FlashReport

Accurate by way of aggregation
Should you trust your intuition-based first impressions?

Noah Eisenkraft

University of North Carolina, Chapel Hill, Campus Box #3490, Chapel Hill, NC, 27599-3490, USA

HIGHLIGHTS

► Scholars disagree about the predictive validity of first impressions.
► These disagreements are the result of studying individual- vs aggregate-level data.
► Individual rater’s first impressions are usually poor predictors of performance.
► Predictive validity increases when one aggregates multiple raters’ first impressions.

ARTICLE INFO

Article history:
Received 9 September 2012
Revised 30 October 2012
Available online 28 November 2012

Keywords:
Intuition
First impressions
Predictive validity
Thin slices
Interviews
Aggregation

ABSTRACT

How much should you trust your intuition about other people’s job performance? Different literatures provide different answers to this question. Social psychological research on “thin slices” suggests that untrained observers can predict a person’s job performance based on a few moments of observation. Industrial/organizational psychologists have found a weaker relationship between job performance and the intuitive judgments that people make following employment interviews. This paper argues that interviewers’ intuitive judgments appear to be weaker predictors than intuitive judgments of thin slices because thin slices research measures predictive validity at the aggregate-level of analysis. Intuition-based first impressions will not usually be valid predictors of job performance unless people have an opportunity to collect and combine the judgments of multiple independent raters.

© 2012 Elsevier Inc. All rights reserved.

Introduction

Social psychologists and industrial/organizational (I/O) psychologists have reached different conclusions regarding the predictive validity of intuition-based first impressions. Social psychologists who study “thin slices” argue that people can use intuitive judgments to accurately predict employee job performance (Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992). During a typical thin slices study, untrained raters evaluate people after watching no more than a few minutes – and, often, only a few seconds – of recorded nonverbal behavior. These intuitive judgments are powerful predictors of how well teachers (Ambady & Rosenthal, 1993), telephone operators (Hecht & LaFrance, 1995), salespeople (Ambady, Krabbenhoft, & Hogan, 2006), and other professionals perform their jobs (Ambady et al., 2000). Intuitive judgments of thin slices explain 15% of the overall variation in job performance ($r = .39$, Ambady et al., 2000).

I/O psychologists who study employment interviews have repeatedly found a weaker relationship between intuitive judgments and job performance (Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994). Although interviews last longer than a few minutes, showcase both verbal and nonverbal behavior, and are conducted with the purpose of predicting job performance, interviewers who listen to their intuition tend to focus on performance-irrelevant factors, apply different standards to different applicants, and do not base their evaluations on consistent criteria (Highhouse, 2008). When people conduct unstructured interviews, their intuitive judgments explain only 4% of the overall variation in applicant job performance ($r = .20$, Huffcutt & Arthur, 1994).

These literatures suggest different conclusions regarding the value of intuition-based first impressions. Research on the predictive validity of thin slice judgments suggests that people should trust their intuition, even if they don’t have much time to form those evaluations (Ambady et al., 2000). I/O psychologists, in contrast, warn that intuition is a threat to predictive validity (Highhouse, 2008). If the conclusions of the interview validity literature are correct, people who trust their intuition often make incorrect predictions and poor selection decisions (Huffcutt & Arthur, 1994; McDaniel et al., 1994). Although there are many contextual differences between thin slices and interviews – e.g., interviews are longer, include both verbal and
nonverbal behavior – these factors do not explain the higher validity correlations in the thin slices literature.

This paper investigates why intuitive judgments of thin slices predict job performance better than intuitive judgments following an employment interview. I argue that the literatures’ different conclusions are the result of studying predictive validity at different levels of analysis. I/O psychologists study predictive validity at the individual-level of analysis; they calculate the correlation between job performance and individual interviewers’ evaluations (Huffcutt & Arthur, 1994). Social psychologists often study the predictive validity of intuitive judgments at the aggregate-level of analysis (Ambady et al., 2000; Kunda, 1999). I draw on the statistics literature to suggest that the benefits of aggregation may explain why there is such a strong correlation between job performance and people’s intuitive judgments of thin slices.

The idea that aggregation can strengthen correlations is as old as modern statistics. The imperfect act of measurement introduces errors that reduce the strength of the associations between any variables (Spearman, 1904). Researchers, however, can remove some of this error by combining multiple ratings into a single measurement. When ratings are averaged together, the error introduced by one idiosyncratically positive rating may be canceled out by the error from a different idiosyncratically negative rating. Correlations between composite measures will often be stronger than the correlations between single-rating measures because the aggregation process removes some of the measurement error that attenuates bivariate correlations (Brown, 1910; Spearman, 1910).

Ostroff (1993) describes the circumstances under which aggregation has the largest effect on the correlation between two variables. Although researchers gravitate toward using the most reliable measures, the benefits of aggregation are inversely related to the reliability of the individual-level data (Ostroff, 1993). When there is more error at the individual-level of analysis, there is more error for the aggregation process to remove. Insofar as there are a sufficient number of ratings, aggregation will have the largest effect when the reliability of the individual-level data is low.

Building on this logic, I hypothesize that aggregate-level intuitive judgments will predict performance better than the intuitive judgments of individual raters. Intuitive judgments tend to be unreliable at the individual-level of analysis; they have a signal to noise ratio of only 15% signal to 85% noise (Albright, Kenny, & Malloy, 1988). The aggregation process should, therefore, minimize the extent to which individuals’ personal biases, idiosyncratic preferences, and random noise decrease the correlation between individuals’ intuitive judgments and job performance. I hypothesize that moving from the individual- to the aggregate-level of analysis will remove the attenuating effects of error, significantly increasing the predictive validity of people’s intuition-based first impressions.

Method

This study measures the extent to which aggregation increases the predictive validity of intuition-based first impressions. Although the laboratory procedure is similar to the procedure used in most research on thin slices – participants independently evaluate 34 ten-second video clips of students being interviewed – I do not focus exclusively on the predictive validity of the participants’ aggregated ratings. I calculate and compare the predictive validity of the participants’ intuition-based first impressions at both the individual- and the aggregate-levels of analysis.

Participants

Forty-one American adults (17 male, 24 female; Average age = 35.6, SD = 13.5) participated in this study. Participants were recruited through Amazon.com’s Mechanical Turk service, an integrated participant recruitment and compensation system that is more diverse and as reliable as traditional samples of American college students (Buhrmester, Kwang, & Gosling, 2011). None of the participants were affiliated with the university where the study materials were created.

Procedure

All participants provided their intuition-based first impressions of 34 undergraduate students (17 male, 17 female) after watching ten-second silent videos of the students being interviewed. The female graduate assistants who conducted the interviews were blinded to the purpose of this study.

The undergraduate student’s grade point average (GPA) was used as an indicator of their job performance. Participants predicted the students’ GPAs on a 0.0 to 4.0 grading scale (e.g., 3.0 = B, 3.1, 3.2, 3.3 = B+, 3.4, etc.) after watching the students’ videos.

Results

The predictive validity of intuition-based first impressions can be measured at the individual- and aggregate-levels of analysis. I measure predictive validity at both levels of analysis to assess whether aggregation increases the relationship between intuitive judgments and performance.

At the individual-level of analysis, predictive validity is the correlation between individuals’ intuitive judgments – each participant’s intuitive judgments of the students’ GPAs – and performance—the students’ actual GPAs. Across all 41 participants, these predictive-validity correlations are relatively small though, on average, significantly greater than zero (Mean $r = .16$, SD $= .18$, $z = .17$, $t(40) = 5.53$, $p < .001$). These individual-level intuitive judgments also show the low reliability (ICC1 = .16) that has been reported in previous research on intuitive judgments (Albright et al., 1988).

At the aggregate-level of analysis, predictive validity is the correlation between the averaged intuitive judgments of multiple raters and performance. The participants’ aggregated judgments positively predict the students’ GPAs ($r = .37$, $t(32) = 2.25$, $p = .03$). As expected, the aggregate-level intuitive judgments are also highly reliable (ICC2 = .89).

I hypothesized that moving from the individual- to the aggregate-level of analysis would increase the predictive validity of intuition-based first impressions. As predicted, the predictive validity correlation between intuitive judgments and performance is significantly stronger than the average, individual-level, correlation between those same variables (Aggregate-level $r = .37$, Average individual-level $r = .16$, Difference $r = .21$, $z = .22$, $t(40) = 7.05$, $p < .001$). Aggregating the unreliable individual-level data significantly increases the predictive validity of intuition-based first impressions.

Discussion

The results of this study show how aggregation can increase the predictive validity of intuition-based first impressions. The difference between the individual- and aggregate-level correlations in this study parallel the difference in conclusions reached by I/O psychologists who study interview validity and social psychologists who study thin slices. The predictive validity of the average participant’s intuitive judgments ($r = .16$) is similar to the predictive validity of the intuitive judgments people form after unstructured interviews ($r = .20$; Huffcutt & Arthur, 1994). The aggregate-level correlation from this study ($r = .37$) is similar to the predictive validity of thin slice judgments ($r = .39$; Ambady et al., 2000). Although unmeasured contextual forces and psychological mechanisms may also affect predictive validity, this pattern of results suggests that social psychologists and I/O psychologists may have reached different conclusions because
they studied the relationship between intuition-based first impressions and performance at different levels-of-analysis.

My conclusion regarding the inferential consequences of aggregation is also supported, albeit indirectly, by Ambady et al.'s (2000) meta-analysis of the thin slices literature. Ambady et al. (2000) report that aggregate-level thin slice judgments are correlated with job performance at $r = .39$ and have an individual-level reliability of .22. Given this information, we can use Spearman's (1904) correction for attenuation formula to estimate the average individual-level correlation. Assuming that the aggregate-level judgments approach perfectly reliability, the correction for attenuation formula suggests that the average individual-level predictive validity of thin slice judgments is $r = .18$. Without often-overlooked benefits of aggregation, individual-level thin slice judgments appear to be no more predictive than the ratings following an unstructured interview.

We now return to the paper's motivating question: Should you trust your intuition? The results of this study suggest that you, as an individual, should not. Individual-level intuitive judgments are too encumbered by idiosyncratic biases and random noise to consistently produce valid predictions of job performance. Intuition-based first impressions should only be trusted when there is an opportunity to collect and combine the judgments of multiple independent raters. There are, however, few situations where this type of aggregation is possible or appropriate. The results of this study add to the growing body of evidence (e.g., Ames, Kammrath, Suppes, & Bolger, 2010; Olivola & Todorov, 2010) that people should question whatever confidence they have in the predictive validity of their first impressions.

My interpretation of these data must, of course, be considered alongside the study's limitations. I investigated only how aggregation influenced the correlation between intuition-based first impressions of undergraduate students and the students' academic performance. The results of this investigation echo the conclusions of researchers who studied how aggregation influences the relationship between inferences based on people's portraits and people's intelligence and personalities (Todorov, Said, & Verosky, 2011). Future studies should both replicate the present study and explore how aggregation influences intuitive judgments that are based on real-world interactions and of people already in the workforce.

It is also important to investigate whether aggregation has a similar effect in other research domains. For example, how much are the conclusions of the group judgments literature — a literature that asks whether groups more or less accurate than individuals (Gigone & Hastie, 1997; Kerr & Tindale, 2004) — affected by the fact that a group's decisions are a function of multiple individual-level inferences? Although the scope of this paper was limited to the relationship between intuitive judgments and performance, the effects associated with moving up a level of analysis may generalize to phenomena in other contexts.

If replicated, this paper's findings have implications beyond understanding the predictive validity of intuition-based first impressions. The results of this study highlight the consequences of studying psychological processes with aggregate-level data. These consequences have been a topic of perennial discussion within psychology (see, e.g., Brand & Bradley, 2012; Gordon, 1924; Monin & Oppenheimer, 2005; Zajonc, 1962), but they bear repeating. Calculating a simple average may strengthen the relationships between variables — making it easier to find statistically significant results — but aggregate-level correlations must be interpreted differently from individual-level results. An aggregate-level correlation describes what would be possible in a world without error; it may not accurately describe what people experience outside of the psychological laboratory. More discussion about a study's level-of-analysis would help clarify whether a study describes what could occur versus what does occur during people's real-world interactions.

### References


