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

Hall, Judith A.; Andrzejewski, Susan A.; Murphy, Nora A.; Schmid Mast, Marianne; Feinstein, Brian A.

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

and body photo PONS, voice PONS), (2) Interpersonal Perception Task (IPT: 30-item and 15-item versions), (3) Diagnostic Analysis of Nonverbal Accuracy (DANVA: faces, postures, voices), (4) Pictures of Facial Affect (POFA), and (5) Japanese and Caucasian Facial Expressions of Emotion (JACFEE). Only one result was entered into the database for each of these tests; when results from multiple perceiver samples were reported in the test manuals or validity studies, they were averaged to yield one result. Two additional tests were considered standard: (6) the Communication of Affect Receiving Ability Test with categorical scoring (CARAT) and (7) the Japanese and Caucasian Brief Affect Recognition Test (JACBART), but for these two tests there were no published normative data, per se. Therefore, for these two tests we averaged the results from all the studies we located (11 and 7, respectively) and entered the mean into the database as a single entry for each of those tests. All of the standard tests were based on categorical judgments, and they all scored accuracy in terms of proportion correct, that is, the number of correct answers divided by the number of test items.

2.1.2. Nonstandard tests

For tests that were not standard (see definition above), all individual results that fulfilled the criteria described below were included as separate entries. Because of the different treatment of standard versus nonstandard tests, the number of entries for a particular test or kind of test is not proportional to the frequency of its use in the literature. For example, the single entry for the standard PONS test consists of the average accuracy over 62 groups of participants reported by Rosenthal et al. (1979), whereas the single entry for a typical nonstandard test consists of the accuracy of a single group of participants. In the present article, this is not a problem because the goal was simply to describe accuracy levels, not to describe how often different tests have been used.

Nonstandard tests used both the proportion-correct scoring method (see above) and a correlational method of scoring which we call the across-stimuli accuracy correlation.[1,2,3] In this method, accuracy for an individual perceiver is scored as the agreement between his/her judgments of a set of stimuli (target persons) and the criterion values for those stimuli as indexed by the Pearson correlation. To illustrate, Lippa and Dietz (2000) obtained perceivers' extraversion ratings of 32 30-s video clips, each showing a different target person's behavior. For a given perceiver, accuracy of judgment was calculated as the correlation between the perceiver's ratings and the targets' actual extraversion (measured with a standard personality scale), across the 32 targets. These correlations were then averaged to describe the accuracy of the group of perceivers.

2.2. Search strategy

Results were obtained from the following sources: (1) studies that had been previously gathered for two unrelated meta-analyses (psychosocial and cognitive correlates of interpersonal sensitivity, respectively), including both the studies that were used in those meta-analyses as well as studies that were rejected for not meeting the original authors' inclusion criteria, which were irrelevant for present purposes (Hall et al., in press; Murphy & Hall, in preparation), (2) studies in our own reprint files, (3) PsycINFO search using the term "accuracy", and (4) reference checking in obtained studies. Because

[1] In the present article, we discuss correlational accuracy only if it was scored as an across-stimuli correlation, because this case is conceptually analogous to the proportion-correct approach and can therefore be directly compared once both are converted to *pi*. Readers should be aware, however, that there are other correlational measurement approaches. Correlational accuracy can be calculated as a correlation across ratings on multiple items for a given target (across-ratings correlation). To illustrate, Kolar, Funder, and Colvin (1996) correlated, for each perceiver judging a given target, 41 trait items rated by the perceiver with the same 41 items self-rated by the target. This correlation expresses how well the profile of perceiver ratings matches the profile of target self-ratings (Hall, Bernieri, & Carney, 2005; Kenny & Winquist, 2001). Though in a strict statistical sense the null value for the across-ratings correlation is 0.00, there has been debate over whether the relevant null value is actually larger than 0.00 because of the possibility that the perceiver can obtain some "accuracy" simply by knowing base rates of the traits being judged (stereotype accuracy; Cronbach, 1955; Funder, 2001b; Kenny & Winquist, 2001). This discussion is beyond the scope of the present article. Because the across-ratings correlation and the across-stimuli correlation ask different questions about accuracy (does the perceiver accurately describe a target in terms of a collection of traits or states, versus can the perceiver accurately distinguish among targets for a given trait or state), these two approaches need not be correlated and therefore cannot be considered interchangeable indices. The across-ratings accuracy correlation is not discussed in the present article.

[2] Across-stimuli correlational accuracy can also be calculated at the group as well as individual level (Hall et al., 2005) by averaging perceivers' ratings *before* the accuracy correlation is calculated, meaning that individual perceivers do not earn accuracy scores; rather, one correlation expresses the accuracy of the whole group of perceivers. Fifty-one retrieved results were not included in the current database because they calculated accuracy in this way. For those 51 results, the overall *pi* was .61 (range = .38–.80), no different from the analogous *pi* for studies that calculated correlational accuracy for individual perceivers (see text). This is not what one would expect based on psychometric theory, as there should be benefits associated with aggregation due to the reduction of measurement error. A more controlled comparison was available in three studies where both of these methods were calculated for the same stimuli. In Watson (1989), individual and group accuracy (*pi*) for judging the Big 5 traits were .57 and .59, respectively. In Zebrowitz, Hall, Murphy, and Rhodes (2002), a meta-analysis of previously published studies of judging intelligence yielded both kinds of calculation; here the individual and group *pi*s were .60 and .65, respectively. Reynolds and Gifford (2001) examined accuracy of judging intelligence in a laboratory study and also found individual and group *pi*s of .60 and .65, respectively (averaged across three channels of presentation). Thus, there is some, though not very strong, evidence that aggregation brings an increase in accuracy. Because results calculated using the group correlation approach should, in principle, be larger than analogous correlations based on individual perceivers, the group correlations should not be combined with, nor compared directly to, accuracy based on any method that is based on individual perceivers (Bernieri, Gillis, Davis, & Grahe, 1996).

[3] Another correlational method that was excluded assessed accuracy by correlating perceivers' ratings with a criterion using *perceivers* as the sampling units (i.e., *N* for the correlation is the number of perceivers in the group; called nomothetic by Kolar, Funder, & Colvin, 1996). For example, Hall, Horgan, Stein, and Roter (2002) assessed patients' accuracy in knowing how liked they were by their physicians by correlating patients' ratings of how much their physician liked them with their respective physicians' actual liking for them. This method, like that described in Footnote 2, yields no individual accuracy scores but rather one correlation that describes accuracy for the entire group (but the *N* for the correlation is perceiver-target pairs, not targets as described in Footnote 2).

the number of potentially eligible results in the published literature is unknown but undoubtedly very large, we did not attempt to exhaustively locate all reports of mean accuracy on an interpersonal sensitivity test. Instead, we reasoned that a representative and unbiased database was achieved by (a) comprehensively including the key standard tests, and (b) including a large number of results based on nonstandard tests using selection criteria that had nothing to do with a given study's accuracy level.

2.3. Inclusion and exclusion criteria

Inclusion/exclusion criteria were: (1) only published studies or test manuals were used, (2) only works in English were used, (3) to minimize variance due to factors unrelated to the test used, studies with perceivers younger than high school, older than approximately 65 years, or identified as nontypical (e.g., clinically diagnosed groups including learning disabled, alcoholic, institutionalized, or mentally ill) were excluded, and (4) for the same reason, studies with either perceivers or targets (expressors) who were mainly from non-Western countries were excluded.

With respect to instrument characteristics, scoring, and reporting, the following criteria were used: (1) any content domain of interpersonal sensitivity except for lie detection was included (studies of lie detection were excluded because Bond & DePaulo's (2006) meta-analysis comprehensively covered this topic), (2) the scoring system was based on proportion correct or on correlation of the across-stimuli type for individuals, (3) studies based on free response were excluded because it was not clear what the guessing level would be (e.g., Ickes, 2001), and (4) studies in which accuracy was measured in people who made their judgments of a partner during or after a live interaction with that person were excluded because, in such a paradigm, judgment accuracy by one person is confounded with the other person's accuracy of expression (e.g., Ickes, 2001; Snodgrass, 1985).

2.4. Definition of sampling units

If a given published source contained more than one independent study or perceiver subgroup, these were counted as separate entries (results) in the database. An exception is for perceiver sex, for which the averaged performance of men and women was entered, when it was reported or could be calculated, in order to minimize variance due to sex. However, if a study consisted of only one sex, that entry was included.

If the same group of perceivers was given more than one accuracy test (which did not happen often), these were counted as separate entries. Because of this, the entries in the database are not completely independent in terms of the perceivers whose accuracy is measured. However, because the focus of interest is not on perceivers but on tests, the small amount of nonindependence thus introduced was ignored. This was additionally not a problem because no significance testing was done as explained below.

2.5. Coded variables

2.5.1. Study characteristics

The characteristics that were coded for each result were: (1) specific test (PONS version; DANVA version; IPT version; POFA; CARAT; JACFEE; JACBART; specific nonstandard test), (2) channel (face; voice; body; face and body; face, body, and voice; other), (3) dynamism (still; moving), (4) content domain (intelligence; personality; emotions/affect; status; other) (specific emotions such as anger were not separately analyzed, and if the domain was personality, the specific trait was noted), (5) stimulus exposure duration, and (6) whether the researcher stated that the stimuli were preselected in a way that the present authors thought was likely to influence accuracy (no, yes, can't tell).

This issue of preselection has important implications for comparing accuracy across tests because the difficulty level of a test is easily altered at the design stage. For example, the PONS test was designed to have a proportion accuracy of about .75 (midway between the guessing level of .50 and perfect accuracy of 1.00), in the belief that scores midway between guessing and perfect accuracy would be optimal for revealing individual differences, a goal that was met by varying the duration of the stimulus exposure during pilot testing until the desired level of accuracy was achieved. As another example, the DANVA faces test was designed to have a proportion accuracy of .80–.90 and this was accomplished by selecting photographs from a larger corpus. Finally, the POFA, JACFEE, and JACBART all included, by design, only prototypic basic facial expressions of emotion known through pretesting and/or anatomical measurement to be very easily judged. In contrast to these examples of stimuli that were selected to have a desired difficulty level, a test developer might have created a test containing relatively unselected stimuli by, for example, using all of the stimuli at hand or selecting stimuli from a larger corpus on grounds other than ease or difficulty of judgment (e.g., Gifford, 1994).

2.6. Calculation of pi

2.6.1. Proportion-correct scoring

One source of incompatibility between tests stems from differing numbers of response options on the answer sheet. As indicated earlier, a given proportion correct does not mean the same thing if the number of response options differs. To

solve this problem, we applied the one-sample effect size estimator called the Proportion Index, or *pi* (Rosenthal & Rubin, 1989). *Pi* converts any mean accuracy that originates as a proportion (or percentage, but we will refer to proportion in this article), no matter how many response options each item had, to its equivalent proportion were it to have been based on two options. Thus, for any multiple-choice instrument, performance can be expressed on the common metric, *pi*.

If a test has two response options per item, the chance level of accuracy expressed as a proportion is .50 (i.e., people should be right half the time if they guess). No conversion is required for such a test because the proportion correct is already *pi*. But if the test has, say, four response options and an associated guessing level of .25, the *pi*-converted guessing level would now be .50, reflecting the fact that .25 accuracy on a four-option test is equivalent to .50 accuracy on a two-option test, and the obtained proportion correct would change accordingly. Below is the formula supplied by Rosenthal and Rubin (1989), where P refers to the obtained proportion correct and k refers to the number of response options:

$$pi = (P(k-1))/(1+(P(k-2))).$$

Applying this formula, a proportion correct of .78 on the four-option test would convert to a *pi* of .91, reflecting the fact that .78 on a four-option test is a much higher level of accuracy than the same .78 would be on a two-option test. After conversion to *pi*, the guessing level of the test is now .50 instead of whatever value it initially had.

2.6.2. Correlational scoring

To convert a mean accuracy score that is based on correlational scoring to *pi*, we applied the logic of the Binomial Effect Size Display (BESD; Rosenthal & Rubin, 1982; see Bosch, Steinkamp, & Boller, 2006, for a similar application). The BESD is a device for depicting the empirical meaning of a correlation coefficient in terms of proportions in a population. The relevance to tests of interpersonal sensitivity is obvious because we would like to estimate the proportion correct for a test that was scored in terms of a correlation. To calculate the BESD, one hypothesizes equal marginal frequencies in a 2×2 table in which the two dimensions are the two variables in question. The correlation is then translated into the cell proportions of the BESD using the formula:

$$\text{BESD values} = .50 \pm (r/2).$$

As an example, consider the accuracy correlation of $r = .28$ from the study of Lippa and Dietz (2000) described above. The two dimensions of the 2×2 BESD table would be dichotomized versions of these two continuous variables (i.e., perceived and actual extraversion), with equal marginal frequencies: perceived extraversion (low, high) and actual extraversion (low, high). The two cells representing accuracy are the "high–high" and "low–low" cells (i.e., those showing concordance between judgment and criterion). In the 2×2 BESD table, these two cells will always have the same proportion, and the other two cells will always have the same (complementary) proportion. Application of the BESD formula in the Lippa and Dietz study reveals that the proportion in the accuracy cells is .64, which can be considered the success or accuracy rate. This figure is an estimate of the proportion accuracy that would have resulted directly from scoring responses as right or wrong without going through the steps of calculating the correlation and applying the BESD formula.

In studies reporting correlationally scored tests, therefore, *pi* can be estimated. These estimates cannot be validated empirically because there were no actual proportion-correct scores reported in the original studies. However, it can be shown that whenever the null (chance) level for the proportion correct is .50 and the row and column marginals are equal, the BESD-estimated proportion correct and the actual proportion correct will coincide exactly, as demonstrated in Appendix B. The appendix shows a hypothetical test involving a two-option multiple choice format. Scoring the test as proportion correct (which is already *pi* because it is a two-choice situation) yields a proportion of .60 (6/10 of the answers are correct). Calculating accuracy as a correlation directly from the raw data (by correlating binary values of the responses and stimuli together) yields $r = .20$, and, finally, estimating the proportion correct (*pi*) from this correlation using the BESD formula yields the same proportion of .60.

Because *pi* and the correlation are linear transformations of each other, it follows that they are correlated perfectly with one another ($r = 1.00$) and that both will produce the same results when entered into further statistical analysis. For example, the single-sample *t*-test to see whether mean accuracy is significantly greater than chance will be the same regardless of which metric is used. Appendix C illustrates this equivalence.

A feature of the BESD is that it does not exactly reproduce the original cell proportions if the original table (if there was one) had highly uneven marginals, that is, the rows and/or columns deviated substantially from a 50:50 split (Hsu, 2004; Thompson & Schumaker, 1997). How much of a problem this is for reaching valid conclusions has been debated, though it is agreed that small deviations have little impact (Crow, 1991; McGraw, 1991; Rosenthal, 1990, 1991; Strahan, 1991). The purpose of the BESD as stated by its developers (see Rosenthal, 1991) is, of course, not to reproduce the original cell proportions but to depict the cell proportions that would be associated with the obtained *r* if there were equal marginal distributions.

In the current context, this debate is moot because, in applying the BESD to an accuracy correlation, we are not able to compare the resulting estimated proportion correct with the actual proportion correct because in such studies perceivers never made any categorical judgments that could directly yield a proportion correct score. Instead, in applying the BESD as we do in this article, the goal is to estimate what the proportion correct *would have been* if the same study had been conducted in categorical format with equal row and column marginals.

Table 1
Description of database (N = 109)

Variable		Result
Scoring method	Proportion correct	55
	Accuracy correlation	54
Type of test	Standard	14
	Nonstandard	95
Cue channel	Face, body, and voice	47
	Face only	24
	Voice only	20
	Face and body	13
	Body only	1
	Other	4
Item preselection	No	71
	Yes	27
	Can't tell	11
Dynamism	Moving	90
	Still	19
Domains	Emotions	52
	Personality	34
	Intelligence	9
	Status	2
	Other	12
Stimulus duration (min)[a]		M = 3.42 (range .001–45)

Note. Result is stated as a frequency unless indicated otherwise.
[a] N = 83 (could not always be coded).

2.7. Coder reliability

The first two authors and the last author divided the coding between them. Reliability checks between the first two authors (n = 13 results) and between the second author and the last author (n = 15 different results) both had a median percentage agreement of 100% (range = 77–100% and 73–100%, respectively).

2.8. Analysis

Results are presented descriptively only, with no significance testing. The goal of the research was to describe accuracy for different instruments, domains, and so forth, in the obtained sample of studies, not to formally test hypotheses about such differences. A further reason not to conduct statistical tests was that the amount of data contributing to a given *pi* varied dramatically, as explained above. One *pi* might be based on the average across numerous large perceiver samples, while another might be based on a single small perceiver sample. The resulting very great variance in standard errors would present a serious problem for significance testing.

3. Results

The search yielded 109 results. Overall, the mean *pi* was .72 (range = .49–.98), indicating that accuracy was approximately midway between guessing (.50) and perfect accuracy (1.00). There were approximately equal numbers of results using proportion correct versus correlational scoring. Other methodological features are shown in Table 1.

3.1. Pi based on correlational scoring

For the 54 results using the correlational method, the mean *pi* was .62 (range = .49–.78). Table 2 shows *pi* according to channel and dynamism. Only for six results (all on judging intelligence from the face; see table note) were photographs used as stimuli; all other results were based on moving stimuli. Results were rather homogeneous across channels, with the face being most accurate and the voice least accurate.[4]

Table 3 breaks down the correlational results according to content domain. There was substantial variation, from a low of *pi* = .49 (slightly below chance) for judging how much in love two targets were to a high of *pi* = .71 for judging teacher effectiveness (one result each). The second highest accuracy was for judging extraversion (*pi* = .70, n = 7 results).

[4] Excluded from the present analyses was a hybrid scoring method in which the accuracy correlation was calculated across the N of constructs X N of targets (Realo et al., 2003).

Table 2
Pi based on correlational scoring (by channel and dynamism)

Channel	Still	Moving
Face	.60(1)[a]	.68(4)
Face and body		.65(3)
Voice		.58(7)
Face, body, and voice		.62(48)
Transcript, face, and body		.62(1)

Note. Only results based on individual across-stimuli accuracy correlations are included. Numbers in parentheses indicate how many results were available for each type.
[a] One entry in database, but based on meta-analysis of six published studies reported in Zebrowitz et al. (2002).

Table 3
Pi based on correlational scoring (by construct judged)

Construct	*pi*
Intelligence	.61(9)[a]
Extraversion	.70(7)
Agreeableness	.56(6)
Conscientiousness	.63(6)
Neuroticism	.59(7)
Openness	.58(6)
Rapport	.61(5)
Positive affect	.68(4)
Masculinity–femininity	.59(1)
Dominance	.57(1)
Teacher effectiveness	.71(1)
Love	.49(1)

Note. Only results based on individual across-stimuli accuracy correlations are included. Numbers in parentheses indicate how many results were available for each type.
[a] Fourteen studies if those described in note to Table 2 are individually counted.

Tests scored correlationally were rarely based on preselected stimuli; indeed, 94% of this kind of study used unselected stimuli according to our coding. Accuracy was greater for the three results that used preselected stimuli (*pi* = .71) than for those that did not (*pi* = .61, *n* = 50 results).

Stimulus exposure durations ranged from 5 s to 45 min (the latter for studies where perceivers and targets spent time together in minimal interaction). Accuracy was weakly associated with exposure duration, $r(51) = -.20$ (for one study, exposure duration could not be ascertained). Accuracy levels for the shortest and longest durations were very similar and there were no trends evident for durations in between. One of the articles in the database presented a more controlled analysis of stimulus durations (Carney, Colvin, & Hall, 2007). Those investigators systematically varied the exposure durations of the same set of stimuli while perceivers judged agreeableness, extraversion, neuroticism, openness, conscientiousness, and intelligence. Averaging over these constructs, there was an increase in accuracy across exposure durations of 5 s, 20 s, 45 s, 1 min, and 5 min (see Carney et al. for more detail).[5]

3.2. Pi based on proportion correct

For the 55 results based on proportion correct, the mean *pi* was .81 (range = .60–.98). The tests considered standard (see Method and Appendix A for list) produced almost the same overall level of accuracy (*pi* = .83, *n* = 14 results) as the nonstandard tests (*pi* = .81, *n* = 41 results). Table 4 shows *pi* broken down by standard versus nonstandard tests, dynamism, and channel. For the most part, a comparison of standard versus nonstandard tests revealed similar levels of accuracy. A deviation from this trend occurred for the upper face channel (i.e., eyes), where the eyes from the POFA produced much higher accuracy than the Eyes Test developed by Baron-Cohen, Jolliffe, Mortimore, and Robertson (1997). However, this is not surprising considering that the POFA eyes were from highly prototypical, basic emotional facial expressions known to be easily judged, whereas the Eyes Test represented a number of different affective states, many of them not basic (e.g., "concerned") and of unknown a priori judgeability.

[5] Ambady and Rosenthal (1992), in a meta-analysis of studies in which meaningful outcomes were predicted from ratings of excerpts of behavior, found that the duration of the excerpt did not moderate those correlations. Though not measuring accuracy of judgment per se, the studies in their review were testing the informational value of the excerpts and therefore their result is relevant to the present results.

Table 4
Pi based on proportion correct (by channel, dynamism, and standard versus nonstandard tests)

Channel	Still		Moving	
	Test	*pi*	Test	*pi*
Face	DANVA	.92	CARAT	.75
	JACFEE	.97	Nonstandard	.78(7)
	JACBART	.89		
	POFA	.98		
	Nonstandard	.92(7)		
Upper face	POFA	.92		
	Nonstandard	.79(1)		
Lower face	POFA	.83		
Face and body	PONS	.76	PONS	.77
	Nonstandard	.77(2)	Nonstandard	.77(6)
Body	DANVA	.85		
Voice			PONS	.68
			DANVA	.89
			Nonstandard	.79(11)
Face, body, and voice			PONS	.79
			IPT	.68
			Nonstandard	.78(7)

Note. See text for explanations of abbreviated names for standard tests. Numbers in parentheses indicate how many results were included for nonstandard tests.

For vocal cues, though the nonstandard tests were very similar to the averaged standard tests (PONS and DANVA), in fact those two tests were very discrepant from each other (Table 4). Two explanations seem possible. First, the PONS uses content-masking techniques (electronic filtering and random splicing) that may make voice samples intrinsically harder to judge than the standard-content method used in many studies including the DANVA (Wallbott & Scherer, 1986) (in the standard-content method, expressors read or say the same affectively neutral material while varying voice tone to convey different affective states). Second, the DANVA stimuli were selected to produce proportion-correct accuracy in excess of .70 (because the test has four response alternatives, this translates to a *pi* of .88) whereas the PONS vocal clips were not selected for high accuracy. The full-length PONS test, which includes those vocal clips, was designed to have midrange accuracy (*pi* of approximately .75), but this goal was not applied to the vocal clips in particular, and in any case selection for a *pi* of approximately .75 is a less extreme selection than that used with the DANVA. Therefore, stimulus selection factors and possibly method of content masking may account for the variation in accuracy for vocal cues across these two tests.

Striking in Table 4 is the very high accuracy of judging facial expressions from photographs. Again, stimulus selection probably plays an important role. The POFA, JACFEE, and JACBART all used highly prototypical, basic, discrete expressions of emotion, preselected to be accurately judged or determined by anatomical coding to show the desired emotions. The DANVA faces test contains basic emotional expressions chosen to produce proportion-correct accuracy in excess of .80 (because this test has four response alternatives, this translates to a *pi* of .92). For their part, the remaining tests using facial photographs also used basic and/or posed, discrete emotions that one would expect to show similarly high levels of accuracy.

Table 4 also compares still to moving stimuli for the visual channels. Mostly, the differences were minor, with the only sharp contrast occurring for the face, where still faces had higher levels of accuracy than moving faces. Inspection of the types of studies that used still versus moving faces suggests that the nature and selection of the stimuli explain this difference rather than dynamism per se. On the CARAT test, perceivers guess which emotionally evocative slide is being viewed by the targets whose faces they see. Because the targets did not know their faces were being recorded, this task therefore involves facial expressions that are spontaneous and probably nonprototypical. Finding a lower level of accuracy than typically found in the still presentation format (i.e., photographs with posed expressions) is therefore not a surprise. The same interpretation can be applied to the other moving face studies because five of the seven involved the same slide- (or film-) viewing paradigm (though not the CARAT per se); the sixth study asked perceivers to guess what kind of task the targets were doing; and the seventh asked perceivers to guess what kind of emotional situation the targets were thinking about. All of these tests involve judgments of spontaneous, nonbasic, nonprototypical expressions for which we might expect lowered levels of accuracy.

The great majority of the multiple-choice tests measured accuracy in judging emotion (50/57 or 88% of the results). When the few non-emotion studies were removed, the results in Table 4 changed very little. An interesting case is the IPT, which was classified as non-emotion because its content covers kinship, intimacy, competition, status, and deception. Though all of these may produce emotional expressions, perceivers are not asked to judge emotion per se. Accuracy for the IPT was lower than found for other tests showing face, body, and voice, and the same was true of the similarly designed Social Interpretations Task, the precursor to the IPT (SIT; Archer & Akert, 1977). The IPT's and SIT's lower accu-

racy could be due to their relatively non-emotional content or to their relatively spontaneous, unscripted nature (i.e., people talking about themselves or interacting with others). Certainly, if emotions are involved in the IPT and SIT they are not of the basic, prototypical type.

The PONS, like the IPT, asks for situational inferences (e.g., is the target person ordering food in a restaurant or talking to a lost child), but it also asks perceivers to make explicitly emotional inferences (e.g., is the target expressing jealous anger, or expressing motherly love). We classified the PONS as emotional because, on balance, the items were more clearly about affect than was the case with the IPT. As Table 4 shows, accuracy on the PONS was well below that found for still facial photograph studies. In all likelihood, this is due to the PONS developers' choice to set overall accuracy (*pi*) at the midrange of about .75 (see above), though it is also the case that the PONS stimuli, while intentionally encoded, were relatively spontaneous and unscripted (Rosenthal et al., 1979). The fact that the moving and still versions of the face and body PONS did not differ in accuracy (Table 4) suggests again that dynamism, per se, is not what matters but rather the nature of the content being shown (e.g., prototypical, basic emotional expressions versus more spontaneous, nonbasic emotional expressions).

The duration of stimulus exposure was not related in a linear fashion to *pi*, $r(28) = -.10$. This analysis is not definitive because exposure duration could not be ascertained in 25 studies (e.g., sentences of unknown duration, self-paced viewing of photographs, stimuli of highly varying lengths in the same test). But it can be noted that for prototypical facial expressions shown for 1 s or less, accuracy was still very high (*pi* = .93, *n* = 3 results).

3.3. Combined analyses

As the preceding sections made clear, the important comparison between accuracy for judging personality versus affect was complexly related to the methods used. Table 5 shows just the 84 results for personality and affect (omitting intelligence). The first thing that is very evident in this table is the gaps—results for judging personality were limited to correlational scoring with unselected stimuli. Although we cannot be sure that there are truly no unretrieved results in the empty cells, it is probably safe to say that such results are not common. The table has more informative data for judgments of affect. Selected stimuli produced higher accuracy than unselected stimuli, and results based on proportion correct produced higher accuracy than results based on correlational scoring. Accuracy was practically the same for judging personality and for judging affect in studies using the correlational method with unselected stimuli.

4. Discussion

By applying simple scoring conversions, accuracy levels from across the interpersonal sensitivity literature could be combined and compared using the common metric *pi* (Rosenthal & Rubin, 1982, 1989). Overall, across all 109 results, the mean *pi* was in the midrange of possible values (.72). Thus, interpersonal sensitivity based on "thin slices" of behavior (Ambady & Rosenthal, 1992) can be dramatically higher than chance (.50). As noted earlier, the present review did not include accuracy in distinguishing truth from lies. Bond and DePaulo (2006) conducted a summary of lie detection accuracy across 292 perceiver samples in which proportion correct was the scoring method and found an overall *pi* of only .54. Though even this relatively low figure was highly significantly above the guessing level of .50, it is clear that accuracy of lie detection as measured in laboratory studies is far lower than accuracy of judging the states and traits reviewed in the present article.

Accuracy varied with the extent to which the stimuli were preselected for specific accuracy levels. The highest accuracy was for photographed facial expressions, which were mostly prototypical, posed, basic, discrete emotions that were typically picked precisely because they were very good exemplars of the desired expressions. Studies that used less selected expressions and behavior samples showed lower levels of accuracy. These less selected stimuli were also much more likely to be spontaneous and to be presented in dynamic, multichannel modalities. These generalizations also apply to studies of lie detection (Bond & DePaulo, 2006), where accuracy is low, researchers are very unlikely to employ preselection of stimuli, and the stimuli are dynamic and multichannel. Because stimuli of this kind have better ecological validity than posed prototypical photographs, the accuracy levels found with the more spontaneous kinds of stimuli are probably a better approximation of accuracy in daily life. Ecological validity does not, however, necessarily reflect on the predictive validity of a test (a topic we do not address in this article).

Table 5
Pi in relation to stimulus selection, content domain, and type of scoring

Content domain and type of scoring	Selected stimuli	Unselected stimuli
Affect		
Proportion correct	.85(21)	.79(19)
Correlation	.71(3)	.59(7)
Personality		
Proportion correct	—	—
Correlation	—	.61(34)

Note. Numbers in parentheses indicate how many results were available for each type.

Accuracy based on proportion correct was higher than accuracy based on correlational scoring. Interpreting this must take into account the high degree of confounding between method and content: proportion-correct scoring was mostly used for judging states (primarily discrete emotions), while the correlationally scored tests were mostly used for judging traits (primarily personality and intelligence). A possible conclusion, therefore, is that traits are more difficult to judge than affective states. However, among correlationally scored studies there was no difference in accuracy between judging personality and judging affect (Table 5). But the gaps that were so evident in Table 5, especially for correlational studies using preselected stimuli, make any conclusions tentative. Whether there would be a difference between judging traits and affective states for correlationally scored studies using preselected stimuli, or for multiple-choice tests of both kinds of content, cannot be assessed on the basis of the present database.

Any insights gained from the current analyses are simply hypotheses to be tested under more controlled circumstances. Such future studies will benefit from the application of the scoring conversions described in the present article. Because the between-studies comparisons we made could not be well controlled, it is important for investigators to make comparisons (e.g., between accuracy for judging personality versus affect) based on the same corpus of behavior samples, thus controlling for stimulus duration, channel, selection, and other potentially confounding factors.

The present database has limitations. The two scoring methods discussed—proportion correct and the across-stimuli correlational method—do not exhaust all the scoring methods that researchers use (e.g., see Footnotes 1–3). Also, the database was not an exhaustive compilation of studies of interpersonal sensitivity. However, even if one were to assemble a much larger and more comprehensive collection of results, there would still be confounded variables and ambiguities. For many research questions, within-study comparisons that can effectively control confounding are much to be preferred. However, the present summary did reveal the potential value of having a common metric of measurement in furthering the development of a theory of interpersonal sensitivity (Zebrowitz, 2001).

Appendix A. Standard tests: Description and data source information for *pi*

A.1. Full-length Profile of Nonverbal Sensitivity (PONS: face, body, and voice)

The PONS is a 220-item test containing 2-s clips of all combinations of face, body, electronically filtered speech, and random-spliced speech (total of 11 channels) of an adult female expressor deliberately portraying 20 different affective situations. *Pi* was based on the mean of 62 non-children samples reported in Rosenthal et al. (1979).

A.2. Face and body video PONS

The face and body video PONS consists of the 20 face-only and the 20 body-only items from the full-length PONS, administered as a separate test. *Pi* was based on one non-children sample reported in Rosenthal et al. (1979).

A.3. Face and body still photographs PONS

The face and body still photographs PONS consists of one still frame from each of the 40 items in the face and body video PONS. *Pi* was based on one non-children sample reported in Rosenthal et al. (1979).

A.4. Voice PONS

The voice PONS consists of the 40 voice-only items from the full-length PONS, administered as a separate test. *Pi* was the mean of two non-children samples reported in Rosenthal et al. (1979).

A.5. Face Diagnostic Analysis of Nonverbal Accuracy (DANVA)

The DANVA-2-AF test consists of 24 posed photographs of adult facial expressions of four basic emotions. *Pi* was based on norm data reported for perceivers ages 15–60 (Nowicki, no date).

A.6. Voice DANVA

The DANVA-2-AV test consists of 24 posed vocal clips consisting of a standard-content sentence expressed as four basic emotions. *Pi* was based on norm data reported for perceivers ages 15–60 (Nowicki, no date).

A.7. Body Postures DANVA

The DANVA-2-POS test consists of 32 posed photographs of bodily expressions of four basic emotions. *Pi* was based on one non-children sample reported in Pitterman and Nowicki (2004).

A.8. Interpersonal Perception Task (IPT)

The IPT items are audiovisual excerpts of individual expressors talking or small groups of expressors interacting. Content covers intimacy, kinship, competition, status, and deception. Although deception was not included in the present review, the IPT was included because only a small minority of items concern deception. *Pi* was the mean of norm data for the IPT-15 (15 items; Costanzo and Archer, no date) and the IPT-30 (30 items; Costanzo and Archer, 1989). (*Note:* IPT items have either two or three response alternatives. The latter were converted to *pi* and then averaged with the two-response items to calculate *pi* for the total test.)

A.9. Communication of Affect Receiving Ability Test (CARAT)

The CARAT consists of 30 spontaneous facial video clips of adult targets watching four categories of emotionally evocative slides. The standard scoring method for the CARAT is proportion correct (i.e., perceivers guess which slide the target was watching); for this, *pi* was the mean of 11 studies that used this method. Accuracy can also be scored using the across-stimuli accuracy correlation, based on correlating pleasantness ratings made by targets with corresponding ratings made by perceivers. When this method was used (in four studies), the entries were left separate because this method was considered nonstandard.

A.10. Japanese and Caucasian Facial Expressions of Emotion (JACFEE)

The JACFEE consists of 56 photographs of prototypical facial expressions of seven basic emotions posed by adult expressors. *Pi* was based on three samples reported in the instrument's manual (Matsumoto and Ekman, 1988). (For the JACFEE we used only data for Caucasian targets judged by Caucasian perceivers.) Two additional published studies that used the JACFEE had very similar results to the norm data and were not included in the quantitative summaries.

A.11. Japanese and Caucasian Brief Affect Recognition Test (JACBART)

The JACBART consists of 56 extremely short (1/5 s or shorter) presentations of JACFEE expressions, each "sandwiched" by neutral expressions by the same expressor. *Pi* was the mean of seven studies that used categorical scoring (Matsumoto et al., 2000).

A.12. Pictures of Facial Affect (POFA)

The POFA consists of 110 facial photographs of adult targets who posed specific muscle configurations and whose expressions were judged very consistently in pretests. *Pi* was based on data in the test manual (Ekman, 1976). (For the POFA we also located 10 studies that used the full-face version of that instrument, but we did not enter them into the database because their mean *pi* was very similar to the POFA normative data reported in the manual.)

Appendix B. Illustration of conversions involving proportion correct, correlational index of accuracy, and binomial effect size display (BESD)

A test shows sad or angry facial expressions and participants choose whether each face is sad or angry. Shown are hypothetical test data for a participant.

Item	Participant's answer	Criterion (correct answer)	Scored item
1	Sad	Sad	Correct
2	Angry	Angry	Correct
3	Sad	Sad	Correct
4	Angry	Angry	Correct
5	Angry	Sad	Incorrect
6	Sad	Sad	Correct
7	Angry	Angry	Correct
8	Sad	Angry	Incorrect
9	Angry	Sad	Incorrect
10	Sad	Angry	Incorrect

Proportion accuracy (*pi*) = .60 (6 out of 10 correct).
Correlational index = .20 (correlation of answers coded 0, 1 with criterion coded 0, 1 where 0 = sad, 1 = angry; $N = 10$).
Proportion accuracy (*pi*) based on correlational index = .60, using BESD formula in text.

Appendix C. Single-sample *t*-tests for proportion correct and correlational index

Seven hypothetical participants took the test shown in Appendix B. The participant depicted in Appendix B is the first participant in the display that follows.

Participant	Proportion correct	Correlational index
1	.60	.20
2	.65	.30
3	.66	.32
4	.54	.08
5	.55	.10
6	.65	.30
7	.70	.40
Mean	.62	.24
t(6)	5.37[a]	5.37[b]
p	<.002	<.002

[a] Single-sample *t*-test against chance level of .50.
[b] Single-sample *t*-test against chance level of .00.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*, 256–274.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). A further advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger Syndrome. *Journal of Child Psychology and Psychiatry, 38*, 813–822.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*, 214–234.

Bosch, H., Steinkamp, F., & Boller, E. (2006). Examining psychokinesis: The interaction of human intention with random number generators—A meta-analysis. *Psychological Bulletin, 132*, 497–523.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York, NY: Russell Sage Foundation.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177–193.

Crow, E. L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology?". *American Psychologist, 46*, 1083.

Davis, M. H., & Kraus, L. A. (1997). Personality and empathic accuracy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 144–168). New York: Guilford.

Funder, D. C. (2001a). *The personality puzzle* (2nd ed.). New York: W.W. Norton.

Funder, D. C. (2001b). Three trends in current research on person perception: Positivity, realism, and sophistication. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 319–331). Mahwah, NJ: Lawrence Erlbaum Associates.

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (in press). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior*.

Hall, J. A., & Bernieri, F. J. (Eds.). (2001). *Interpersonal sensitivity: Theory and measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hall, J. A., Bernieri, F. J., & Carney, D. R. (2005). Nonverbal behavior and interpersonal sensitivity. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 237–281). Oxford: Oxford University Press.

Hall, J. A., Horgan, T. G., Stein, T. S., & Roter, D. L. (2002). Liking in the physician–patient relationship. *Patient Education and Counseling, 48*, 69–77.

Hsu, L. M. (2004). Biases and success rate differences shown in binomial effect size displays. *Psychological Methods, 9*, 183–197.

Ickes, W. (2001). Measuring empathic accuracy. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 219–241). Mahwah, NJ: Lawrence Erlbaum Associates.

Jenness, A. (1932). The effects of coaching subjects in the recognition of facial expression. *Journal of General Psychology, 7*, 163–178.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*, 770–814.

Kenny, D. A., & Winquist, L. (2001). The measurement of interpersonal sensitivity: Consideration of design, components, and units of analysis. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 265–302). Mahwah, NJ: Lawrence Erlbaum Associates.

McGraw, K. O. (1991). Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?". *American Psychologist*, 1084–1086.

Murphy, N. A., & Hall, J. A. (in preparation). Cognitive ability and interpersonal sensitivity: A meta-analysis.

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior, 18*, 9–34.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (revised ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775–777.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166–169.

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.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. *Psychological Bulletin, 115*, 102–141.

Snodgrass, S. E. (1985). Women's intuition: The effect of subordinate role on interpersonal sensitivity. *Journal of Personality and Social Psychology, 49*, 146–155.

Thompson, K. N., & Schumaker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Educational and Behavioral Statistics, 22*, 109–117.

Vernon, P. E. (1933). Some characteristics of the good judge of personality. *Journal of Social Psychology, 4*, 42–58.

Zebrowitz, L. A. (2001). Groping for the elephant of interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 333–350). Mahwah, NJ: Lawrence Erlbaum Associates.

Further reading

Note: Studies that were included in the quantitative analyses reported in the text and footnotes are marked with ∗ if they were scored by proportion correct and ∗∗ if they were scored as a correlation. Some studies contributed both kinds of scoring and/or multiple results to the database. Among correlational studies, those marked with "R" (for across-ratings correlation; see Footnote 1) and "G" (for group-level correlation; see Footnote 2) were not included in the main analyses presented in Tables 1–5.**Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55*, 387–395 (G).

**Aloni, M., & Bernieri, F. J. (2004). Is love blind? The effects of experience and infatuation on the perception of love. *Journal of Nonverbal Behavior, 28*, 287–295.

*,**Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology, 83*, 947–961.

*Archer, D., & Akert, R. M. (1977). Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of Personality and Social Psychology, 35*, 443–449.

*Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*, 614–636.

*Barnes, M. L., & Sternberg, R. J. (1989). Social intelligence and decoding of nonverbal cues. *Intelligence, 13*, 263–287.

*Bastone, L. M., & Wood, H. A. (1997). Individual differences in the ability to decode emotional facial expressions. *Psychology: A Journal of Human Behavior, 34*, 32–36.

**Bernieri, F. J., Gillis, J. S., Davis, J. M., & Grahe, J. E. (1996). Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology, 71*, 110–129.

**Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65*, 546–553 (G).

*,**Buck, R. (1976). A test of nonverbal receiving ability: Preliminary studies. *Human Communication Research, 2*, 162–171.

*Buck, R., & Lerman, J. (1979). General vs. specific nonverbal sensitivity and clinical training. *Human Communication, Summer*, 267–274.

*,**Buck, R., Losow, J. I., Murphy, M. M., & Costanzo, P. (1992). Social facilitation and inhibition of emotional expression and communication. *Journal of Personality and Social Psychology, 63*, 962–968.

**Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*, 1054–1072.

*Costanzo, M., & Archer, D. (1989). Interpreting the expressive behavior of others: The Interpersonal Perception Test. *Journal of Nonverbal Behavior, 13*, 225–245.

*Costanzo, M., & Archer, D. (no date). *The Interpersonal Perception Task-15 (IPT-15): A guide for researchers and teachers*. Berkeley, CA: University of California Extension.

*Danish, S. J., & Kagan, N. (1971). Measurement of affective sensitivity: Toward a valid measure of interpersonal perception. *Journal of Counseling Psychology, 18*, 51–54.

*DiMatteo, M. R., Hays, R. D., & Prince, L. M. (1986). Relationship of physicians' nonverbal communication skill to patient satisfaction, appointment noncompliance and physician workload. *Health Psychology, 5*, 581–594.

*Ekman, P. (1976). *Pictures of Facial Affect*. San Francisco: University of California Medical Center.

**Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology, 69*, 656–672 (R).

**Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology, 66*, 398–412 (G).

*Goldstein, N. E., & Feldman, R. S. (1996). Knowledge of American Sign Language and the ability of hearing individuals to decode facial expressions of emotion. *Journal of Nonverbal Behavior, 20*, 111–122.

**Grahe, J. E., & Bernieri, F. J. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior, 23*, 253–269.

*Halberstadt, A. G. (1983). Family expressiveness styles and nonverbal communication skills. *Journal of Nonverbal Behavior, 8*, 14–26.

*Hall, J. A., Halberstadt, A. G., & O'Brien, C. E. (1997). "Subordination" and nonverbal sensitivity: A study and synthesis of findings based on trait measures. *Sex Roles, 37*, 295–317.

**Hall, J. A., & Matsumoto, D. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion, 4*, 201–206 (R).

*Hall, J. A., Rosip, J. C., Smith LeBeau, L., Horgan, T. G., & Carter, J. D. (2006). Attributing the sources of accuracy in unequal-power dyadic communication: Who is better and why? *Journal of Experimental Social Psychology, 42*, 18–27.

*Hess, U., & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology, 40*, 129–141.

*Hodgins, H. S., & Belch, C. (2000). Interparental violence and nonverbal abilities. *Journal of Nonverbal Behavior, 24*, 3–24.

*Izard, C. E. (1971). *The face of emotion*. New York: Appleton-Century-Crofts.

*Keeley-Dyreson, M., Burgoon, J. K., & Bailey, W. (1991). The effects of stress and gender on nonverbal decoding accuracy in kinesic and vocalic channels. *Human Communication Research, 17*, 584–605.

*Kirouac, G., & Doré, F. Y. (1983). Accuracy and latency of judgment of facial expressions of emotions. *Perceptual and Motor Skills, 57*, 683–686.

*Kirouac, G., & Doré, F. Y. (1985). Accuracy of the judgment of facial expression of emotions as a function of sex and level of education. *Journal of Nonverbal Behavior, 9*, 3–7.

*Kiss, I., & Ennis, T. (2001). Age-related decline in perception of prosodic affect. *Applied Neuropsychology, 8*, 251–254.

**Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311–337 (R).

**Lippa, R. A., & Dietz, J. K. (2000). The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior, 24*, 25–43.

*Matsumoto, D., & Ekman, P. (1988). *Japanese and Caucasian Facial Expressions of Emotion (JACFEE) and Neutral Faces (JACNeuf)*. San Francisco: Department of Psychiatry, University of California, San Francisco.

*Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., et al (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior, 24*, 179–209.

*Mazurski, E. J., & Bond, N. W. (1993). A new series of slides depicting facial expressions of affect: A comparison with the Pictures of Facial Affect series. *Australian Journal of Psychology, 45*, 41–47.

*McKelvie, S. J. (1995). Emotional expression in upside-down faces: Evidence for configurational and componential processing. *British Journal of Social Psychology, 34*, 325–334.

*Morand, D. A. (2001). The emotional intelligence of managers: Assessing the construct validity of a nonverbal measure of "people skills". *Journal of Business and Psychology, 16*, 21–33.

**Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality, 71*, 465–493 (G).

**Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4*, 681–691 (G).

*Nowicki, S. (no date). *DANVA 2: Instruction manual for the receptive tests of the Diagnostic Analysis of Nonverbal Accuracy 2*. Atlanta: Department of Psychology, Emory University.

*Nowicki, S., & Richman, D. (1985). The effect of standard, motivation, and strategy instructions on the facial processing accuracy of internal and external subjects. *Journal of Research in Personality, 19*, 354–364.

*Phillips, L. H., MacLean, R. D. J., & Allen, R. (2002). Age and the understanding of emotions: Neuropsychological and sociocognitive perspectives. *Journal of Gerontology, 57B*, P526–P529.
*Pitterman, H., & Nowicki, S. Jr., (2004). A test of the ability to identify emotion in human standing and sitting postures: The Diagnostic Analysis of Nonverbal Accuracy-2 Posture Test (DANVA2-POS). *Genetic, Social, and General Psychology Monographs, 130*, 146–162.
*Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M., et al (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality, 37*, 420–445.
**Reynolds, D. J., Jr., & Gifford, R. (2001). The sounds and sights of intelligence: A lens model channel analysis. *Personality and Social Psychology Bulletin, 27*, 187–200 (G; also individual correlations that were included in Tables 2–4).
*Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore: The Johns Hopkins University Press.
*Sabatelli, R. M., Buck, R., & Dreyer, A. (1980). Communication via facial cues in intimate dyads. *Personality and Social Psychology Bulletin, 6*, 242–247.
*Sabatelli, R. M., Buck, R., & Dreyer, A. (1982). Nonverbal communication accuracy in married couples: Relationship with marital complaints. *Journal of Personality and Social Psychology, 43*, 1088–1097.
**Schmid Mast, M., & Hall, J. A. (2004). Who is the boss and who is not? Accuracy of judging status. *Journal of Nonverbal Behavior, 28*, 145–165 (G).
**Schmid Mast, M., Hall, J. A., Murphy, N. A., & Colvin, C. R. (2003). Judging assertiveness. *Facta Universitatis, 2*, 731–744.
*Shapiro, B. A. (1982). Relation of facial expressions and activities: A study of attensity differences in events. *Perceptual and Motor Skills, 54*, 1199–1211.
*Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30*, 526–537.
*Sternberg, R. J., & Smith, C. (1985). Social intelligence and decoding skills in nonverbal communication. *Social Cognition, 3*, 168–192.
Strahan, R. F. (1991). Remarks on the binomial effect size display. *American Psychologist, 46*, 166–169.
*Stricker, L. J., & Rock, D. A. (1990). Interpersonal competence, social intelligence, and general ability. *Personality and Individual Differences, 11*, 833–839.
*Toner, H. L., & Gates, G. R. (1985). Emotional traits and recognition of facial expression of emotion. *Journal of Nonverbal Behavior, 9*, 48–66.
*Toomey, R., Seidman, L. J., Lyons, M. J., Faraone, S. V., & Tsuang, M. T. (1999). Poor perception of nonverbal social-emotional cues in relatives of schizophrenic patients. *Schizophrenia Research, 40*, 121–130.
*Trimboli, A., & Walker, M. (1993). The CAST test of nonverbal sensitivity. *Journal of Language and Social Psychology, 12*, 49–65.
**Vogt, D. S., & Colvin, C. R. (2003). Interpersonal orientation and the accuracy of personality judgments. *Journal of Personality, 71*, 267–295 (R).
*,**Wagner, H. L., MacDonald, C. J., & Manstead, A. S. R. (1986). Communication of individual emotions by spontaneous facial expressions. *Journal of Personality and Social Psychology, 50*, 737–743.
*Wallbott, H. G. (1991). Recognition of emotion from facial expression via imitation? Some indirect evidence for an old theory. *British Journal of Social Psychology, 30*, 207–219.
*Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology, 51*, 690–699.
**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 (G; also individual correlations that were included in Tables 1–3 and Table 5).
**Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin, 28*, 238–249 (G; also individual correlations that were included in Tables 1–3).
*Zuckerman, M., DeFrank, R. S., Hall, J. A., & Rosenthal, R. (1976). Encoding and decoding of spontaneous and posed facial expressions. *Journal of Personality and Social Psychology, 34*, 966–977.