composite represents the mean accuracy of the four possible 3-rater combinations.

However, the primary aims of the study were replication and extension of previous findings regarding independent and group judgments of personality. I first hypothesized a conceptual replication of Watson's (1989) findings regarding the value of aggregated independent observers. The first column of Table 1 provides the average accuracy of the four separate independent peer ratings across targets. Each single peer accuracy correlation was Fisher-transformed, and the four estimates were averaged and then inverse-transformed for the estimates in the table. The second, third, and fourth columns represent aggregation of independent judgments group sizes of 2, 3, and 4. To calculate these estimates, I computed the accuracy correlations for each possible combination of 2 (6 possible combinations), 3 (4 possible combinations), and 4 (only one possible combination) raters, and then, in the case of the 2- and 3-rater aggregates, averaged them in similar fashion to the estimates in the first column. The results are uniform and clear: adding independent raters to an aggregate increases the validity of peer judgments in this near zero acquaintance setting. Furthermore, these gains take the form of a negatively accelerating curve, with the average increase in accuracy decreasing with each additional rater. Specifically, the average increase across all five traits from a single rater to a 2-rater composite was .07, greater than the difference between 2- and 3-rater composites (.04) and between 3- and 4-rater composites (.02). Thus, Hypothesis 1 was confirmed. In addition, judgments of Neuroticism seemed to benefit most from additional raters: the jump in validity from a single rater to a 4-rater composite was .21 versus .13, .12, .10, and .08 for Agreeableness, Extraversion, Conscientiousness, and Openness, respectively.

I next hypothesized that increasing group size would not enhance the accuracy of trait judgments. The fifth, sixth, and seventh data columns in Table 1 provide the accuracy estimates for groups of 2, 3, and 4 members, respectively. Average accuracy (across trait domains) was relatively similar across group size, with 2-member groups producing an average accuracy estimate of .45 (versus .47 and .45 for 3- and 4-member groups, respectively). Contrary to the independent aggregation data, however, there was not a clear, uniform pattern across traits. First, it is worth noting that in only one case (Openness) did groups of 4 demonstrate the highest level of observed accuracy. Three-member groups performed better than 2- or 4-member groups for Neuroticism, Agreeableness, and Conscientiousness, and 2-member groups showed relatively greater accuracy for Extraversion. Overall, these data are supportive of Hypothesis 2a with the caveat that the differential patterns across dimensions were unforeseen.

Finally, I hypothesized that group estimates would be similar to single independent judgments in terms of accuracy. Two-member groups performed similarly to single raters on Neuroticism (.42 versus .44), Agreeableness (.52 versus .50), and Conscientiousness

(.41 versus .43), slightly outperformed single raters for Openness (.35 versus .29), and substantially outperformed single raters for Extraversion (.54 versus .34). These patterns did not hold across group sizes, however. Three-member groups showed a much more uniform pattern of results, with estimates for each domain exceeding the relevant single-rater estimates; differences ranged from .04 (Extraversion) to .10 (Neuroticism). Four-member groups slightly underperformed single-rater estimates for Neuroticism (.40 versus .44) and Agreeableness (.44 versus .50), slightly outperformed single raters for Conscientiousness (.45 versus .43), and substantially outperformed single raters for Openness (.47 versus .29) and Extraversion (.48 versus .34). Overall, groups showed greater accuracy on 11 of the 15 possible comparisons between group judgments and single raters. Although these differences are less consistent than those observed between single raters and aggregated independent observers (for which every estimate for every trait shows an increase in accuracy with added raters), it is unlikely to have occurred by chance: a binomial test indicates that the probability of 11 or more of these 15 comparisons being greater in the group judgments given the assumption of equality across conditions is only .059. Thus, Hypothesis 2b is unsupported, as it seems that groups engaging in conversation are generally slightly more accurate than a single individual. However, this advantage disappears with the addition of one more single independent rater: of the 15 possible comparisons between groups of varying size and 2-rater composites, groups only outperform the composites six times.

Fig. 2 plots the accuracy estimates, by judgment dimension, for single-raters, 2-, 3-, and 4-rater composites, and groups of 2, 3, and 4 members. The error bars on the diamonds provide the range of observed accuracy for the single-, 2-, and 3-rater estimates. It is noteworthy that the average of the multi-rater composites of each size (represented in Fig. 2 by the diamonds) typically exceeds the highest observed single-rater accuracy (represented in Fig. 2 by the upper limits of the error bars for one-rater estimates), particularly with the addition of the 4th rater (Openness is an exception). The relatively random pattern of the circles reiterates the prior discussion: there is no consistent pattern across group size, and overall, groups perform slightly better than the average single rater and slightly worse than multi-rater composites. Another useful way to compare groups versus individuals would be by comparing the best single independent judgment to those made by groups of varying size. In the fifteen comparisons between group judgments and the highest observed accuracy of a single judge within a domain, groups only outperformed the best single rater four times, which again would violate the assumption that these judgment conditions were equal with respect to accuracy ($p = .059$).

4. Discussion

I designed a study to examine two primary methods of consensus building: aggregation of independent judgments versus group discussion. This design was particularly advantageous for a few reasons. First, to my knowledge this is only the third study to examine these two types of group judgments simultaneously. Furthermore, this is the only study also to systematically evaluate the influence of group size in each judgment condition. The between-subjects nature of the judgment manipulation served to eliminate potential carryover effects, and the design also allowed for diversity in both the targets and judges. These features allowed me to test three key hypotheses.

First, I expected to replicate the previous finding (Watson, 1989) that additional independent raters would increase the accuracy of personality judgment across all domains. Indeed, I observed a clear and consistent pattern across domains in the form of a negatively accelerating curve. Second, I expected that I would not observe such a trend in group judgments, as aggregated judgment via discussion is unlikely to present the same advantages as aggregated independent judgments in relation to reducing random error. The data also supported this hypothesis, as group judgments did not systematically increase (or decrease) with added size. If one were forced to make a conclusion about a general pattern, it seemed that 3-member groups slightly outperformed 2- and 4-member groups overall, but this effect was small and inconsistent across trait domains. Third, I expected that groups would be no more accurate than individual raters across all trait domains. This hypothesis was unsupported, as groups, in general, outperformed the average single rater. In fact, groups on the whole tended to perform similarly to 2-rater composites. However, this advantage was in comparison to the average single-rater. The best single rater tended to outperform the groups. The general trends in the data are plotted in Fig. 2.

Confirmation of Hypothesis 1 was hardly surprising given the general tenets of reliability theory, and Hypothesis 2a had not been

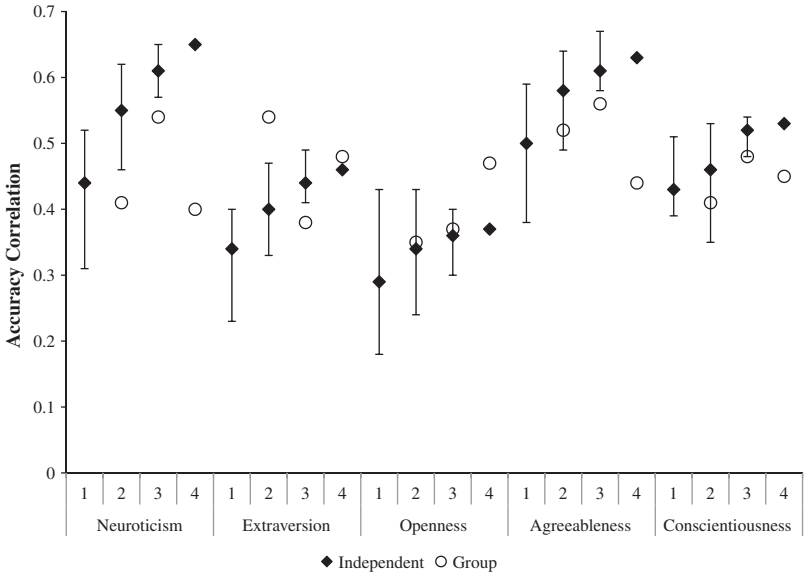

Fig. 2. Aggregated independent judgment versus group discussion accuracy by trait. Error bars represent the range of correlations observed across different rater combinations in the independent judgment condition. There was only one possible combination of four raters and thus no observed range.

previously tested with respect to personality judgment. However, findings related to Hypothesis 2b do not conform to previous results. Borkenau and Liebler (1994) found that groups performed worse than the average independent rater, and Leising et al. (submitted for publication) found that groups performed relatively equivalently to the average independent rater whereas in the current study, groups slightly outperformed the average single rater across trait judgments. It is important to note that the statistical test that supports this distinction is somewhat imperfectly applied—one could argue that the comparisons are not wholly independent—and does not speak to the magnitude of the effect. In addition, this design did differ from previous endeavors in some important ways. Borkenau and Liebler's (1994) participants had access to dynamic audio and visual cues, whereas participants in this study were provided with only a photograph and thus had no access to such cues. Also, participants in this study received ancillary information, which may have fundamentally changed the decision making process. Nevertheless, there is much work to be done to elucidate this particular issue, as its practical implications may be among the most important: should we even confer with others about trait judgments? These results indicate that—at the very least—group discussion does not hinder accuracy.

Another important issue that remains unresolved involves the effect of group size. Although there was no clear pattern with respect to the impact of group size on accuracy in personality judgment, it is worth noting that these data do not follow the same trajectory as aggregated independent judgments. One limitation of this study was that the comparison halted at 4-member groups. If one were to extrapolate from the given data, a 5-member group likely would have performed worse than even the average single rater, which is noteworthy given that Borkenau and Liebler's (1994) study (in which groups performed worse than the average single rater) involved 5-member groups. Future research should aim to determine whether there is an optimal group size (perhaps 2 or 3 members), after which we might observe a downward trend in accuracy, as opposed to the asymptotic pattern observed for aggregated independent judgments of increasing group size (see lower portion of Fig. 1).

Another limitation that this study shares with all others to date is its focus on a near zero-acquaintance judgment setting. Although many consequential group personality judgments occur naturally under conditions of unfamiliarity between group and target (e.g., a job interview), it would be useful to examine these effects in groups of already-acquainted individuals or under more natural interaction conditions. Specifically, although the group is often unacquainted with the target when consensual decisions about personality are of particular practical importance, the group itself may often consist of people with at least some familiarity with each other. This may lead to different information-sharing strategies or different dynamics in terms of general judgment process. Future studies of group personality judgment might focus more carefully on the group decision making process, borrowing paradigms from social psychology to identify the most advantageous strategies for arriving at an accurate consensual impression of a given target. In sum, varying the level of familiarity both between group and target and within group could potentially lead to different conclusions about the efficacy of group discussion as it pertains to judging personality.

A final concern about the current study lies in its relatively small sample size. Ultimately, the cost of utilizing a more reliable accuracy criterion was a restriction in the number of eligible targets from the original pool. This precluded some potentially useful analytic strategies and lowers confidence in some of the observed effect size estimates. Future studies might endeavor to either (a) expand the number of targets in a similar variable-centered analysis or (b) expand the number of items addressed per target to facilitate a person-centered analytic approach. Each of these options presents different logistical problems, which may contribute to the fact that there have been such few studies of this phenomenon in the literature to date.

In conclusion, we frequently consult others prior to making important decisions about the nature of a given target individual, but there is a paucity of data that speak to the efficacy of consensus judgments of personality. The current study suggests that although aggregating independent judgments of personality may be a superior strategy, group discussion may still produce more accurate personality judgments than those provided by a single independent rater.

References

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

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

Borkenau, P., & Liebler, A. (1994). Effects of communication among judges on the validity of their judgments. *European Journal of Psychological Assessment, 10*, 10–14.

Ehrhart, M. G., Ehrhart, K., Roesch, S. C., Chung-Herrera, B. G., Nadler, K., & Bradshaw, K. (2009). Testing the latent factor structure and construct validity of the ten-item personality inventory. *Personality and Individual Differences, 47*, 900–905. http://dx.doi.org/10.1016/j.paid.2009.07.012.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*, 149–167. http://dx.doi.org/10.1037/0033-2909.121.1.149.

Gosling, S. D., Rentfrow, P. J., & Swann, W. R. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*, 504–528. http://dx.doi.org/10.1016/S0092-6566(03)00046-1.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.). *Decision research* (vol. 2, pp. 129–157). Greenwich, CT: JAI Press.

Ingham, A. G., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Experimental Social Psychology, 10*, 371–384. http://dx.doi.org/10.1016/0022-1031(74)90033-X.

Janis, I. L (2007). Groupthink. In R. P. Vecchio (Ed.), *Leadership: Understanding the dynamics of power and influence in organizations* (2nd ed., pp. 157–169). Notre Dame, IN US: University of Notre Dame Press.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality* (2nd ed., pp. 102–138). New York: Guilford.

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

Leising, D., Fritz, U., & Borkenau, P. (submitted for publication). *Communication between raters does not improve the accuracy of judgments of people's intelligence.* Manuscript.

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

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 108*, 9020–9025. http://dx.doi.org/10.1073/pnas.1008636108.

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12*, 125–135. http://dx.doi.org/10.1037/h0027568.

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

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48*, 1467–1478. http://dx.doi.org/10.1037/0022-3514.48.6.1467.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York, NY, US: Doubleday & Co..

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

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

Watson, D., & Humrichouse, J. (2006). Personality development in emerging adulthood: Integrating evidence from self-ratings and spouse ratings. *Journal of Personality and Social Psychology, 91*, 959–974. http://dx.doi.org/10.1037/0022-3514.91.5.959.