Journal of Research in Personality 42 (2008) 1476–1489

Contents lists available at ScienceDirect

Journal of Research in Personality

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

Accuracy of judging others' traits and states: Comparing mean levels across tests ☆

Judith A. Hall [a,*], Susan A. Andrzejewski [a], Nora A. Murphy [b], Marianne Schmid Mast [c], Brian A. Feinstein [a]

[a] Northeastern University, Department of Psychology, 125 NI, 360 Huntington Avenue, Boston, MA 02115-5096, USA
[b] Loyola Marymount University, Department of Psychology, 1 LMU Drive, Suite 4700, Los Angeles, CA 90045-2659, USA
[c] University of Neuchâtel, Department of Work and Organizational Psychology, Rue de la Maladière 23, CH-2000, Neuchâtel, Switzerland

ARTICLE INFO

Article history:
Available online 4 July 2008

Keywords:
Interpersonal sensitivity
Personality judgment
Emotion recognition
Accuracy
pi
Binomial Effect Size Display

ABSTRACT

Tests of accuracy in interpersonal perception take many forms. Often, such tests use designs and scoring methods that produce overall accuracy levels that cannot be directly compared across tests. Therefore, progress in understanding accuracy levels has been hampered. The present article employed several techniques for achieving score equivalency. Mean accuracy was converted to a common metric, *pi* [Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106*, 332–337] in a database of 109 published results representing tests that varied in terms of scoring method (proportion accuracy versus correlation), content (e.g., personality versus affect), number of response options, item preselection, cue channel (e.g., face versus voice), stimulus duration, and dynamism. Overall, accuracy was midway between guessing level and a perfect score, with accuracy being higher for tests based on preselected than unselected stimuli. When item preselection was held constant, accuracy was equivalent for judging affect and judging personality. However, comparisons must be made with caution due to methodological variations between studies and gaps in the literature.

1. Introduction

Interpersonal sensitivity—defined as accuracy in judging others' traits and states—has been a topic of research for a very long time (e.g., Jenness, 1932; Vernon, 1933). Interpersonal sensitivity is correlated with many aspects of psychological functioning (Davis & Kraus, 1997; Hall, Andrzejewski, & Yopchick, in press) and is embraced as an important skill in both personality and social psychology (Funder, 2001a; Hall and Bernieri, 2001; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979; Vogt and Colvin, 2003).

However, there are still significant gaps in understanding. These include the origins of interpersonal sensitivity and the nature of causal paths between interpersonal sensitivity and other variables. Another gap, which is the subject of the present article, concerns an understanding of mean levels of accuracy on tests of interpersonal sensitivity. Specifically, we asked whether accuracy for judging traits, such as extraversion, is different from accuracy for judging states, such as emotions. These two traditions of research have had little contact with each other, and the two kinds of accuracy have hardly ever been

☆ The authors are grateful to Robert Rosenthal for advice during the formative stages of this project and David A. Kenny, Frank J. Bernieri, and Charles F. Bond, Jr. for feedback on an earlier version of the article.

* Corresponding author. Fax: +1 617 373 8714.
E-mail address: j.hall@neu.edu (J.A. Hall).

0092-6566/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jrp.2008.06.013

measured in the same group of perceivers (for an exception, see Realo et al., 2003). However, even if they are measured in the same perceivers, the use of incompatible scoring metrics would still prevent direct comparison of accuracy. We also asked how accuracy varied as a function of cue channel (e.g., face versus voice), still versus dynamic stimuli, length of stimulus exposure, and the preselection of stimuli by the test makers. The analysis was based on a database of 109 published results representing many standard and nonstandard interpersonal sensitivity tests.

These questions have not been asked on any scale up until this time. A major contributing reason for this is the wide variation in test designs and scoring systems that exists in this literature. Most crucially, different tests produce scores on different metrics, making comparison difficult to impossible. There are so many ways to measure interpersonal sensitivity that, according to Zebrowitz (2001), it is hard to develop an empirical understanding of what the field shows and ultimately to develop a coherent theory of this kind of skill. In fact, Zebrowitz compared those who study interpersonal sensitivity to the blind men who all declared they were touching a different animal when, in fact, they were all touching the same elephant.

Accuracy of judging traits (e.g., personality or intelligence) is nearly always measured by asking perceivers to make scalar ratings of stimuli, such as videotaped interpersonal interactions, with accuracy calculated as a correlation between judgments and criterion values (e.g., Murphy, Hall, & Colvin, 2003; Watson, 1989). Accuracy of judging states (e.g., emotions) is nearly always measured by having perceivers make categorical judgments of stimuli such as photographs of facial expressions, using a multiple-choice answer format. On such tests, accuracy is calculated as the proportion or percentage correct (e.g., Bond & DePaulo, 2006; Nowicki & Duke, 1994; Rosenthal et al., 1979). The correlational approach and the proportion-correct approach are each well suited to the nature of the content being judged—continuous and categorical, respectively—but they create an "apples and oranges" problem because the scores that each method yields are not on the same metric and therefore cannot be combined or compared directly. This incompatibility is addressed in the present article.

Another incompatibility problem, also addressed here, applies to the proportion-correct approach. For tests within that methodological tradition, accuracy levels often cannot be compared directly because the number of response options varies from test to test. A proportion correct of .50, for example, does not mean the same thing when it is based on a test with two response options versus a test with six options. On the former test a proportion correct of .50 is right at the guessing level whereas on the latter test the same proportion is far above the guessing level.

In the present review, both of these sources of incompatibility were resolved by applying simple conversion procedures whereby all results in the database were expressed using a common metric. This metric was the Proportion Index, or *pi*, developed by Rosenthal and Rubin (1989). Applying a common metric was an essential step before meaningful comparisons across tests could be made.

Lack of a common metric is not always a problem for research synthesis. When the goal is to combine or compare *associations between variables*, as in a typical meta-analysis, standard effect size indices such as Cohen's *d* or the Pearson correlation can be used that do not require variables to be measured on the same metric from study to study (Cooper & Hedges, 1994; Rosenthal, 1991). However, for any application in which *means* will be combined or compared, a common metric is necessary.

Summarizing and comparing means across studies is a recognized, though infrequently applied, method of research synthesis (Rosenthal, 1991). In the field of interpersonal sensitivity measurement, Russell (1994) averaged the mean accuracy of decoding basic facial expressions of emotion across studies in order to examine accuracy as a function of the national origins of the perceivers. Bond and DePaulo (2006) averaged the mean accuracy of detecting deception across studies in order to look at overall accuracy levels and also to relate accuracy to study characteristics. In both of those reviews, the authors limited their summaries to tests that used compatible metrics. Juslin and Laukka (2003) used the *pi* statistic to compare accuracy rates for judging vocally expressed emotions across tests with different designs, and to compare accuracy rates for vocally expressed emotion versus musical renditions of emotion.

In the present summary of published test results, we describe the comparison between accuracy in judging personality versus affect, as well as other methodological comparisons, and we discuss issues that are important when considering accuracy levels on interpersonal sensitivity tests.

2. Method

2.1. Database

In the tests of interpersonal sensitivity included here, perceivers made judgments of adult strangers' recorded expressions or behavior, or of adult strangers in live, though minimal, interaction, after which the researcher scored the judgments for accuracy against a criterion that independently described the targets on the construct in question.

2.1.1. Standard tests

Certain tests were considered to be *standard*, that is, they were established instruments that were supported by psychometric and validity studies. For these standard tests, results used in the present article were the normative data reported in test manuals or in large validity studies. Tests treated in this manner, described in more detail in Appendix A (along with citation information), were: (1) Profile of Nonverbal Sensitivity (PONS: full-length PONS, face and body video PONS, face