The Value of Pooling “Naive” Expertise

Marc S. Schulz
Bryn Mawr College

Robert J. Waldinger
Harvard Medical School

We read with great interest the article by Westen and Weinberger (October 2004) exploring the benefits and limitations of clinical observation and judgment. They make a persuasive case for using statistical methods to aggregate clinicians’ observations in clinical and personality research. Just as increasing the number of items on an assessment instrument can improve reliability, statistically aggregating the judgments of multiple observers is likely to yield a judgment more reliable than that of a single observer. The advantages of pooling multiple observations, however, are not limited to research in clinical and personality psychology. Westen and Weinberger identify two categories of informants—clinicians and participants—but these categories could be expanded to include other observers who might have particular expertise or experience related to the phenomenon of interest. The type of expert best suited to provide observations depends on the type of expertise required. For reports about subjective psychological or physical experience, the individual him- or herself is likely to be most expert. For assessment of clinically observable phenomena, clinicians may indeed bring value-added expertise. There are some domains, however, in which those with the greatest expertise are neither specially trained observers nor self-reporters but, rather, lay observers who have a native or learned ability to detect complicated social or psychological phenomena and make subtle discriminations. This type of expertise is often thought of as intuitive because it uses implicit knowledge that is not always accessible to conscious awareness or capable of being fully articulated. One way to harness this intuitive expertise effectively is to pool the judgments of multiple lay observers.

Using lay expertise in observational research runs counter to the prevailing belief that manualized coding systems are required to obtain valid, reliable, and reproducible results. To be sure, manualized approaches have contributed greatly to research in psychology. However, some phenomena do not lend themselves easily to codification and rule-driven observation. For example, emotions, power dynamics, validation, and empathy have all been shown to be of great relevance to interpersonal functioning, but these are challenging to operationalize. Yet they are phenomena that individuals routinely evaluate in their everyday interactions with others. When using manual-based systems, coders may find that the rules and procedures designed to facilitate rating of these constructs do not fully utilize—and may even interfere with—the natural decoding expertise that they have developed. Needing, for example, to process all of the information in a manual and simultaneously attend to the behavioral segment being coded may overload the cognitive capacities of many individuals and interfere with their ability to utilize important cues.

Consider the example of assessing emotion. Because it is socially adaptive to do so, most human beings develop a sophisticated and almost instantaneous ability to recognize others’ emotions. People use complex internal algorithms to synthesize disparate cues about emotion from body language, facial expression, vocal qualities, speech content, and social context. Of course, individual differences in this proficiency exist. Just as aggregating judgments across clinicians removes idiosyncratic biases and yields a more reliable composite judgment, aggregating the individual judgments of multiple untrained judges yields similar benefits.

Our research has led us to believe that lay observers’ intuitive judgments about emotions may in fact capture important information that is lost when coders depend on more commonly used manualized approaches such as the Specific Affect Coding System (SPAFF; Gottman, McCoy, Coan, & Collier, 1996) and the Facial Action Coding System (Ekman & Friesen, 1978). In our laboratory, we experimented for a year with manualized approaches to coding emotion and then decided to explore the utility of what is often (though perhaps misleadingly) termed naive coding. Inspired by the work of Rosenthal and colleagues on coding nonverbal behaviors (e.g., Ambady & Rosenthal, 1993), we provided untrained, unmarried college-age individuals a list of 16 emotion descriptors without instructing them on how to interpret these descriptors or on markers of specific emotions. We asked them to watch videotapes of 10-min couple interactions and to code the emotional expression of each partner. For each 30-s segment of the interaction, coders rated the intensity with which each of the 16 emotions was displayed. Despite the fact that each rater coded in slightly idiosyncratic ways, we obtained good reliabilities on factor-analytically derived scales by combining the ratings of 5–6 coders. The aggregated ratings of these “naive” coders were linked in expected ways with concurrent marital satisfaction and independent assessments of marital adjustment. More important, they predicted with more than 80% accuracy a real-world consequence of great significance: whether couples in committed relationships would remain together after 5 years (Waldinger, Schulz, Hauser, Allen, & Crowell, 2004). When we compared emotion ratings of couple interactions using this approach with ratings of the same interactions by experienced coders using the manualized SPAFF, we found a high degree of correlation between the two sets of ratings (Waldinger et al., 2004). We also...
found evidence that the untrained raters were more sensitive to subtle or low-intensity expression of emotions.

Our experience with using lay expertise in assessing emotion is consistent with the research cited by Westen and Weinberger (2004) suggesting that “with increasing experience in a given domain, people are typically able to make more subtle discriminations, process information more efficiently, and automatize procedures that initially required conscious attention and hence consumed working memory resources” (p. 599). To succeed in the social world, individuals must learn to make judgments about many complex psychological constructs, such as emotion, that are complicated to define but are nevertheless meaningful and commonly observed phenomena. They must make these judgments quickly, often while simultaneously engaging in other cognitively demanding activities. Empirical research indicates that people can learn to recognize complex connections in the social world without having an explicit awareness of, or ability to, articulate this implicit knowledge (Westen & Weinberger, 2004). In our zealously to reduce variability among coders and to make our methods more exportable to other investigators, researchers risk losing the intuitive expertise that people naturally develop in making judgments about the world.

REFERENCES


Science as a History of Corrected Mistakes

James M. Wood
University of Texas at El Paso

M. Teresa Nezworski
University of Texas at Dallas

According to Westen and Weinberger (October 2004), “The history of science can be viewed as the history of confirmatory biases” (p. 609). However, we prefer an alternative formulation: The history of science can be viewed as a constant and largely successful struggle to overcome confirmatory biases. The scientific method is a tool to help people progress toward the truth despite their all-too-human susceptibility to confirmation bias and other errors (Haack, 2003; Sagan, 1995). Put another way, science is a history of corrected mistakes. With this in mind, we wish to correct misrepresentations of our work made by Westen and Weinberger.

As part of a series on the Rorschach Inkblot Test that appeared in Psychological Assessment, we and some colleagues (Garb, Wood, Nezworski, Grove, & Stejskal, 2001) were invited to comment on a series of articles, including one that described the results of a meta-analysis by Hiller, Rosenthal, Bornstein, Berry, and Brunell-Neuleib (1999). Claiming that confirmatory bias characterizes science, Westen and Weinberger (2004) criticize our commentary on this meta-analysis:

The published responses to this (nonparanoid) meta-analysis were decidedly partisan, with some commentators noting approvingly at the study and others dismissing it with a plethora of post hoc methodological complaints, including (our favorite) that its authors (which included Robert Rosenthal, whose chapter on meta-analysis was published around the same time in the Annual Review of Psychology; Rosenthal & DiMatteo, 2000) did not understand the basics of meta-analytic technique (Garb, Wood, Nezworski, & Grove [sic], 2001). (p. 608)

However, Westen and Weinberger’s unflattering description of our commentary, which they ironically single out as their “favorite,” is a misrepresentation of what we actually said. We emphatically deny that we engaged in the ad hominem criticism that Westen and Weinberger attribute to us. Although we described methodological limitations and problems with the Hiller et al. meta-analysis, including poor interrater reliability and failure to blind coding judges, our article neither claimed nor implied that Rosenthal and his colleagues “did not understand the basics” of meta-analysis.

Westen and Weinberger (2004) also misrepresent our conclusions in a chapter on clinical assessment that we wrote for Annual Review of Psychology (Wood, Garb, Lilienfeld, & Nezworski, 2002). According to Westen and Weinberger, we advocated that “clinicians replace their standard diagnostic practices with structured interviews that inquire about each diagnostic criterion for each disorder in the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders [DSM–IV]” (p. 596). In fact, our chapter offered a very moderate recommendation: “Psychologists should probably increase their use of structured interviews in making diagnoses” (Wood et al., 2002, p. 531). We definitely did not advise clinicians to inquire about “each diagnostic criterion for each disorder” in the DSM–IV. Commonly used structured interviews do not usually inquire about relatively rare disorders, such as narcolepsy or factitious disorder. And for the more common disorders, structured interviews typically include screening questions so that interviewers do not have to waste time by inquiring about “each diagnostic criterion.” A structured interview that inquired mindlessly about each criterion for each disorder would be extremely time-consuming and impractical in both clinical and research settings.

The science of psychology will progress most quickly if disagreements are conducted in a respectful manner, with close attention to accuracy. Misrepresenting other psychologists’ conclusions and arguments will lead to needlessly conflicts. If science is a history of corrected mistakes, as we believe, then we hope we have contributed to science with this comment.

REFERENCES


