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

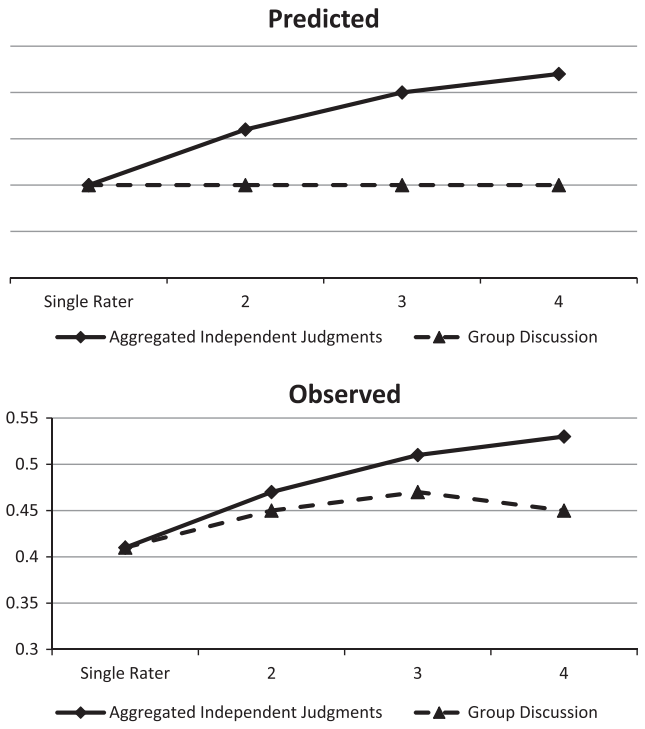

Fig. 1. Predicted and observed accuracy for independent judgments versus group discussion collapsed across traits.

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