RESEARCH APPLICATIONS IN NONVERBAL BEHAVIOR

CHAPTER 6

NONVERBAL BEHAVIOR AND INTERPERSONAL SENSITIVITY

JUDITH A. HALL, FRANK J. BERNIERI, AND DANA R. CARNEY

Although questions about how people respond to others' nonverbal cues have always been central to the study of nonverbal communication, the study of individual differences in accuracy of nonverbal cue processing, or interpersonal sensitivity, is a more recent endeavor. This chapter focuses on assessment of individual differences, emphasizing the major paradigms and instruments for assessing accuracy of nonverbal cue processing, and discussing characteristics of the stimuli and judgment methodologies (e.g. what state or trait is being judged, who is being judged, what cue channels are available, whether the cues are posed or spontaneous, whether judgment is done in live interaction or from standard stimuli, what judgment format is used, what criteria and methods are used for scoring). Relative advantages of different approaches are discussed in terms of psychometric qualities, validity, and utility.

Introduction

Interpersonal sensitivity is a complex concept that is subject to many definitions and many methods of measurement. This chapter describes approaches to the measurement of accuracy in processing interpersonal, mostly nonverbal, cues. By 'cues' we mean perceptible behaviors, such as facial expressions and tones of voice, that have the capacity to give insight into the expressor's attributes or condition. We define interpersonal sensitivity as accuracy in judging the meanings of cues given off by expressors, as well as accuracy in noticing or recalling cues. We discuss methods of measuring such accuracy at both the group level (i.e. mean level of accuracy for a particular social group or in a particular experimental condition) and at the level of the individual test taker (i.e. for assessing individual differences).

One could argue that a proper definition of interpersonal sensitivity would also include behavior that is emitted in response to another's cues, on the grounds that such responses are crucial to truly sensitive interpersonal interaction (Bernieri 2001). For example, empathy defined as the ability to commiserate effectively with a sad friend can be seen as a manifestation of interpersonal sensitivity. However, such a broad definition of the construct cannot be handled in the space allowed. This chapter is limited to the study of the receptive aspect of interpersonal sensitivity. In this research, perceivers make judgments about cues or about people whose cues they see and/or hear, and such judgments are then scored for accuracy.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 238 17.6.2005 1:59am

In daily life, we are constantly processing and evaluating cues that are conveyed by others through the face, body, and voice or embodied in their appearance, and we can do so with surprising accuracy based on fairly small amounts of information (e.g. Ambady & Gray 2002; Ambady *et al.* 2001; Carney *et al.* 2004; Lippa & Dietz 1998). Unless we are very distracted, it is likely that we are continuously monitoring and processing cues emanating from the people around us, but even when we are distracted or not consciously attending, cues in the periphery of our attention are often processed unconsciously. Research shows that when people are subliminally shown different facial expressions, their behavior varies in keeping with the affective connotations of the primed cues (Murphy & Zajonc 1993). Thus, cues of which people have no awareness appear to be processed accurately. Strangers are capable of making personality judgments of each other at levels greater than chance within minutes of laying eyes on each other for the first time and without hearing each other say anything (Marcus & Leatherwood 1998). Infants respond to nonverbal cues in ways that suggest at least rudimentary comprehension, for example by noticing when the affective tone is similar or different between visual and auditory modalities (Phillips *et al.* 1990). As further evidence for infants' attunement to nonverbal cues, they mimic adults' facial expressions shortly after birth (Meltzoff & Moore 1977) and show categorical perception of faces (Bornstein & Arterberry 2003).

Animals, of course, also respond to cues conveyed by each other and by human beings. The communication of information on sex, age, health, vulnerability, attractiveness, mating readiness, affiliation and reconciliation, territoriality, dominance, and threat, conveyed by static cues such as size and by dynamic cues such as facial expressions, vocalizations, or body movements, is crucial to social life throughout the animal kingdom (de Waal 2001). The biological value of being able to make such discriminations is obvious, as life, death, and reproductive success may hang in the balance (McArthur & Baron 1983). One can hardly imagine a functional social order in which the members are not supremely sensitive to information of this sort.

Lest one assume that the cues animals can respond to are all simple, gross, or wired in, one needs only talk to dog trainers and owners to hear many accounts of their dogs' (sometimes excessive) sensitivity to each other and to humans. In social psychology, the most famous demonstration of the subtlety of nonverbal cue processing by an animal was the horse, Clever Hans. In the early 1900s in Berlin, Clever Hans attracted a wide following for his apparent ability to count, solve mathematical problems, and answer apparently any question to which he could answer by tapping his hoof the appropriate number of times, even if his owner was not the questioner and even if the questioner did not know the answer (see Spitz 1997 for an excellent account). Hans' abilities were so prodigious that many careful observers, including his owner, were persuaded that he possessed conceptual thinking.

Under attack as a fraud, Hans' owner agreed to let a commission of experts (including experimental psychologists) conduct an investigation. Their experiments led to the conclusion that Hans was not a fraud, and that he was indeed a remarkable horse. However, what made him remarkable was not that he could solve mathematical problems (he could not), but rather that he combined uncanny sensitivity to cues with a quick intelligence for learning reward contingencies. Hans could perceive

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 239 17.6.2005 1:59am

changes in the questioner's (or audience members') face, head, eyes, posture, and respiration that were often not visible to the naked human eye, and he learned how to respond to them in ways that earned him food snacks and approval. Thus, he knew that he would be rewarded if he started tapping after hearing a question and then stopped when the questioner, or someone in the audience who knew the answer, inadvertently cued him by moving his or her head by as little as one millimeter, or by some other tiny cue. The relevance of the Clever Hans phenomenon to the development of interpersonal sensitivity in human beings and to our appreciation of the ubiquity of nonverbal cue processing in everyday life cannot be overstated.

But how do we measure something as intangible as our sensitivity to each other? In Clever Hans' case, experimentation determined that he was sensitive, indeed very sensitive. But *how* sensitive? How do we put numbers on such accuracy? And how do we even define what it is we are interested in?

Overview of chapter

In this chapter, we deal both conceptually and practically with methodologies for measuring interpersonal sensitivity. On the conceptual side, we discuss definitional and methodological issues that are generic to this area of research, such as the definition of interpersonal sensitivity and the determination of scoring criteria. On the practical side, we describe specific instruments, including their psychometric characteristics, validity (as much as space permits), and utility. We describe measurement approaches in terms of characteristics of the stimuli and judgment methodologies, including what state or trait is being judged, who is being judged, what cue channels are available, whether the cues are posed or spontaneous, whether judgment is done live or from standard stimuli, what judgment format is used, what operational criteria are used for scoring, and what techniques are used for scoring accuracy.

Conceptual issues

Sensitivity to what?

It takes only a little thought to recognize that sensitivity involves numerous processes and cues that vary along multiple dimensions (Bernieri 2001; Hall & Bernieri 2001). Such variations include:

- the depth of cue processing that one engages in (attentional versus inferential);
- the degree of awareness of cue processing (not conscious versus conscious);
- stimulus dynamism (static, such as physiognomic; semi-static, such as clothing style; or dynamic, such as hand gestures);
- stimulus channel (such as face versus voice);
- spontaneity of encoding (posed versus spontaneous);
- construct domain (such as states versus traits);
- what specific construct is being measured (such as different specific emotions).

It is important to state at the outset that distinctions such as given in these examples are oversimplifications. Awareness, for example, is a continuum, not a dichotomy, and the

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 240 17.6.2005 1:59am

same is true for stimulus dynamism (a given cue may appear static only because it does not change rapidly enough for the change to be evident in short behavioral samples). Below we discuss these dimensions, as well as various qualifications, in broad terms, leaving the specifics of measurement to a later section.

Depth of processing

As the Clever Hans example indicated, one kind of sensitivity is simply noticing. This *attentional accuracy* (Hall *et al.* 2001), which can apply to either behavior or appearance, is typically the first step in making accurate interpretations of the meanings or significance of cues. Interpretation can occur right away ('From the way she is looking at me, I can tell she knows I'm lying!'). Sometimes, however, what is noticed takes on significance at a later date: 'Oh, you didn't get the job. That's why you were so quiet at breakfast' or 'Now that I think of it, I realize my boss doesn't talk to me as much as he used to'. Sometimes it's the noticing itself that matters, rather than a higher order interpretation: 'I've noticed that my friend has pierced ears, so I will buy her that kind of earring' or 'Remember Jane? She's the one who smiles all the time'.

Though the concept of noticing cues appears to be very simple, conceptually one can distinguish between paying attention, noticing, and recalling cues. These three processes, listed in the order of likely occurrence in practice, are not synonymous. A person may be paying attention but might still miss cues, or may notice cues but not remember them later (or may remember them incorrectly). The field of eyewitness testimony demonstrates how fallible the recall of behavior that is seen and heard can be (Fruzzetti *et al.* 1992).

In contrast to attentional accuracy, the term *inferential accuracy* refers to judgments made about the meaning or significance of cues (Hall *et al.* 2001). Empirically, almost all research on interpersonal sensitivity, at both the group and individual level, has been based on inferential accuracy. The range of different things about which people can draw inferences will be discussed later.

Awareness of processing

In the example of a priming experiment in which expressive cues are presented subliminally, one can be sure that the cues were processed without conscious awareness. It is common to read that the processing of others' nonverbal cues is, by nature, tacit and out of awareness (e.g. Ambady & Gray 2002; Gilbert & Krull 1988). Edward Sapir famously referred to nonverbal communication as 'an elaborate secret code that is written no where, known by none, and used by all' (Sapir 1949). In other words, people are very skillful in their use and interpretation of nonverbal cues, but they have little or no explicit insight into cue usages and meanings (e.g. Grahe & Bernieri 2002).

However, in daily life we process cues with different *degrees* of awareness, from unaware to completely aware. Sometimes people are aware that they are processing specific cues: 'I could tell from the tears in his eyes that my husband was very moved by the movie'. Other times, they are aware that they are processing cues, but they are not very aware of the specific cues they are using: 'I don't know why, but I was sure she was telling the truth'. Of course, even if we are very deliberate about noticing and are very

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 241 17.6.2005 1:59am

sure of what we think we saw or what cues we think we relied on, we might still be wrong. People might have judgment policies (rules for decoding meaning; Bernieri 2001) that they consciously apply, but the judgment policies may be wrong and, therefore, their judgments will be wrong (Hall *et al.* 2004). Or, the judgment policy may be wrong but their judgments wind up being generally correct because the erroneous cue happens to covary with a cue that is valid for the judgment in question.

In daily life, as well as in the research laboratory, it is very difficult to know how aware people are of the process of judging cues. Research on self-fulfilling prophecies generally assumes that neither party in an interaction is aware of the biasing cues being conveyed and responded to (Rosenthal 1976; Rosenthal & Rubin 1978; Snyder *et al.* 1977) and, of course, any study using subliminally presented stimuli demonstrates unconscious processing. In the tradition of measuring sensitivity to cues (the subject of this chapter), the problem of determining awareness is usually skirted by making perceivers fully aware that they are judging cues and then focusing on accuracy of judgment rather than the process of judgment. Thus, whether perceivers use an implicit or explicit process is not of concern to the researcher.

But people do undoubtedly possess many beliefs about nonverbal communication (e.g. Carney *et al.* (in press); Rosip & Hall 2004) and these beliefs, independent of their application in the judgment of actual nonverbal stimuli, can be considered an explicit kind of interpersonal sensitivity so long as the investigator can score them for accuracy. An example of a correct belief would be that a facial expression of genuine enjoyment is likely to involve the muscles at the outer corners of the mouth and eyes (Ekman *et al.* 1990). (Of course, one could also have an implicit understanding of such a relation that one could apply in practice even if one could not articulate it explicitly.) Explicitly held beliefs about the meanings of nonverbal cues may certainly contribute to the accurate processing of nonverbal cues. To what extent such explicit knowledge contributes to accurate judgment of others is an empirical question about which not much is known at present (Rosip & Hall 2004). One study that begins to explore this general issue distinguished between cues that were abstract, subjective, and molar (e.g. apparent dominant behavior, nervousness) and cues that were more concrete, objective, and molecular (e.g. proximity, eye contact) in determining judgments of dyad rapport based on exposure to short excerpts of interaction (Grahe & Bernieri 2002). Judgments of rapport were influenced equally by both categories of cues. However, whereas perceivers were generally aware of how their judgments were influenced by abstract cues, they were much less aware of how the concrete, objective, molecular cues affected their judgments.

Stimulus characteristics

The question of what form the cues presented to perceivers should take is a complex one. In principle, one could develop a typology of interpersonal sensitivity measurements that represents the crossing of many factors including those already mentioned and those still to be discussed. A given method would represent a cell in this many-celled matrix. Here, we will be content with listing different conceptual factors.

1. Cue dynamism

Here again, a continuum is present but, pragmatically, one can talk about different categories of dynamism. Some cues are intrinsically static, at least within reasonable time frames; height being an example. Others can vary gradually over time (weight, hair length while it grows), while others that seem intrinsically static can actually change abruptly with surgical or cosmetic intervention (body shape, facial features, haircuts). Some cues are relatively static within an interaction (seated distance from an interaction partner), while others are likely to change very often (mouth or hand movements). Sometimes 'static' versus 'dynamic' is artificially defined by varying how long the perceiver is allowed to view the stimulus. A still photograph isolated from the ongoing behavioral stream is thus made into an artificially static stimulus, or posture may seem static only because the video clips are short and the person's posture did not change during the time allowed.

2. Channels

The concept of channel is useful heuristically, as a way of dividing up the sources of nonverbal information, but its specific definition may vary. Thus, 'face' may be a convenient channel for one researcher but regions of the face may be conceptualized as different channels by another researcher. Commonly discussed channels are facial expressions, eye behavior, head movements, upper body movements, lower body movements, hand/arm movements, posture, proxemic variables, touch (self and other), vocal behavior, face or body physiognomy, hair and make-up, and clothing and accessories (Knapp & Hall 2002). Within each of these categories, multiple subcategories can be identified.

The distinction between the verbal and nonverbal behavior channels, seemingly an easy one, can actually be unclear. Hand emblems such as the A–OK sign or the 'gun-to-temple' gesture in U.S. culture (Ekman & Friesen 1972) have such distinct verbal translations that they might almost be considered verbal, even though technically they are not. Manual sign language systems are generally considered so far toward the linguistic end of the continuum that nonverbal communication researchers typically do not study them. Fluid hand movements emitted during speech are closely tied to the language encoding process (McNeill 1985). Behaviors such as interruptions and back-channel responses (e.g. 'mm-hmm', 'yeah', 'I see') are often considered to be functionally nonverbal because their significance does not depend on the linguistic content per se.

Pragmatically, researchers of interpersonal sensitivity sometimes keep the verbal and nonverbal channels integrated, so that perceivers are exposed to both at once (e.g. Costanzo & Archer 1989; Ickes 2001; Vogt & Colvin 2003), and sometimes they separate them (e.g. silent video, content-masked speech, transcript) (Gesn & Ickes 1999; Murphy *et al.* 2003; Rosenthal *et al.* 1979; Scherer *et al.* 1977).

3. Spontaneity

Cues to be noticed or judged may vary in how spontaneously or deliberately they occur. This too is a continuum, with totally spontaneous, unrehearsed, and unplanned behavior at one end (e.g. facial expressions one doesn't even know one is making)

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 243 17.6.2005 1:59am

and completely deliberate, planned, or posed behavior at the other (e.g. putting on a display of good cheer when you actually feel sad, displaying hand emblems such as the 'A–OK' sign, deliberately looking at your watch to indicate to someone that you need to leave). This dimension corresponds roughly to the conscious–unconscious distinction made earlier with respect to the perceiver's awareness of the nonverbal perception process. There has been debate over whether one can ever be sure that expressions are unplanned or unintended (Fridlund 1997).

4. Construct domain

Bernieri (2001) identified numerous categories of meaning about which a perceiver can be interpersonally sensitive. Inferences can be about *states* or *traits* (another distinction that has an underlying continuum). Most commonly studied is sensitivity to *states* and, within this, affective states have received the most attention. These can be measured directly (is the person showing disgust on his face?) or indirectly (is the person using the kind of cues she would use if she was talking about her divorce/thanking someone/talking to a lost child?). Another state commonly studied is truth versus falsehood, which also can be inquired about directly (is she lying?) or indirectly (does this person make me feel uncomfortable?).

Many other states can, of course, be the objects of sensitivity, though affect and deception have received the most research attention. We often make judgments about others' thoughts, intentions, needs, physical states, and likely future behavior from observing their outwardly expressed behavior (Bernieri 2001).

The domain of *traits* that may be judged from nonverbal cues is also large. In this category, ability to judge personality traits has been studied most often. Other individual differences that people judge in daily life include intelligence, specific competencies, status or dominance, ethnicity, culture, sexual orientation, mental health, and social adjustment.

As alluded to above, when we introduced the concept of 'attentional accuracy', another kind of sensitivity is memory for (relatedly, noticing or attending to) cues. Such cues can be static (e.g. appearance; Horgan *et al.* 2004) or dynamic (e.g. nonverbal behavior; Hall & Murphy 2004).

Table 6.1 shows (not exhaustively) specific interpersonal sensitivity constructs that have been measured, with illustrative studies cited for each.

What is the accuracy criterion?

In daily life, when we make the kinds of judgments listed above, sometimes we find out whether the judgments are right or wrong: 'I can't believe I fell for his lies' or 'You're 21? You look so much younger'. But much of the time, we never know for sure. However, researchers who set out to measure accuracy *must* know or else they cannot score their instruments. Deciding what is the 'right answer' to an interpersonal sensitivity question has been called the criterion problem (Archer *et al.* 2001; Bernieri 2001; Kenny 1994; Rosenthal *et al.* 1979).

Kenny (1994) categorized criteria used in sensitivity research as self-report, consensus, expert judgments, behavioral observations, and operational criteria. Establishing a

Table 6.1 Constructs assessed in interpersonal sensitivity research

Construct	Illustrative research
Situationally determined affect	Costanzo & Archer 1989; Rosenthal *et al.* 1979
Emotions	Matsumoto *et al.* 2000; Nowicki & Duke 1994
Relationships	Barnes & Sternberg 1989; Costanzo & Archer 1989
Love	Aloni & Bernieri 2004
Rapport	Bernieri *et al.* 1996
Deception	Ekman & O'Sullivan 1991; deTurck 1991
Personality	Blackman 2002; Borkenau & Liebler 1995
Status	Barnes & Sternberg 1989; Schmid Mast & Hall 2004
Others' interpersonal sensitivity	Carney & Harrigan 2003
Intelligence	Murphy *et al.* 2003; Reynolds & Gifford 2001
Thoughts and feelings	Ickes 2001; Thomas & Fletcher 2003
Prejudice	Carney 2004; Rollman 1978
Sexual orientation	Ambady *et al.* 1999
Interactional enjoyment	Carney *et al.* 2004
Ethnic group	Allport & Kramer 1946; Dorfman *et al.* 1971
Recall of appearance	Horgan *et al.* 2004
Recall of nonverbal behavior	Hall *et al.* 2001; Hall & Murphy 2004
Explicit knowledge of nonverbal cues	Rosip & Hall 2004; Vrij & Semin 1996
Explicit knowledge of gender differences	Hall & Carter 1999; Swim 1994

valid criterion can be a difficult epistemological enterprise because often there is not an unimpeachable 'gold standard'. We will mention some of these difficulties as we proceed. The summary below is brief and the reader should consult the sources named above for an expanded discussion.

Self-report

Self-report, or what the target (i.e. the person being judged) says about him/herself, is mostly used when the target is deemed to have valid knowledge about the state, trait, or characteristic in question. Examples of each of these would be 'my current mood', 'my extraversion', and 'my age'. Sometimes such self-reports are measured using well-validated instruments (such as for measuring personality). Setting aside the possibility that the target would fabricate the answer, self-report criteria are fallible to the extent that the target may not actually know the correct answer. The tendency to engage in self-enhancing distortions is ubiquitous and may not always involve a cynical attempt to deceive (Colvin *et al.* 1995).

Consensus

Consensus, or what observers agree is the right answer, is often used as the criterion for labeling emotional expressions (e.g. Ekman *et al.* 1987; Zuckerman *et al.* 1975). Consensus judgments are fallible to the extent that observers share an erroneous judgment policy (association between the cue and its attributed meaning) even when they show high inter-observer reliability of judgment. For example, observers may agree, but erroneously, that if a woman speaks relatively loudly she is more likely to be addressing a woman than a man (Hall & Braunwald 1981).

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 245 17.6.2005 1:59am

The desirability of consensus as a criterion may depend on the nature of the construct being judged. If the construct is defined as something residing within the target person, observers' consensus may not be a good criterion. For example, physical pain exists within the target and is defined independently of what observers think. However, if the construct is socially defined (e.g. expressed hostility or politeness), consensus may be the most appropriate criterion.

Expert judgments

Expert judgments may be provided by respondents who are considered to have the best possible knowledge about the target's state or trait. For example, a clinical psychologist's opinion of a target's degree of psychoticism would likely be more trustworthy than the target's own opinion. Of course, such 'experts'(whether they be clinicians, teachers, supervisors, parents, or friends) can still be biased or ignorant.

Behavioral observations

Behavioral observations are especially relevant for sensitivity defined as attentional accuracy. Thus, trained coders or a computer might provide data on a target's length of gaze or average fundamental vocal frequency and this could then serve as the basis for determining whether perceivers are accurate in their recall of gaze or pitch. Trained coders can also be used to establish the criterion for perceivers' judgments of personality (Funder *et al.* 2000). The reliability and validity of such criteria are obviously relevant here (as is always the case, no matter what the criterion). If behavioral observations are highly impressionistic or inferential (e.g. if a group of naïve judges were to rate the friendliness of the targets), such a criterion might better be called consensus.

One important illustration of behavioral observation as criterion is the application of the Facial Action Coding System or FACS (Ekman *et al.* 2002). The FACS produces a detailed identification of what facial muscles have moved, in what combinations, and how much. By itself, the FACS is simply a descriptive system. However, when paired with empirical findings indicating what muscle activity is associated with what kind of emotional experience or intention, the FACS can be used to establish scoring criteria for expressions. For example, if a given configuration of facial movements is deemed, according to FACS research, to show 'disgust', then those movements are the criterion against which judgments of that expression are scored (Ekman & Rosenberg 1997; Ekman *et al.* 2002).

Behavioral observations have the appeal of being relatively concrete to define but they become problematic if the researcher wants to draw a higher-order inference about the behavior in question (Bernieri 2001). For example, it might be easy to count interruptions, but not easy to know whether interruptions mean dominance or simply active participation.

Operational criteria

Finally, operational criteria are used when some externally verifiable fact can be identified as the operational definition of the target's state or trait. If the researcher chooses when the target will lie or tell the truth, then the researcher's choice is the

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 246 17.6.2005 1:59am

operational criterion to be used when scoring the accuracy of perceivers' lie–truth judgments. As other illustrations, Buck (1979) showed emotionally evocative pictures to the target and then asked perceivers to guess what pictures were being shown just from watching the target's face (thus, the pictures are the operational criterion). Costanzo and Archer (1989) arranged for a (real) boss and subordinate to interact and then asked perceivers to guess which person was the boss.

In early research, some experimenters created unexpected experiences for participants as a controlled method of producing specific emotional states. So, for example, having to cut the head off a dead rat was assumed to produce disgust and being told (falsely) that one's loved one had died was assumed to produce grief (Dunlap 1927; Landis 1924). The common technique of asking targets to deliberately pose the expression of certain emotions (Noller 1980, 2001; Nowicki & Duke 1994; Rosenthal *et al.* 1979) and then using the posed intention as the criterion, can also be included in this category. In the assessment of explicit cue knowledge (Rosip & Hall 2004), the scoring criterion is also operational in nature because it treats findings available in previous research as the 'gold standard' against which people's beliefs are compared and scored for accuracy.

Each of these criterion definitions has limitations, some of which we have identified. Each criterion definition is likely to have its own construct validity problems. Sometimes, researchers combine more than one of these criterion-setting methods in order to reduce error to a minimum. As examples, Vogt and Colvin's (2003) criteria for targets' personality included self-reports, parental reports, peer reports, and direct observation; and Scherer and Ceschi's (2000) criteria for targets' emotions included self-reports, independent observers, and behavioral measures. What to do when criteria conflict remains a problem? In practice, the best test of the validity of accuracy criteria may be whether the accuracy scores generated by a given method or instrument produce findings suggestive of construct validity.

Questions at the intersection of theory and method

We started with general observations about the nature of interpersonal sensitivity and then progressed toward empirical issues by discussing different kinds of criteria that can be used for determining accuracy. Now we move further into empirical territory by asking several questions that relate to actual research on interpersonal sensitivity. It is not our intention to review all, or even much, of what is known from this body of research. But it is essential to ask several fundamental questions about this research in order to set the stage for a description of specific methodologies.

Can cues be judged accurately?

If researchers could not obtain above-chance levels of accuracy with their measuring instruments, one could justifiably challenge the tests' adequacy. If, for example, perceivers trying to identify a 'fear' expression are systematically wrong in their judgments or are no more accurate than they would be if they were just guessing, one might question whether accuracy of judging 'fear' had adequately been tested. Therefore, most researchers look for overall accuracy at above-chance levels.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 247 17.6.2005 1:59am

Numerous techniques can be applied to manipulate difficulty level and thereby achieve a desired overall level of accuracy. These include varying the amount of the stimulus the perceiver is exposed to (e.g. 2 seconds vs. 10 seconds); altering the wording of item alternatives on a multiple choice response to make discriminations easier or harder (e.g. a choice between 'happy' and 'fear' as opposed to 'surprise' and 'fear'); or manipulating signal clarity (e.g. high vs. low-intensity vocal expressions; Baum & Nowicki 1998). Sometimes, researchers calibrate their instruments to achieve a level of accuracy that is not merely better than chance but that falls at a level that optimizes variance in scores in order to create the best opportunity for detecting individual differences (Guilford 1954). For example, if the guessing level is 50%, then optimal accuracy is around 75% (Rosenthal *et al.* 1979; Rosip & Hall 2004).

Because investigators can exercise a great deal of control over the mean level of accuracy obtained on a given test (and also because the judged stimuli, judgment task, and scoring methods can be noncomparable between tests), interpretations of absolute levels of accuracy and comparisons of accuracy across different constructs or tests must be undertaken with great caution. To help researchers compare across studies, Rosenthal and Rubin (1989) developed an effect-size index for standardizing mean percentage accuracy across studies that differ in the number of response alternatives provided to perceivers.

Broadly speaking, perceivers have above-chance levels of accuracy, though the range is great. Judgments of deception are not much above the guessing level (Malone & DePaulo 2001), whereas accuracy in identifying prototypical facial expressions of emotion is often extremely high (e.g. Biehl *et al.* 1997; Ekman *et al.* 1987). Whether the overall level is 'too high' or 'too low' depends on the purposes of the research. It is important to note, however, that obtaining levels above chance is not necessarily required in order to have a valid test. When researchers cannot obtain above-chance levels it is important to ascertain whether the observed scores contain true-score variance or are due completely to random error. It is entirely possible for a measure to have a large true-score variance component even when the sample mean is below average. For example, imagine a Spanish reading comprehension test that is given to a sample of 100 people of whom only 10 or so know the language. By the same token, individual items on a test that fall below the chance level may still be valid items if they correlate with other items on the test. Such an item, though difficult for most people, is still likely to be judged correctly by good decoders (Rosenthal *et al.* 1979). Demonstrating the validity of such a test or items requires carrying out construct validity studies to show (for example) that the test or item correlates as predicted with other variables.

Are group effects worth studying?

Showing an adequate level of interpersonal sensitivity at the group level may be comforting to a researcher, but it does not answer the question of whether such sensitivity is an interesting social psychological variable. This question depends on whether sensitivity, as it is measured in research, is related to other real-world or experimentally manipulated group-level variables. This should not be considered a

foregone conclusion. Perhaps interpersonal sensitivity in 'real life' does not vary with situational, task-specific, or group-based factors; or, perhaps the instruments that researchers have devised fail to capture that sensitivity even if such relations actually exist.

Research has, in fact, produced many group-level results—far too many to review here. Based on reviews and primary sources (e.g. Ambady *et al.* 2000, 2001; Ambady & Gray 2002; Baum & Nowicki 1998; Carney *et al.* 2004; Elfenbein & Ambady 2002; Hall 1978, 1984; Malone & DePaulo 2001; Rosenthal *et al.* 1979), it has been shown that accuracy varies with the channel being judged (e.g. face vs. voice), the length of stimulus exposure, the intensity of the cues being expressed, the specific construct being judged, the gender of perceivers and targets, the culture and ethnicity of perceivers and targets, perceivers' occupational characteristics, and perceivers' manipulated mood. This sampling of results indicates that interpersonal sensitivity is a meaningful social psychological construct.

Do individual differences exist?

Many researchers of interpersonal sensitivity are interested in questions relating to individual differences, such as where such differences come from and their cognitive, social, and personal correlates. Therefore it is crucial to demonstrate that research instruments and methods are capable of detecting individual differences. This question takes us to questions about the psychometric adequacy of measuring approaches, because individual differences that are not associated with reliable measurement are due to random error (noise) and are therefore not true individual differences. Also related to the question of reliability is the question of how intercorrelated different tests are. We address both of these questions in this section.

Reliability

One index of true individual difference variance is retest reliability: do test takers maintain their relative rank compared to each other when tested again? Established tests report adequate retest reliability, with median retest correlations of 0.69 across six samples taking the PONS test (Rosenthal *et al.* 1979), and retest correlations of 0.80 for the CARAT (Buck 1976), 0.70 for the IPT (Costanzo & Archer 1989), 0.80 and above for the DANVA (Nowicki & Duke 2001), and 0.88 for an omnibus paper-and-pencil test of nonverbal cue knowledge (Rosip & Hall 2004). (Individual tests will be described in more detail in later sections.) These figures indicate that there is considerable stability in measured interpersonal sensitivity.

Another index of reliability is internal consistency, which is a joint function of the strength of inter-item correlations and the number of items on the test. The widely used PONS test (Rosenthal *et al.* 1979) has good internal consistency (KR-20 = 0.86), but this good reliability is achieved by having a large number of items on the test (220 to be exact) that in fact have an average inter-item correlation of only about 0.03. Therefore, though the full-length test has good internal consistency, the short forms of the PONS (such as the 40-item face and body test, or the 40-item voice test), as well as several other tests, including the IPT (Costanzo & Archer 1989), have poor internal consist-

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 249 17.6.2005 1:59am

ency—typically, Cronbach's alphas of less than 0.40 and, sometimes, much less (see review in Hall 2001). On the other hand, some tests, including the DANVA (Nowicki & Duke 1994) and the JACBART (Matsumoto *et al.* 2000), have internal consistency (alpha) in the 0.70–0.80 range. Rosip and Hall's (2004) test of explicit nonverbal cue knowledge also has internal consistency in this range.

It is interesting that nonverbal decoding tests (that is, tests involving the judgment of cues emitted by targets) with the best internal consistency tend to be those testing a single content domain, namely emotions. Decoding tests with weak inter-item correlations (e.g. PONS and IPT) cover a much broader domain of content, in that the test taker must judge cues that are associated with a range of affective, role, and situational circumstances.

In principle, the problem of weak inter-item correlations in sensitivity instruments can be rectified as long as the inter-item correlations are greater than zero and there is a sufficient number of items (as on the full-length PONS test for which internal consistency is acceptable). Unfortunately, however, real-world constraints can limit the realization of psychometric goals. The full-length PONS test is often foregone in exchange for shorter but less reliable forms of the test, and some tests with questionable internal consistency are already rather long (e.g. the full-length IPT test). Because it is typically not feasible in practice to use extremely long tests, low internal consistency may remain an issue.

Another intriguing possibility that sensitivity researchers and theoreticians may need to deal with is that the various components, skills, and competencies that constitute global sensitivity —sometimes even as they exist within a single instrument such as the IPT or PONS—are simply orthogonal to one another. To the extent that this is true, the internal consistency of a single test instrument becomes less relevant, as will be discussed below.

Intercorrelations among tests

Repeatedly, correlations between different interpersonal sensitivity tests have been found to be very low or even close to zero (reviewed by Bernieri 2001; Colvin & Bundick 2001; Hall 2001). Furthermore, analysis of the factor structure of the PONS test suggested that the major channels of face, body, and voice are relatively orthogonal on that test (Rosenthal *et al.* 1979). At present, it is not known whether this pattern of low intercorrelations stems from weak psychometric qualities of the instruments (i.e. weak internal consistency) or reflects the true structure of the interpersonal sensitivity domain. It is quite possible that the domain of interpersonal sensitivity consists of several, perhaps many, discrete skills that are not well predicted one from another, as proposed by one prescient writer in the early days of interpersonal sensitivity research (Buck 1976). Thus, ability to detect lies may be unrelated to ability to judge emotion in the voice, and these may both be independent of many other abilities (e.g. to judge the kind of relationship two people have, to judge someone's personality traits).

Alternatively, there may exist a structure to the sensitivity domain that can only be detected with larger test batteries and latent variable analysis. Possible structures could be based on channels, the types of constructs being judged (e.g. state vs. trait), the

breadth or specificity of test content, specific design methodologies or scoring systems, and so forth. (For discussion of the structure of the emotional intelligence construct, see Mayer *et al.* 2003 and Roberts *et al.* 2001.)

Are poor internal consistency and poor between-test correlations necessarily bad?

As we have noted, the domain of interpersonal sensitivity may include quasi-independent discrete variables that collectively define the higher-order construct. One might think of either whole tests or items within tests in this way. To think in this manner is analogous to seeing income, education, and occupational prestige as collectively defining socioeconomic status (Bollen & Lennox 1991). Socioeconomic status is the empirical consequence of one's standing on these indices rather than being a latent construct of which each index is simply an indicator. Thus, the fact that income, education, and occupational prestige are not strongly intercorrelated would not lead one to conclude that these variables are flawed indicators of the latent construct or that one had created a psychometrically bad scale of socioeconomic status. Rather, socioeonomic status is *defined* by the component variables, which might even have a compensatory relation to each other (e.g. high income may compensate for less education).

Applying the same logic to the case of interpersonal sensitivity, one could argue that tests that are more omnibus in their content may actually gain validity by including items that represent a number of different skills. The IPT (Costanzo & Archer 1989), for example, includes items relating to deception, kinship, competition, status, and intimacy. On the other hand, a test that includes items from distinct meaning domains could have attenuated correlations with external variables if not all of the meaning domains actually bear a relation to those variables. For example, total score on a test that includes items on lie detection as well as role relations might correlate weakly with success as a police interrogator if only the lie-detection component of skill is relevant to police interrogation skills. Thus, there may be both good and bad aspects to tests with diverse content.

Do individual differences matter?

Good reliability does not necessarily mean that what is measured is meaningful or useful. Conversely, as discussed above, poor internal consistency does not necessarily mean that nothing meaningful or useful is being measured. To ask whether something meaningful is being measured is to ask about construct validity. Fortunately, tests of interpersonal sensitivity have many correlates that both support the claim that interpersonal sensitivity is being measured and reveal a great deal about the place of interpersonal sensitivity in the daily lives of children and adults. This chapter is too short to do justice to this literature, but a brief summary based on reviews as well as primary sources will demonstrate how rich the network of findings is (e.g. Ambady *et al.* 2001; Archer *et al.* 2001; Bernieri 1991; Carney & Harrigan 2003; Costanzo & Archer 1989; DiMatteo *et al.* 1979; Funder & Harris 1986; Hall 1998 (Table 7–4); Hall & Carter 1999; Hall *et al.* 1997, 2000; Knapp & Hall 2002; Nowicki & Duke 2001; Rosenthal *et al.* 1979). We limit this summary to research based on decoding tests (i.e. inferential accuracy) because nearly all research is on that topic.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 251 17.6.2005 1:59am

Better interpersonal sensitivity, as measured with inferential tests, has been found to be associated with increasing age through childhood; better mental health; more well adjusted personality; less shyness and social anxiety; more dominant personality; higher self-esteem; higher social competence based on sociometric ratings, as well as peer, teacher, and parent ratings; higher ratings of sensitivity by peers or supervisors; better ability to judge a friend's interpersonal sensitivity; more democratic attitudes among teachers; more social inclusion needs; more cognitive complexity; more self-monitoring; more internal locus of control; more popularity; less aggression; higher academic achievement (when test takers are children); better supervisor ratings of job performance (when test takers are clinicians, foreign service officers, and teachers); more satisfaction among medical patients (when test takers are the patients' physicians); quicker ability to learn in a dyadic teaching situation; possession of more accurate knowledge of differences between men's and women's behavior; higher reports of marital satisfaction; and being rated by peers as more likeable, honest, and open, and less hostile and manipulative. However, self-assessment of one's own sensitivity (including confidence in one's performance) shows little relationship to measured accuracy (Aloni & Bernieri 2004; Carney & Harrigan 2003; DePaulo *et al.* 1997; Riggio & Riggio 2001; Smith *et al.* 1991; Zuckerman & Larrance 1979).

The discriminant validity of interpersonal sensitivity tests also requires examination. The variable of most concern in this regard is overall intelligence as measured by IQ or achievement measures. Correlations with such measures range from negligible to moderate in magnitude (Davis & Kraus 1997; Halberstadt & Hall 1980; Nowicki & Duke 2001; Rosenthal *et al.* 1979; Rosip & Hall 2004), with the trend suggesting a satisfactorily small contribution of general cognitive ability to interpersonal sensitivity. Thus, tests of interpersonal sensitivity are not simply measuring overall cognitive ability.

Although the range of correlates listed above indicates that individual differences do matter, it is important to acknowledge that many of the correlations are modest in magnitude, and sometimes they do not support predictions. It is also important to note that in listing correlates of interpersonal sensitivity, we are not making claims about causal relations. Not much is known about the causal antecedents and consequences of interpersonal sensitivity.

Major paradigms

Standard cue sets

By far the most common paradigm for assessing interpersonal sensitivity involves the use of standard cue sets, by which we mean stimuli that are stored in the form of photographs, drawings, audiotapes, or videotapes and which can therefore be judged by groups of perceivers and re-used on more than one occasion. The cue set contains multiple items showing one or more targets (the term 'target' is synonymous with 'encoder' or 'expressor'). Structuring the cue set involves many choices including the number of targets, length of clips, what channels to include, what constructs to include. At this stage, researchers can only speculate about the implications of these choices (Colvin & Bundick 2001).

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 252 17.6.2005 1:59am

Sometimes a researcher creates a stimulus set for a particular study and does not use it again (e.g. Zuckerman *et al.* 1975), and sometimes a researcher invests much time and energy in developing, refining, and standardizing stimuli for repeated use. When stimuli are standardized and the researcher also wants to promote the stimulus set as the basis for a standardized test, he or she is also likely to conduct psychometric and normative analyses as well as undertake to demonstrate convergent and discriminant validity. Other sources can be consulted on psychometric theory and the design of tests (e.g. Cronbach 1990; Kline 2000*a*,*b*; Loewenthal 2001).

The validity of using standard cue sets is well established through empirical research. The term 'thin slices' was coined by Ambady and Rosenthal (1992) to describe cue sets containing very brief excerpts of behavior (less than five minutes in length but often less than one minute). In addition to a wealth of research showing that accuracy judgments made from thin slices have construct validity as predictors of personal characteristics, states, and outcomes (as indicated by the partial list offered earlier), research shows that accuracy of judgments based on thin slices can significantly predict accuracy based on a longer stream of the same behavior (Archer & Akert 1980) and that accuracy based on a thin slice may be as high as accuracy based on a longer slice (Carney *et al.* 2004).

There are many advantages to using standard cue sets. The stimuli are easily transportable, and administering the test can typically be done with simple equipment in a group setting. A scoring key needs to be developed only once. If the research question requires the researcher to code the stimuli for different cues (e.g. how much the targets smiled or details of the targets' clothing), the behavior needs to be coded only once (e.g. Borkenau & Liebler 1995; Hall & Murphy 2004; Schmid Mast & Hall 2004). Using a standard cue set also facilitates valid comparisons among perceivers and between groups because accuracy is measured against a common stimulus. Finally, with this approach one can easily separate verbal from nonverbal cues as well as different nonverbal channels of communication, as described later.

There are also limitations to using standard cue sets. One is ecological validity, because the behavior represented in a standard cue set is likely to be shown briefly and out of context, which could have a negative impact on the accuracy of judgments. Also, there are limitations to the researcher's ability to generalize beyond the specific features of the measurement paradigm used. The extent to which validity is jeopardized by design features of the stimulus set is an empirical question. Though the PONS test was criticized for having only one target (LaFrance & Henley 1994), the accumulated results for the test suggest that it has validity (e.g. Funder & Harris 1986; Hall 2001; Rosenthal *et al.* 1979). Whether it would have even more validity if it had more targets, or had different content, or used a different criterion as the basis of scoring, is not known.

Face-to-face assessment

Much more rarely undertaken is the assessment of interpersonal sensitivity between people who interact *with each other*. There are two reasons for doing this kind of research. One is that one can achieve a higher level of 'real lifeness' than can be obtained using standard cue sets. Communication that occurs during live interaction entails

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 253 17.6.2005 1:59am

cognitive and motivational processes that are difficult, perhaps impossible, to create in a standardized decoding paradigm (Bernieri 2001; Patterson 1995). The second reason is that the live paradigm permits the investigation of theoretical questions not easily handled in the standardized paradigm—for example, the accuracy of husbands' and wives' communication with each other (Noller 1980) or the accuracy of superior–subordinate communication (Hall *et al.*, in press; Snodgrass 1985, 1992; Snodgrass *et al.* 1998).

Because assignment of roles controls for pre-existing skills and experiences, the live interaction paradigm is especially suited to studying the impact of motivational processes on accuracy. To illustrate, Hall *et al.* (in press) adapted the dyadic communication task of Noller (1980) by randomly assigning dyad members to high- and low-power roles and then having them deliberately send nonverbal affective messages to each other. Accuracy of decoding was scored by comparing judgments to the affect being intentionally communicated.

However, there are significant difficulties with this research paradigm. The first is the labor intensiveness of recruiting and running participants in live interactions. The second is interpretational ambiguity that is intrinsic to a within-dyad communication situation. In the Hall *et al.* (in press) study, as in Snodgrass (1985, 1992), it was not clear whether a difference in the decoding accuracy of the assigned groups was due to one group making special efforts to decode well or to the other group producing messages that were especially easy to judge. These sources of accuracy are fully confounded in a dyadic situation.

To understand the source of this difference, Hall *et al.* (in press) showed the videotaped expressions to naïve judges (as did Noller 1980 and Snodgrass *et al.* 1998). Comparison of the naïve group's accuracy to the original groups' accuracy makes it possible to disentangle the confounded sender and receiver effects because the naïve group's performance can be considered a pure indicator of the accuracy with which the original expressions could be decoded. Thus, the dyadic communication methodology has significant ambiguities, the resolution of which can have substantial theoretical importance (in the example case, whether low-power individuals' accuracy of judging superiors' cues was due to their decoding efforts or to the clarity of the messages they were given to judge).

If the researcher is able to gather judgment data (with appropriate criteria for scoring) from several people in a round-robin or other appropriate design (in which each person in a group serves as both target and judge; Kenny 1994; Kenny *et al.* 1996; Kenny & Winquist 2001), then well-developed statistical methods, based on Kenny's Social Relations Model, are capable of separating different sources of variance. Specifically, perceiver, target, and perceiver X target effects (analogous to main effects and interaction in the analysis of variance) can be isolated for both judgments and criteria and compared to derive different kinds of accuracy scores. Difficulties with this approach include the logistics of recruiting participants in groups and running enough groups for meaningful analysis, and mastering the requisite statistical tools. It is, furthermore, a method best suited to the study of judgment processes among strangers. However, as Bernieri (2001) argues, the intrinsically componential nature of measured accuracy is a fact to be reckoned with, though some researchers may be more interested

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 254 17.6.2005 1:59am

than others in decomposing an accuracy score into different sources of variance such as rating biases, general knowledge of people and situations, and accuracy of judging specific targets with other (artifactual) influences removed.

The hybrid paradigm: retrospective tape review

Another methodology for measuring interpersonal sensitivity is the retrospective tape review ('empathic accuracy') paradigm developed by Ickes and colleagues (Ickes 1997, 2001; Ickes *et al.* 1990). We refer to this paradigm as 'hybrid' because participants both engage in face-to-face interaction and make judgments of recorded behavior. After participants interact in live interaction, each participant individually reviews the videotape of the interaction two times. The first time, the participant stops the tape whenever he or she remembers having had a 'thought or feeling' and writes down the content of the thought or feeling. The second time, the participant watches his or her partner, with the experimenter stopping the tape at each of the partner's previously identified thoughts and feelings, at which point the participant guesses what the partner's thought or feeling was. Then the two lists of thoughts and feelings (the partner's self-reports and the participant's guesses) are compared and scored for accuracy. Findings to date suggest that accuracy in this paradigm depends more on verbal than on nonverbal cues (Gesn & Ickes 1999).

As with the face-to-face accuracy paradigm, dyadic retrospective tape review confounds perceiver and target effects (Ickes 2001). However, it is possible to avoid this difficulty by enlisting naïve viewers to make judgments, as well as the original interactants. It is also possible to use videotape clips from this paradigm as a standard cue set that shows multiple targets, for which each target's self-reported thoughts and feelings serve as the criteria of accuracy (Gesn & Ickes 1999; Marangoni *et al.* 1995).

Explicit knowledge assessment

As mentioned earlier, a potentially interesting approach to measuring interpersonal sensitivity consists of assessing people's explicit knowledge about social behavior. In this approach, the test taker is asked directly (on a paper-and-pencil test) about the meanings or correlates of nonverbal cues or about a domain of social behavior. Vrij and Semin (1996) measured knowledge of cues to deception, Murphy (2003) measured knowledge of cues to intelligence, Hall and Carter (1999) measured knowledge of a range of gender differences, including gender differences in nonverbal communication, and Rosip and Hall (2004) measured knowledge about a wide range of nonverbal cues and usages. In each case, the explicit knowledge test was scored by comparing responses to a 'gold standard' developed from the research literature.

This approach may have utility in itself, and is likely to expand understanding of the proximal determinants of sensitivity as measured with a performance test. For example, Rosip and Hall (2004) found that women scored higher on their 81-item Test of Nonverbal Cue Knowledge (TONCK) and that scores on the TONCK had a modest but significant correlation with performance on nonverbal decoding tests (PONS and DANVA). Recent work is emerging to demonstrate that performance on the IPT is also

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 255 17.6.2005 1:59am

predicted by one's explicit knowledge of the relevant cues related to the interpersonal domains highlighted in that video task (McLarney–Vesotski 2003). How strong the correlations are likely to be between explicit knowledge and performance-based sensitivity could depend on many factors, including the degree of overlap in item content between the tests (i.e. whether tests are concerned with similar or different domains of meaning), perceivers' ability to describe explicitly the tacit knowledge base they use when making judgments of people, the adequacy of the 'gold standard' used for scoring accuracy on the paper-and-pencil test, and the contribution of transient factors such as motivation and distraction.

Operational issues

Separating channels

If only visual behavior is to be judged, it is easy to use photographs or silent videotapes. Visible features can be selectively obscured using electronic masking (e.g. obscuring facial expressions so that the viewer can see only body movements). When words only are to be judged, transcripts of what is said can be prepared for perceivers to read and judge (e.g. Murphy *et al.* 2003). When vocal nonverbal characteristics are to be judged, it is necessary to obscure the verbal content. This has been done in several ways, including the following (see Scherer *et al.* 1985; Scherer 2003; and Chapter 3 for more detailed descriptions and comparisons). *Randomized splicing* consists of dividing the voice sample into short segments, rearranging the segments, and playing the new sample to perceivers (Rosenthal *et al.* 1979). The meaning of a spoken sentence is no longer evident but the voice retains many of its acoustic properties. The use of a low bandpass filter produces what is known as *electronically filtered speech* by removing the highest tones, thereby making consonants hard to identify and making the voice sound muffled and the words unintelligible (Rosenthal *et al.* 1979). Both of these methods can be used with naturalistically recorded voice samples.

Standard-content methodology consists of asking targets to read something with neutral affective content such as the alphabet, a weather report, or a standard sentence (e.g. Borkenau & Liebler 1995; Dimitrovsky 1964; Noller 1980; Nowicki & Duke 1994). If emotional meaning is to be judged, the targets may be asked to deliberately vary their voices to convey the desired emotions. If traits (e.g. intelligence, extraversion) of the targets are to be judged, targets may be asked to behave in their normal way while reading or reciting the standard content.

Instruments vary in whether they are designed to be scored in only one channel (e.g. facial expressions on the JACBART; Matsumoto *et al.* 2000), whether they can be scored in multiple channels (e.g. face, body, and voice on the PONS test; Rosenthal *et al.* 1979), or whether verbal and nonverbal channels (and various nonverbal channels) are typically not distinguished (e.g. IPT of Costanzo & Archer 1989; empathic accuracy task of Ickes 2001). However, even in the latter case, the researcher can still experimentally separate the channels if this is desired. Doing so brought Costanzo and Archer (1989) to the conclusion that verbal information contributes little to accuracy on the IPT, while a similar analysis led Gesn and Ickes (1999) to conclude that verbal

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 256 17.6.2005 1:59am

information matters more than nonverbal information in the empathic accuracy paradigm.

Response formats

The most common response format for testing interpersonal sensitivity is *multiple choice*, with the number of options ranging from two to approximately seven. As examples, for each item on the PONS test there are two (out of a total of 20) situational labels (e.g. 'talking about the death of a friend' and 'expressing jealous rage'), with the pairings varying from item to item (Rosenthal *et al.* 1979). The DANVA presents the options 'anger, fear, sadness, or happiness' for all items (Nowicki & Duke 1994). The JACBART lists seven emotions as the choices (Matsumoto *et al.* 2000). Most lie detection tasks involve two choices (truth or lie) (Malone & DePaulo 2001). The TONCK (omnibus explicit knowledge test) presents a 'true–false' option for all items (Rosip & Hall 2004).

Corrections have been offered to control for different numbers of response options (Rosenthal & Rubin 1989), similarities and differences among response alternatives (e.g. more negative than positive emotion options; Ekman 1994), and response bias (Wagner 1993, 1997). Studies that compare results with and without such corrections are much needed (e.g. Elfenbein & Ambady 2002; see Scherer *et al.* 2003 for a discussion of corrections).

When the responses are dichotomous, signal detection methods (Green & Swets 1966) may be applied to distinguish response bias from accuracy, as was done in Dorfman *et al.*'s (1971) study of accuracy in distinguishing Jewish from non-Jewish faces. This method has good potential applicability to research on lie detection, in that it distinguishes among hits (correctly saying truth when the item does indeed show truth), false alarms (saying truth when it is actually a lie), misses (saying lie when it is actually truth), and correct rejections (correctly calling a lie a lie) (Malone & DePaulo 2001). Accuracy (sensitivity) is the excess of hits over false alarms after standardization of the relevant percentages. Sometimes the biases themselves are of theoretical interest as, for example, the so-called truth bias whereby people tend to overestimate how often targets are telling the truth or biases in how romantic partners view each other (Kenny & Acitelli 2001).

Less often used are dimensional response formats, whereby perceivers can respond on a rating scale. Sometimes the purpose of using a rating scale is to assess accuracy indirectly, as for example in lie detection research, by asking perceivers to rate how ambivalent a target seems rather than to state explicitly whether the communication is truthful or not (Malone & DePaulo 2001). Accuracy is higher to the extent that the average ambivalence rating for deceptive messages is higher than that for truthful messages.

An implicit rating scale methodology is used in some research on judging personality, wherein perceivers perform a Q-sort on adjectives describing a target (Colvin & Bundick 2001; Vogt & Colvin 2003). In a Q-sort, the perceiver places descriptive statements or adjectives into ordered piles to reflect how much each one describes the target. One can think of a Q-sort as a rating task on which mean and variance are

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 257 17.6.2005 1:59am

controlled at the outset. More explicit ratings are also sometimes made. For example, when perceivers rate the extraversion of each target represented in the stimulus set (Borkenau & Liebler 1995; Lippa & Dietz 2000), or partners in a face-to-face interaction rate the other's feelings about the self and other (Snodgrass 1985, 1992), or perceivers rate facial expressions on the degree to which each of seven emotions is present (Hall & Matsumoto 2004).

Scoring methods

On first glance, it would seem that scoring interpersonal sensitivity tasks would be a straightforward matter of comparing responses to the criterion of accuracy. For example, did the perceiver say 'anger' to an expression that actually is 'anger' according to the researcher's criterion? However, scoring accuracy tasks is not always simple, because there are alternative ways of scoring the same set of data and one way may be no more 'correct' than another. Also, the scoring options often depend on the structure of the dataset. Finally, some statistical methods for scoring accuracy are very complex and not fully accessible to many researchers. Below are some of the scoring choices facing researchers.

Percentage or mean (sum) accuracy

Using this approach, accuracy is represented by the percentage correct or the mean (or sum) of correct items. To maximize interpretability and usefulness to other researchers, whenever possible the confusion matrices should be included in research reports. To illustrate, if the stimuli consisted of six facially depicted emotions and the response alternatives consisted of the same six emotions, then the confusion matrix shows the data for all cells of this six by six array of responses. Accuracy is determined by comparing the diagonal values to off-diagonal values.

The great majority of studies of interpersonal sensitivity use percentage or mean (or sum) accuracy, without the correction (e.g. they count up the number of correct answers). The guessing or chance level of accuracy depends, of course, on how many response options are provided. For analysis of individual differences, accuracy is calculated for individual perceivers, and for group-based analysis, the mean accuracy across perceivers is used. Investigators may have further choices to create subtotal scores. For example, at the group level, accuracy can be calculated across perceivers and separately for each target, yielding encoding accuracy scores for use in analysis of individual differences among targets in how accurately they were judged. How far to subdivide accuracy scores according to individual targets or according to other variables such as sex or the different constructs being judged (e.g. different emotions) is a decision based on theoretical goals as well as the impact on reliability of decreasing the number of items included in an accuracy score. Assessment of overall accuracy typically consists of comparing the mean accuracy against the guessing level by a one-sample *t*-test. As with any such test, assuming a mean that deviates at all from the null hypothesis value, the *p*-value will become smaller with increasing sample size.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 258 17.6.2005 1:59am

Absolute discrepancy

Early accuracy researchers often calculated accuracy as the absolute difference between a perceiver's rating of a target and the criterion (typically, the target's self-rating). Cronbach (1955) and others pointed out hazards with this method, identifying several distinct reasons relating to the use of rating scales for why accuracy might be artifactually high or low. Researchers have not often used this approach in recent years. Hall *et al.* (2001) used discrepancy scores in a somewhat different way, calculating the absolute difference between how much a perceiver said the partner displayed a certain cue (rated on a scale) and how much the partner actually displayed that cue (as counted or timed by coders), after each was *Z*-scored. With this method, more accuracy is defined as smaller absolute discrepancy between these two values.

One cannot employ this metric to address the question of whether there is, or is not, a significant level of accuracy. The problem is that testing the null hypothesis that people have no accuracy tests whether a sample of judgments differs significantly from *perfect* accuracy, which is a null hypothesis of little scientific value. Testing whether judgments have some accuracy is not the same as testing whether judgments have perfect accuracy. Another problem with the difference metric is that there is no value for it that corresponds to *zero* accuracy. The point at which one moves from low accuracy (difference $>$ zero) to zero accuracy is not addressable quantitatively. However, significance tests between sample means can be meaningfully interpreted, allowing one to use this metric to determine whether accuracy increases or decreases significantly (e.g. over time after intervention, between groups).

Range scoring

Hall and Murphy (2004) measured accuracy of recalling the nonverbal behavior of a person being interviewed on videotape by asking perceivers to estimate how often each of 16 nonverbal behaviors were emitted. Because scoring an estimate as 'correct' only if it was an exact match with the criterion value would make the test prohibitively difficult, ranges were established such that approximately 50% of the participants' estimates fell within the range for a given behavior. To illustrate, if the exact count of hand gestures was 16, the range of estimates scored as accurate might fall between 13 and 19. One advantage of this method is that the investigator has good control over the difficulty level of items. Range scoring suffers from the same limitation as absolute discrepancy scoring in the sense that there is no defined 'zero-accuracy' value against which tests can be made.

Profile correlation

Profile correlation (Bernieri *et al.* 1994; Blackman 2002; Carney *et al.* 2004; Funder 1980; Vogt & Colvin 2003)—also called sensitivity correlation (Judd & Park 1993) and idiographic analysis (Kenny & Winquist 2001)—is defined as the correlation between judgments and criterion values for a given decoder. Choice is required between calculating the profile correlation across items (e.g. traits/states being judged) within targets in order to generate a separate decoding accuracy coefficient for each target, or across targets within traits/states in order to generate a separate decoding accuracy coefficient

for each trait/state. Either way, the resulting correlation is used as an accuracy score that can range from −1.0 to +1.0. (Typically, for any statistical analysis, it is normalized using the Fisher-z transformation.)

The profile correlation across items indicates how well the profile of item judgments made by the perceiver matches the profile of criterion values in the target. Using this approach, Vogt and Colvin (2003) had perceivers make Q-sorts of 100 personality-relevant adjectives after seeing a target on videotape. The perceivers' ratings (determined by which pile each trait was sorted into) were then correlated with the target's criterion values for the same 100 items, producing a separate profile accuracy correlation for each combination of perceiver and target. In this case, the *N* for the correlation was the number of traits in the analysis, or 100.

Although the metric calculated is a Pearson *r* and is interpreted as such, significance testing is problematic in that the units of analysis (items within target) are not independent. Therefore, researchers do not use this trait profile correlation to answer the question 'is *this* perceiver significantly accurate in judging this target?'. Instead, it is most often used as an effect size estimate to track how accuracy might be influenced by various other things.The present authors are not aware of any investigations that have described the precise distribution of this metric calculated in this manner, but we have little reason to doubt that the sample means of this statistic are distributed normally. Therefore, as with any other quantitative measure of accuracy, sample means can be subjected to parametric tests of significance. For example, an appropriate test of whether group mean accuracy is significantly greater than $r = 0$ is simply a one-sample *t*-test against zero, with *N* being the number of perceivers in the group. (For an alternative method for testing whether overall accuracy exceeds chance, see Kenny & Winquist 2001.) Note that for all such tests against zero, the *p*-value associated with any given non-zero value will become smaller as the number of perceivers increases.

The choice of a null hypothesis reflecting 'no accuracy' for the trait profile correlation statistic, however, is anything but straightforward. The most obvious null value for the profile correlation would be $r = 0$. Generally, a perceiver would be accurate to the extent he or she judged traits correctly as high or low. A significantly negative profile correlation would indicate a perceiver attributed the *opposite* or complementary personality to the target. But as Cronbach (1955) and others (Bernieri *et al.* 1994; Kenny & Winquist 2001) have pointed out, positive profile correlations will be achieved, in part, by some variance components that are not associated with rating a specific target uniquely. Consider the case where a perceiver is judging a target on the traits warmth, honesty, laziness, and hostility. We might assume that across the general population, the traits of warmth and honesty are more pronounced than are laziness and hostility. Without reading a target uniquely, a perceiver would register a positive profile accuracy correlation merely by rating the target more highly on the positive traits than on the negative ones. Under these circumstances, perceivers would be accurate against a null hypothesis of $r = 0$ but only because of an appreciation of how traits are manifest across the population and not because they perceived the given target accurately. In other words, they would be accurate due to the accuracy of their implicit theory of personality or *stereotype* accuracy (Bernieri *et al.* 1994; Cronbach 1955; Kenny & Winquist 2001). Psychologically mature and emotionally stable perceivers who believe people generally

like social contact and are not severely neurotic will tend to produce positive trait profile accuracy correlations across all targets. Psychotic or emotionally unstable perceivers who believe that people are generally sadistic and misanthropic would likely generate trait profile accuracy correlations that are negative or near zero, no matter who they may be judging. For this reason, it is likely that an ill-timed review of accuracy that did not consider the impact of variance components concluded that, in fact, the good judge of personality was psychologically healthy and mature (Taft 1955). (Interestingly, this conclusion appears to be correct even if it was based on questionable methodology.)

Procedures to measure and/or remove this contributing source of variance to trait profile accuracy correlations can be found in Bernieri *et al.* (1994) and in Kenny and Winquist (2001). The common theme behind these procedures is to determine the extent to which the profile correlation found is significantly higher than that which can be attributed to a generic reading of the typical human being's trait profile. The null hypothesis for these tests shifts from $r = 0$ to a correlation value that is somewhere above zero. Exactly where above zero one sets the null will differ depending on the design of the study.

In a 'perception of roommate' study where all perceivers judged one target—their own roommate—Bernieri *et al.* (1994) calculated the average trait values across all targets for each trait in the study. This generated a 'mean target trait profile'. The extent to which the perceivers' judgments happened to correlate with this mean target trait profile represented a kind of *chance* accuracy and served as a null hypothesis. In that context, a perceiver (let us say it is a female) would be accurately judging her roommate only to the extent that her judgments correlated more strongly to the target profile than to the statistically derived pseudo roommate.

A problem with this particular technique is that if a target happens to have a very typical trait profile (i.e. it happens to match closely the statistically derived pseudo target trait profile), then a perceiver will not be able to show any accuracy above the null even when her judgments match the target criterion perfectly. In other words, perceivers only have the potential for accuracy above the null to the extent that the target personality differs from the norm.

Yet another alternative null hypothesis that could be created, which is similar in spirit but addresses this problem, would be to compare the trait profile correlation to the target with all of the correlations that could be statistically derived with nontargets. In the roommate study described above, one would correlate a perceiver's single set of trait judgments made of her roommate to each of the other target profiles recorded in the study. If there were N roommate pairs in the entire sample, then each perceiver would have one trait profile correlation for her roommate, and $N-1$ trait profile correlations with the remaining pseudo roommates. The central tendency of the distribution of pseudo roommate trait profile correlations would represent a null value above which accuracy for reading the true roommate would be inferred.

Alternatively, the profile correlation can be calculated across targets within a trait or other construct. This correlation indicates how well a perceiver's profile of ratings of targets matches the actual profile of the targets on the trait. To illustrate, Lippa and Dietz (2000) correlated perceivers' ratings of extraversion for each of 32 targets (seen on

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 261 17.6.2005 1:59am

videotape for approximately 30 seconds) with those targets' self-reported extraversion. In this case, the *N* for the correlation was the number of targets (or 32), with each perceiver receiving one correlation (i.e. accuracy score) which represented his/her accuracy across all targets for the trait of extraversion. For this design, the null value of $r = 0$ is appropriate.

As noted by Bernieri *et al.* (1994), Kenny and Winquist (2001), and others, profile correlations and many other scoring methods can be subjected to componential analyses (conceptually, if not empirically) to acknowledge different sources of variance in the accuracy scores (see Cronbach 1955). Methods for correcting or adjusting scores for stereotype accuracy and elevation are discussed by Bernieri *et al.* (1994) as well as by Kenny (1994; Kenny & Winquist 2001) and others.

Group accuracy correlations

All of the preceding methods begin by computing interpersonal accuracy for individual perceivers before averaging or summarizing them to arrive at a group mean estimate. Accuracy can also be computed at a sample level directly, without ever calculating any one perceiver's level of accuracy. Such metrics can be appropriate for exploring group-level hypotheses (e.g. investigating gender, age, and role effects). For example, Hall *et al.* (2001) compared the accuracy of high-status versus low-status participants in their recall of their partner's nonverbal behaviors using such a method.

When group-level effects are of primary concern and individual differences within groups are considered error variance, then group accuracy correlations are the appropriate metric. In some instances, a sample or group of individuals might only generate one single judgment outcome, as in the case of a search committee's final assessment of a job candidate. The group in essence becomes a single perceiver. Group correlations of this type are conceptually similar to those formed for any single perceiver. Below we describe two different kinds of group accuracy scoring methods.

Pooled consensus accuracy: group accuracy correlations reflecting the accuracy of a group's mean judgment

In this method, the ratings made by perceivers are averaged before any other calculations are made. Conceptually, a group, committee, or sample is considered 'the perceiver', from which a single judgment is generated. The group consensus can be derived by simply averaging the individual ratings or it could conceivably be the outcome of a group discussion. When a group judges either a sample of targets on one item or judges one target across a set of items (i.e. profile), a correlation is computed with the criterion values where the null hypothesis of no accuracy is defined as $r = 0$. As illustration of this method, Borkenau and Liebler (1995) averaged perceivers' ratings of extraversion for each of 50 targets and then correlated these mean ratings with the criterion values derived from targets' self-ratings of extraversion. (For further illustrations of this method, see Zebrowitz *et al.* 2002 and Hall & Carter 1999.)

A notable feature of this method is that the magnitude of the resulting group correlation tends to be larger than the average accuracy correlations individually

calculated (Ambady *et al.* 2000; Kunda 1999). As examples, Bernieri *et al.*(1996) found that pooled consensus judgments of target rapport correlated around $r = 0.29$ with targets' own ratings of rapport, whereas the average accuracy for individual perceivers was only around $r = 0.20$. Watson (1989) found that perceivers' ratings of target persons' personality traits correlated with the targets' self-ratings more strongly when five perceivers' ratings were pooled before calculating the profile correlation than when profile correlations were calculated for individual perceivers and then averaged (e.g. for agreeableness, $r = 0.16$ vs. $r = 0.10$). Finally, Hall and Carter (1999), in studying the accuracy of people's beliefs about sex differences, found the profile accuracy of pooled judgments to be $r = 0.70$, whereas the average individual profile accuracy correlation was only around $r = 0.45$.

The relative gains in accuracy attained by pooled consensus judgments will be driven by reliability issues. In general, these group accuracy correlations will increase as the N of perceivers within the group increases, due to the cancellation of random error. They might also increase to the extent that the different perceivers in the group are accurately detecting different sources of criterion variance from each other. For example, imagine a five-member committee where each member's judgment across targets correlates $r = 0.30$ with the criterion but they each employ a different and orthogonal cue or strategy to achieve this degree of accuracy. The resulting consensus judgment might then be understood as a composite of five orthogonal valid predictors. The gain in accuracy from pooling in this case would result more from the combination of orthogonal sources of true accuracy variance than it would from simply increasing N (i.e. the cancellation of random errors) (Guilford 1954).

The group accuracy correlation with perceivers as the units of analysis

This approach is commonly referred to as nomothetic analysis because one focuses primarily on the situational factors that influence the accuracy of a given sample of perceivers (Kenny & Winquist 2001). The judgments of a sample of perceivers are correlated with the criterion values for a set of targets, items, or relationships, with each perceiver–target dyad being a unit of analysis and N being the number of such dyads. Examples of this are studies that investigate accuracy of roommates' or friends' perceptions of one another or accuracy of doctors' and patients' impressions of how much they are liked by the other (e.g. Bernieri *et al.* 1994; Funder *et al.* 1995; Hall *et al.* 2002).

In this method, a single correlation coefficient assesses the degree of correspondence between perceiver judgment and criterion across the sample of perceivers. The group accuracy correlation is a sample statistic. There is no component of it that relates to the accuracy of any individual perceiver. The null hypothesis reflecting no accuracy is represented by $r = 0$, where the df is $N-2$ perceivers. For any given positive non-zero value of r, the p-value for the significance test against zero will increase as a function of N.

Lens model analysis

Brunswik (1956) was concerned with modeling the accuracy with which perceivers could assess their physical environments. He proposed a methodology where, after a

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 263 17.6.2005 1:59am

perceiver judged a set of stimuli, his or her judgments would be subjected to a regression analysis to learn how they correlated with the criterion as well as with a series of potential judgment mediators or cues. The correlation between a perceiver's judgment of an attribute (e.g. size) over a series of stimuli and the criterion values for those stimuli constitute a perceiver's accuracy, which Brunswik termed *achievement*.

Heider (1958) was the first to embrace the lens model as a conceptual framework with which to describe interpersonal perception. However, it was not until many years later that the employment of the lens model for interpersonal perception began in earnest (e.g. Funder & Sneed 1993; Gifford 1994; Scherer 1982). This model has been useful in documenting not only how accurately people make judgments, but also what cues influence a perceiver's judgments and how the cues themselves are related to the criterion (i.e. ecological validity).

The lens model has been applied to study vocal behavior (Scherer 1978), personality traits (Borkenau & Liebler 1995; Gifford 1994; Lippa 1998), intelligence (Borkenau & Liebler 1995; Murphy *et al.* 2003; Reynolds & Gifford 2001), and rapport (Bernieri & Gillis 2001; Bernieri *et al.* 1996; Gillis & Bernieri 2001; Gillis *et al.* 1995). To apply this model, the researcher:

1. establishes criterion values of the state or trait in question for each target in the stimulus set;
2. gathers perceivers' judgments of the targets on the state or trait in question, and;
3. codes or otherwise assesses the targets on cues that may be relevant to the state or trait.

(See Chapter 3 for more detailed description.)

To illustrate, Gifford (1994) established the personality dominance of 60 targets through self-ratings and had perceivers rate the targets on dominance from videotaped interactions. He also measured how much the targets gestured and how much they manipulated an object (e.g. jewelry, shirt sleeve, pen). The correlation between perceiver judgments and the targets' true dominance, as defined by the criterion, constituted their level of achievement (i.e. dominance judgment accuracy) and was around $r = 0.26$. The correlation between target dominance and coded behavior constituted the ecological validity of the cues examined and was $r = 0.29$ for gestures and $r = -0.46$ for object manipulation. This suggested that target dominance could be perceived by observers looking for more gestures and little object manipulation. In fact, perceiver judgments of dominance correlated highly with gestures ($r = 0.66$), suggesting that this is precisely how perceivers looked for dominance. Interestingly, perceiver judgments did not correlate significantly with object manipulation, indicating that observers overlooked an important cue, given that this cue had the strongest relationship to the dominance criterion (Gifford 1994).

The lens model can be constructed either for a perceiver group as a whole or for individual perceivers whose coefficients (accuracy correlations and correlations indicating their judgment policies) can be averaged for purposes of summary display (Bernieri *et al.* 1996; Bernieri & Gillis 2001). A few critical points need to be made when a lens model analysis is performed on pooled consensus judgments. First, pooling perceiver judgments before correlating them with a target criterion defines a group

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 264 17.6.2005 1:59am

correlation (see p. 00). Therefore, increases in the reliability of the perceiver judgment should increase achievement (accuracy) as the number of perceivers being pooled increases. Second, the cue utilization output of the lens model analysis based on pooled consensus judgments cannot be taken as representative or typical of the individual perceivers that contributed to it. There is no mathematical reason to necessitate that any *one* perceiver utilized the cues in the manner consistent with the apparent cue utilization of the pooled judgment to which they contributed. To illustrate, suppose there were three perceivers who were all judging the dominance of a set of targets. Imagine further that all had a different single-cue judgment such that perceiver A was influenced by head nods, perceiver B by gestures, and perceiver C by forward lean. The lens model of their pooled judgments in this case would likely reveal a complex judgment policy consisting of all three cues—a policy employed by no one. Finally, to the extent that individual perceivers have judgment policies that consist of *opposite* cue dependencies (i.e. some perceivers are tracking in a positive direction while others are tracking the same cue in a negative direction), lens models of pooled consensus judgments will not detect the actual judgment policies being employed at the individual perceiver level.

Sometimes, investigators using the lens model are interested in determining which cues can account for accuracy— in other words, which cues can be called mediators of the judgment–accuracy link. Inspection of the corresponding correlations (as in the comparison of 0.29 to 0.66 in the Gifford example given above) gives an indication; if both correlations are positive, that cue (in this case, gestures) is a good candidate as a mediating cue. Two points must be made about such a conclusion. First, an eyeball conclusion is not as sound as a statistical test. For this purpose, statistical tests are available—for example, the Sobel test (Kenny *et al.* 1998; for applications in the context of the lens model, see Murphy *et al.* 2003 and Zebrowitz *et al.* 2002). Second, even if such a test shows that accuracy can be statistically accounted for with a given cue, because of the correlational nature of such data, there is no way to know for sure that this cue was actually used. An alternative possibility is that perceivers used a different (possibly unmeasured) cue, but that cue was highly correlated with the measured cue. In the Gifford example of gestures described above, such a correlated cue might have been speaking time. Because people gesture when speaking and hardly at all when not, perceivers may have been making their judgments based on seeing how much the targets spoke and not on their gesturing.

Variance components approach

The final scoring method we will discuss is the variance components or Social Relations Model (SRM) approach developed by Kenny (Kenny 1994, 2001; Kenny & Winquist 2001). Among all the methods described, the variance components approach has the most potential for comprehensiveness in that the analysis is limited only by the amount and quality of the data available. Inspired by Cronbach's (1955) admonition that accuracy researchers should be mindful of the sources of variance that contribute to a given accuracy metric, the approach attempts to identify and partition, exhaustively, all of the variance components possible, given the data matrix.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 265 17.6.2005 1:59am

Perhaps the best way to understand the SRM is to imagine what the ideal or perfect data set for learning about interpersonal perception accuracy would look like from an analysis of variance (ANOVA) framework. If one could imagine every person in a population rating every other person in that population, on every possible attribute, in every possible situation/role, over an infinite number of trials, then one could theoretically identify precisely what contributed to a given perceiver's accuracy in any single cell within that perfect data matrix, given one had the corresponding criterion data. One would be able to partition perfectly the extent to which the accuracy was due to such things as:

1. this particular perceiver's judgment biases;
2. the perceiver's general knowledge of how various traits and other attributes hang together in humans;
3. how well this particular perceiver accurately perceives this particular target, trait, or situation/time uniquely.

In this hypothetical data matrix, no significance testing would be needed because we would simply be describing the parameters of the population and individual score components.

In essence, the various statistical procedures developed from the SRM are all designed to estimate as many sources of variance as possible given the completeness of the data matrix provided (Kenny & Winquist 2001). Although the general method is expandable to allow the examination of almost any theoretical effect on accuracy one could conceive, its most representative form partitions accuracy into the four components identified by Cronbach (Cronbach 1955; Kenny 1994; Kenny & Albright 1987). When multiple targets are judged by multiple perceivers on a single measure, and the criterion data are recorded on a similar scale, then one can determine accuracy due to:

1. *elevation*—the degree of correspondence between the perceiver's judgment grand mean and the criterion grand mean;
2. *differential elevation*—the degree of correspondence between a target's variation from the sample target mean and the variation of the perceiver's judgment of that target from the perceiver's mean;
3. *stereotype*—understood in terms of a perceiver's knowledge of the population mean value of the trait/attribute in question;
4. *dyad*—the extent to which a perceiver's judgments of a target correspond to that target's behavior as it is uniquely expressed towards the perceiver.

Calling the variance components approach a 'scoring method' hardly does it justice, as it essentially attempts to marry the proper statistical analysis to any given experimental design and item scaling, and results in a unique output of variance components for that particular paradigm. The most notable feature of this approach is that the interpretation of the outputs from these analyses can vary greatly in their utility. To date, the SRM has been employed primarily for determining the relative magnitudes of accuracy components (e.g. when people are accurate is it because they know the generic person stereotype or is it because they perceive the uniqueness of the target?). It was not developed to assess an individual perceiver's accuracy. It describes accuracy at the level

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 266 17.6.2005 1:59am

of a sample and, thus, is conceptually similar to *group accuracy* discussed earlier. As such, this approach is not well suited to exploring trait moderators of an individual's accuracy. However, SRM should be ideal for exploring situational and role moderators of accuracy, although this particular use for this technique has yet to be fully exploited.

Another issue influencing the utility of this approach involves the assumption and necessary interpretation of accuracy as being componential. Although the various sources of variance are mathematically defined and easily discussed in terms of the rows and columns of a data matrix, some have wondered whether the decomposition of accuracy undermines the integrity and interpretability of the accuracy construct itself (Funder 2001; Vogt & Colvin 2003). What one researcher might label artifact, another might consider accuracy. The theoretical utility and validity of an accuracy construct that is left after several of its variance components have been removed will have to be defended.

Nevertheless, ways of analyzing accuracy within the SRM are evolving, to fit the increasing number of experimental designs (Kenny & Winquist 2001). But, more importantly, the framework is expanding to handle an increasing number of factors that were never before incorporated into accuracy theory. For example, Kenny and Winquist (2001) describe the procedures one would use if targets were judged on multiple measures instead of one. The analysis of this design extension doubles the number of accuracy components from four to eight. The potential of this method is great because there are other factors not yet explored (e.g. roles, situation, time, medium, domain of attribute being assessed) that could be incorporated into this model, generating even more variance components. For example, in the future, one might use SRM to address the questions 'to what extent is a perceiver accurate in judging Professor Jones because of an intuitive understanding of how professors typically behave?' and 'are people more accurate in assessing traits, states, or the behavioral cues that express them?'. The SRM model is flexible enough to be modified to handle many such research questions not yet framed.

Instruments for measuring interpersonal sensitivity

Thus far, we have led the reader from broad conceptual issues to more concrete, operational issues (and problems) in the assessment of interpersonal sensitivity. It is now time to describe specific instruments in more detail, some of which we have mentioned in the preceding pages. These instruments vary greatly in how frequently they have been used in published research. (Methodological descriptions of other measurement paradigms that we described earlier can be found in the citations provided with those descriptions.)

Profile of Nonverbal Sensitivity Test (PONS; Rosenthal *et al.* 1979)

This test measures accuracy of inferring the emotional tone of scenes acted out by one female expressor. The full-length PONS test consists of 220 2-second audio clips, video clips, or combined audio and video, and a printed answer sheet containing 220 pairs of brief verbal descriptions that are responded to in multiple-choice style. The PONS test

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 267 17.6.2005 1:59am

has an *a priori* structure consisting of 11 channels: three visual channels alone, with no voice; two vocal channels alone, with no visual cues; and six combinations of the three visual and two vocal cues. These 11 channels are crossed by 20 affective scenes (e.g. returning a faulty item to a store, expressing jealous rage) which themselves fall into a two-by-two configuration (with five scenes each) representing the dimensions of positivity (positive vs. negative) and dominance (more dominant vs. less dominant). The test was made in 1971, and the monograph describing the test's development and validation was published in 1979 (Rosenthal *et al.* 1979). Internal consistency (KR-20) for the full-length test is 0.86, and test–retest reliability is 0.69 (median over six samples). Short forms of the test include the 40 audio-only items, the 40 face and body items, and the 60 face, body, and face-plus-body items shown as still photographs.

Among the many validity results published in Rosenthal *et al.* 1979, and since, are the following: people who scored higher on the PONS had healthier, more well-adjusted personalities, were rated as more interpersonally sensitive by peers or supervisors, were more democratic teachers, and were rated as better in their job performance as clinicians and teachers. The test also shows developmental and sex differences (females score higher).

Empathic cccuracy standard-cue methodology (Ickes 1997, 2001)

Empathic inference is the 'everyday mind reading' that people do whenever they attempt to infer other people's thoughts and feelings, and empathic accuracy is the extent to which such inferences are accurate. The basic methodology was described earlier in this chapter. In the standard-cue variant of this methodology, a collection of videotaped expressors can be shown to perceivers who were not the original interaction partners (e.g. Gesn & Ickes, 1999). The tape is stopped at the precise moments that the expressor indicated that he or she thought or felt something, and perceivers indicate, in an open-ended fashion, what the expressor was feeling at that moment. Responses are then scored for accuracy on a 0 (*not accurate at all*) to 2 (*maximally accurate*) response scale. A perceiver's responses can then be summed across judged expressors and used as an individual difference measure of empathic accuracy. Inter-rater alphas tend to be around 0.90 and the general empathic accuracy paradigm shows predictive validity (Marangoni *et al.* 1995; Stinson & Ickes 1992).

Interpersonal Perception Task (IPT; Costanzo & Archer 1989)

The IPT shows videotaped clips of varying lengths (approximately one minute) in which people are shown in interaction or speaking to the camera.. Both nonverbal and verbal cues are presented in a full-channel audiovisual mode (verbal cues, though present, are designed not to be informative). The IPT has both 30-item and 15-item versions (Archer *et al.* 2001). Criteria for scoring this multiple-choice test are objective facts about the expressors or the circumstances under which their cues were expressed. An example would be a videotaped clip of a woman talking to someone on the telephone—is she talking to her mother or her boyfriend? Or, a person is telling aspects of her life story—is it a true story? For each IPT scene there is an interpretive question,

and, for each question, there is an objectively correct answer. Each of the IPT scenes taps one of five interpretive domains: kinship, lie-telling, competition, status, and intimacy. Test–retest reliability is $r = 0.70$ (Costanzo & Archer 1989).

Diagnostic Analysis of Nonverbal Accuracy (DANVA; Nowicki & Duke 1994, 2001)

Several tests to measure sensitivity to nonverbal cues of emotion are collectively called the DANVA. The basic test stimuli of the DANVA are predominantly posed photographs and audio recordings (of a single standard sentence). The DANVA measures are 24-item measures which tap four different emotions: happy, sad, angry, and fearful. In the most widely used test, perceivers view pictures of 24 adults posing a facial expression of emotion for several seconds and then choose the emotion word that best represents the facial expression. The audio version of the task is the same except that the stimuli are voices reading the standard sentence. Responses are then scored for accuracy against correct answers. Correct answers are determined by selecting items that had a high degree of inter-judge agreement (i.e. $> 80\%$). Additional versions of the DANVA include tests with children and African Americans as expressors. Internal consistency is usually higher than 0.70 for all DANVA tests, and the tests have predictive validity and correlations with an array of personality constructs (Nowicki & Duke 2001).

Missing Cartoons Test (deMille *et al.* 1965)

This test is a 28-item measure of social situation decoding ability in which respondents are asked to choose the missing cartoon segment that belongs in a four-segment cartoon strip. Each four-segment strip depicts an ambiguous social situation where one of the four segments is missing and the correct cartoon segment that completes the sequence is listed below the strip along with three incorrect choices. The ambiguous social situations contain overt cues such as those associated with behavior, and less overt cues such as those associated with thoughts and feelings. The number of items answered correctly is the accuracy score. This measure has recently shown adequate internal consistency (alpha = 0.76) and predictive validity (e.g. Carney & Harrigan 2003).

Child and Adolescent Social Perception Measure (CASP; Magill–Evans *et al.* 1995)

The CASP consists of 10 naturalistic scenes acted out by children and adolescents, presented on videotape with electronic filtering of the soundtrack to prevent verbal comprehension. Test takers do not mark a preprinted answer sheet but, rather, respond in an open-ended fashion to probes about which emotions they perceive. A standardized scoring system is used. This test shows age and gender differences in a normative sample of children, differentiates normally functioning adolescents from a sample with Asperger's syndrome, and correlates with parent and teacher ratings of functioning (Koning & Magill–Evans 2001). Internal consistency is 0.88 (Magill–Evans *et al.* 1995).

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 269 17.6.2005 1:59am

CARAT (Buck 1976)

The 'slide-viewing technique' (Buck 1979) measures spontaneous expression of emotional cues (i.e. cues that are revealed on the target's face without awareness or intention). Although variations on this technique exist, the basic paradigm consists of showing affectively arousing color slides to individuals whose faces are surreptitiously recorded on a videotape as they watch the slides. The videotape is then shown to naïve judges, who are typically asked which slide was being viewed. If the slide can be accurately identified, one can infer that the expressor unintentionally revealed his or her emotional response to the slide. Test–retest reliability is 0.80 (Buck 1976).

Most research using the slide-viewing technique is concerned with accuracy of expressing, not accuracy of judging. However, Buck (1976) incorporated one set of these facial expressions into a judgmental accuracy test. Test takers view a series of faces and can respond either categorically, as described earlier (i.e. which slide was being viewed), or dimensionally, by rating how pleasant the expressor rated his or her own experience. Accuracy for the dimensional measure (which Buck called the 'pleasantness' measure) consists of correlating these ratings with the original expressor's ratings across the different expressor's expressions. Thus, there are two different criteria for accuracy: the category of the slide (sexual, unusual, scenic, and unpleasant) and the expressor's self-ratings of pleasant affect.

The CARAT has not been extensively used in individual difference research, though validational findings have been reported (Buck 1976; Hall *et al.* 1997).

Japanese and Caucasian Brief Affect Recognition Test (JACBART; Matsumoto *et al.* 2000)

The JACBART was developed as an improved version of the Brief Affect Recognition Test (BART; Ekman & Friesen 1974). The BART consists of photographs of faces showing 'basic' emotions identified in the research program of Paul Ekman (happiness, sadness, disgust, fear, surprise, and anger), presented tachistoscopically (e.g. less than 1/25 second; Ekman & Friesen 1974). Test takers choose which emotion was shown from a multiple choice. The JACBART (Matsumoto *et al.* 2000) is similar to the BART, except that it contains equal numbers of Japanese and Caucasian, and male and female, expressions of seven different emotions (happiness, sadness, disgust, fear, surprise, anger, and contempt) and each expression is 'sandwiched' between two neutral expressions made by the same expressor. The entire test is on videotape (not requiring special equipment). Exposures are 1/5 second or shorter.

Reliability is high (internal consistency >0.80, retest $r = 0.78$), and scores correlate with an array of personality measures. Scoring can be done categorically or using individual profile correlations (a separate correlation for each perceiver for each item, where the correlation is between the perceiver's ratings of how much each of seven emotions was shown in the expression and how much a normative sample who viewed the expressions for 10 seconds said each emotion was shown in the expression; Hall & Matsumoto 2004).

Pictures of Facial Affect (POFA; Ekman & Friesen 1976)

The POFA is a set of 110 black and white photographs of 14 different individuals (eight female and six male) expressing six different 'basic' emotions—happy, sad, angry, fearful, surprised, and disgusted, plus a neutral picture for each individual expressor. The development and validation of the POFA was based on Ekman and Friesens's Facial Action Coding System (FACS; Ekman & Friesen 1978). FACS is a coding system that maps facial muscle configurations to different emotional experiences, and was used to select photographs for the POFA—that is, FACS-determined muscular configurations were used as criteria in choosing stimuli that accurately represent each of the six basic emotions.

The POFA is considered to be more of a stimulus set than a standardized test of emotion decoding ability. However, it can be used as an individual difference measure of emotional sensitivity and has been used in a large number of studies. When the POFA is used in research, a selection of the photographs (or sometimes the whole set) is presented to participants either in photograph, slide, or digitized (i.e. on a computer screen) form. Participants are then asked to judge the emotion being expressed in the photograph and respond on a forced-choice scale (e.g. McAndrew 1986; Niit & Valsiner 1977). In some studies though, participants were allowed to respond in an open-ended response format (e.g. Boucher & Carlson 1980).

The validity of the POFA expressions is supported by the numerous studies that have shown consensus within and across cultures (see Russell 1994). Generally, the psychometric properties of the POFA are largely assumed, since its construction was based on an extremely elaborate and highly reliable and valid system for determining and labeling the intensity of facial expressions of affect (i.e. FACS). Recent work has added validity data as described in the following sampling of results.

The POFA shows developmental effects (Lenti *et al.* 1999), relations to personality (Larkin *et al.* 2002), and differences between psychiatric and learning disabled groups and normally functioning groups (Holder & Kirkpatrick 1991; Rojahn *et al.* 2002). Wallbott (1991) showed that people could accurately identify which of the POFA slides they were viewing from videotapes of themselves while viewing the pictures. Dimberg and Thunberg (1998) showed that the POFA predicted facial activity (measured with EMG) consistent with the POFA pictures shown at 300–400 milliseconds (almost subliminal) of exposure. (Also see Dimberg 1997 and Dimberg *et al.* 2000 for additional evidence of this effect.)

Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT; Mayer *et al.* 2003)

Emotional intelligence (EI) is defined as a set of skills concerned with the processing of emotion-related information, with emotion decoding accuracy at its theoretical (Mayer & Salovey 1997) and measurement (Mayer *et al.* 2003) core. The construct of EI is most reliably and validly measured with the MSCEIT, Version 2.0 (Mayer *et al.* 2003). The MSCEIT measures each component and associated sub-components of the four-branch model proposed by Mayer and Salovey (1997):

1. the accurate perception of emotion;
2. the use of emotion to facilitate cognitive activities;

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 271 17.6.2005 1:59am

3. the ability to understand emotion;
4. regulation/management of emotion.

Because Mayer and Salovey's (1997) model of EI hypothesized that emotional knowledge is embedded within a social context of communication and social interaction, the criterion scores, or 'right answers', for each of the items on the MSCEIT are based on hundreds of raters' average scores on each item. Split-half reliabilities for each of the four MSCEIT branches and associated subtests ranged from 0.64 to 0.93 on a large diverse sample of individuals (Mayer *et al.* 2003). The test–retest reliability of the total MSCEIT score has been reported as $r = 0.86$ (Brackett & Mayer 2001). The components of the MSCEIT are moderately intercorrelated yet distinct. Indeed, confirmatory factor analytic results support Mayer and Salovey's (1997) four-component EI model (Mayer *et al.* 2003).

Although the MSCEIT has been the subject of considerable development and reliability work, the validity of this test is just beginning to emerge. Theoretically, EI should be related to a number of adaptive interpersonal consequences, and the usefulness of the MSCEIT in attempting to predict such positive outcomes is without question. However, whether such relations exist is only beginning to receive attention now that the MSCEIT has been fully developed.

For our purposes, the emotion-decoding aspect of the MSCEIT has the greatest relevance, as only this portion of the test is concerned with interpersonal sensitivity as opposed to other emotion-related traits and tendencies.

Conclusions

Researchers wanting to measure interpersonal sensitivity have many choices to make, both conceptual and operational. It is obvious that theory should guide method—in this case, that the questions one wants to ask should guide the choice of instruments and methods. However, this is easier said than done. Not enough is yet known about the landscape of interpersonal sensitivity to know with any certainty how research questions do map onto instruments and methods. If we want to measure the impact of some variable on interpersonal sensitivity, how do we choose which domain of sensitivity to examine? If we know what domain we want (e.g. sensitivity to emotions), which of the several available instruments should we pick? Which emotions should we include? Should we use still photos or moving video, or posed or spontaneous expressions? Is it all right to develop one's own instrument? Which of many possible criteria and scoring methods should we pick? Regrettably, research at this stage in the field's development does not offer good answers to such questions. Therefore, researchers can hardly be faulted if they just use the instrument they are most familiar with, or that someone recommends to them, or that is most often cited, or that is most convenient to acquire and use. Often we lack the empirical knowledge on which to make a more empirically grounded choice.

With the passage of time, this situation should improve. In the meantime, we would caution researchers not to generalize too much beyond the instruments, methods, and operational definitions that they have used. Findings for one kind of accuracy (e.g.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 272 17.6.2005 1:59am

domain, method) may not apply to another kind of accuracy. If resources allow, multiple methods should be used. Meta-analyses that examine results for different methodologies are especially valuable (Elfenbein & Ambady 2002; Hall 1978; McClure 2000). Understanding the impact of methodology on results is, of course, an important step toward developing a theory of interpersonal sensitivity (Zebrowitz 2001).

References

Allport, G.W. & Kramer, B.M. (1946). Some roots of prejudice. *Journal of Psychology,* **22**, 9–39.

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,** 287–96.

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**, 947–61.

Ambady, N. & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin,* **111**, 256–74.

Ambady, N. & Rosenthal, R. (1993). Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology,* **64**, 431–41.

Ambady, N., Bernieri, F., & Richeson, J.A. (2000). Towards a histology of social behavior: judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology,* **32**, 201–71.

Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology,* **77,** 538–47.

Ambady, N., LaPlante, D., & Johnson, E. (2001). Thin-slice judgments as a measure of interpersonal sensitivity. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 89–101. Mahwah, NJ: Erlbaum.

Archer, D. & Akert, R.M. (1980). The encoding of meaning: a test of three theories of social interaction. *Sociological Inquiry,* **50**, 393–419.

Archer, D., Costanzo, M., & Akert, R. (2001). The Interpersonal Perception Task (IPT):alternative approaches to problems of theory and design. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 161–82. Mahwah, NJ: Erlbaum.

Barnes, M.L. & Sternberg, R.J. (1989). Social intelligence and judgment policy of nonverbal cues. *Intelligence,* **13**, 263–87.

Baum, K.M. & Nowicki, S., Jr. (1998). Perception of emotion: measuring decoding accuracy of adult prosodic cues varying in intensity. *Journal of Nonverbal Behavior,* **22**, 89–107.

Bernieri, F.J. (1991). Interpersonal sensitivity in teaching interactions. *Personality and Social Psychology Bulletin,* **17**, 98–103.

Bernieri, F.J. (2001). Toward a taxonomy of interpersonal sensitivity. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 3–19. Mahwah, NJ: Erlbaum.

Bernieri, F.J. & Gillis, J.S. (2001). Judging rapport: employing Brunswik's lens model to study interpersonal sensitivity. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 67–88. Mahwah, NJ: Erlbaum.

Bernieri, F., Gillis, J.S., Davis, J.M., & Grahe, J.E. (1996). Dyad rapport and the accuracy of its judgment across situations. *Journal of Personality and Social Psychology,* **71**, 110–29.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 273 17.6.2005 1:59am

Bernieri, F.J., Zuckerman, M., Koestner, R., & Rosenthal, R. (1994). Measuring person perception accuracy: another look at self-other agreement. *Personality and Social Psychology Bulletin*, **20**, 367–378.

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., *et al.* (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): reliability data and cross-national differences. *Journal of Nonverbal Behavior,* **21**, 3–21.

Blackman, M. (2002). The employment interview via the telephone: are we sacrificing accurate personality judgments for cost efficiency? *Journal of Research in Personality,* **36**, 208–23.

Bollen, K. & Lennox, R. (1991). Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin,* **110**, 305–14.

Borkenau, P. & Liebler, A. (1995). Observable attributes and manifestations and cues of personality and intelligence. *Journal of Personality,* **63**, 1–25.

Bornstein, M.H. & Arterberry, M.E. (2003). Recognition, discrimination and categorization of smiling by 5-month-old infants. *Developmental Science,* **6**, 585–99.

Boucher, J.D. & Carlson, G.E. (1980). Recognition of facial expression in three cultures. *Journal of Cross-Cultural Psychology,* **11**, 263–80.

Brackett, M. & Mayer, J.D. (2001). *Comparing measures of emotional intelligence.* Paper presented at the Third Positive Psychology Summit, Washington, DC, October 2001.

Brackett, M. & Mayer, J.D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin,* **29**, 1147–58.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley, CA: University of California Press.

Buck, R. (1976). A test of nonverbal receiving ability: preliminary studies. *Human Communication Research,* **2**, 162–71.

Buck, R. (1979). Measuring individual differences in nonverbal communication of affect: the slide-viewing paradigm. *Human Communication Research,* **6**, 47–57.

Carney, D.R. (2004). *In the face of prejudice: nonverbal expression and accurate detection of explicitly and implicitly measured anti-Black attitudes.* Unpublished doctoral dissertation, Northeastern University, Boston, MA.

Carney, D.R. & Harrigan, J.A. (2003). It takes one to know one: interpersonal sensitivity is related to accurate assessments of others' interpersonal sensitivity. *Emotion,* **3**, 194–200.

Carney, D.R., Colvin, C.R., & Hall, J.A. (2004). *What, when, and for how long? A look at judgmental accuracy from thin slices of social behavior.* Manuscript submitted for publication.

Carney, D.R., Hall, J.A., & Smith LeBeau, L. (in press). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior.*

Colvin, C.R. & Bundick, M.J. (2001). In search of the good judge of personality: some methodological and theoretical concerns. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 47–65. Mahwah, NJ: Erlbaum.

Costanzo, M. & Archer, D. (1989). Interpreting the expressive behavior of others: the Interpersonal Perception Task. *Journal of Nonverbal Behavior,* **13**, 225–45.

Cronbach, L.J. (1955). Processes affecting scores on 'understanding of others' and 'assumed similarity'. *Psychological Bulletin,* **52**, 177–93.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 274 17.6.2005 1:59am

Cronbach, L.J. (1990). *Essentials of psychological testing*, 5th edn. New York: Harper Collins.

Davis, M.H. & Kraus, L.A. (1997). Personality and empathic accuracy. In *Empathic accuracy* (ed. W. Ickes), pp. 144–68. New York: Guilford.

deMille, R., O'Sullivan, M., & Guilford, J.P. (1965). *Missing Cartoons—Form A*. Beverly Hills: Sheridan Supply Company.

DePaulo, B.M., Charlton, K., Cooper, H., Lindsay, J.J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review,* **1**, 346–57.

deTurck, M.A. (1991). Training observers to detect spontaneous deception: effects of gender. *Communication Reports,* **4**, 79–89.

de Waal, F.B.M. (ed.). (2001). *Tree of origin: what primate behavior can tell us about human social evolution.* Cambridge, MA: Harvard University Press.

DiMatteo, M.R., Friedman, H.S., & Taranta, A. (1979). Sensitivity to bodily nonverbal communication as a factor in practitioner–patient rapport. *Journal of Nonverbal Behavior,* **4**, 18–26.

Dimberg, U. (1997). Facial reactions: rapidly evoked emotional responses. *Journal of Psychophysiology,* **11**, 115–23.

Dimberg, U. & Thunberg, M. (1998). Rapid facial reactions to emotional facial expressions. *Scandinavian Journal of Psychology,* **39**, 39–45.

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science,* **11**, 86–9.

Dimitrovsky, L. (1964). The ability to identify the emotional meaning of vocal expressions at successive age levels. In *The communication of emotional meaning* (ed. J.R. Davitz). New York: McGraw–Hill.

Dorfman, D.D., Keeve, S., & Saslow, C. (1971). Ethnic identification: a signal detection analysis. *Journal of Personality and Social Psychology,* **18**, 373–9.

Dunlap, K. (1927). The role of eye muscles and mouth muscles in the expression of the emotions. *Genetic Psychology Monographs*, **2**, 199–233.

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological Bulletin,* **115**, 268–87.

Ekman, P. & Friesen, W.V. (1972). Hand movements. *Journal of Communication,* **22**, 353–74.

Ekman, P. & Friesen, W.V. (1974). Nonverbal behavior and psychopathology. In *The psychology of depression: contemporary theory and research* (ed. R.J. Friedman & M.M. Katz). Oxford, UK: Wiley.

Ekman, P. & Friesen, W.V. (1976). *Pictures of Facial Affect.* Palo Alto, CA: Consulting Psychologists' Press.

Ekman, P. & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist,* **46**, 913–20.

Ekman, P. & Rosenberg, E. (1997). *What the face reveals: basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* New York: Oxford University Press.

Ekman, P., Davidson, R.J., & Friesen, W.V. (1990). The Duchenne smile: emotional expression and brain physiology II. *Journal of Personality and Social Psychology,* **58**, 342–53.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial Action Coding System.* Salt Lake City, UT: A Human Face.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 275 17.6.2005 1:59am

Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni–Tarlatzis, I., Heider, K., *et al.* (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology,* **53**, 712–17.

Elfenbein, H.A. & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin,* **128**, 203–35.

Elfenbein, H.A. & Ambady, N. (2003). When familiarity breeds accuracy: cultural exposure and facial emotion recognition. *Journal of Personality and Social Psychology,* **85**, 276–90.

Fridlund, A.J. (1997). The new ethology of human facial expressions. In *The psychology of facial expression* (ed. J.A. Russell & J.M. Fernández–Dols), pp. 103–29). Paris: Cambridge University Press.

Fruzzetti, A.E., Toland, K., Teller, S.A., & Loftus, E.F. (1992). Memory and eyewitness testimony. In *Aspects of memory, practical aspects* (2nd edn) (ed. M.M. Gruneberg & P.E. Morris), pp. 18–50. Florence, KY: Taylor & Frances.

Funder, D. (2001). Three trends in current research on person perception: positivity, realism, and sophistication. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 319–31. Mahwah, NJ: Erlbaum.

Funder, D.C. & Harris, M.J. (1986). On the several facets of personality assessment: the case of social acuity. *Journal of Personality,* **54**, 528–50.

Funder, D.C. & Sneed, C.D. (1993). Behavioral manifestations of personality: an ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology,* **64**, 479–90.

Funder, D.C., Furr, R.M., & Colvin, C.R. (2000). The Riverside Behavioral Q-Sort: a tool for the description of social behavior. *Journal of Personality,* **68**, 451–89.

Funder, D.C., Kolar, D., & Colvin, R.C. (1995). Agreement among judges of personality: interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology,* **69**, 656–72.

Gesn, P.R. & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: channel and sequence effects. *Journal of Personality and Social Psychology*, **77**, 746–61.

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.

Gilbert, D.T. & Krull, D.S. (1988). Seeing less and knowing more: the benefits of perceptual ignorance. *Journal of Personality and Social Psychology*, **54**, 193–202.

Gillis, J.S. & Bernieri, F.J. (2001). The perception and judgment of rapport. In *The essential Brunswik: beginnings, explications, applications* (ed. K.R. Hammond & T.R. Steward), pp. 380–83. New York: Oxford University Press.

Gillis, J., Bernieri, F., & Wooten, E. (1995). The effects of stimulus medium and feedback on the judgment of rapport. *Organizational Behavior and Human Decision Processes,* **63**, 33–45.

Grahe, J.E. & Bernieri, F.J. (2002). Self-awareness of judgment policies of rapport. *Personality and Social Psychology Bulletin,* **28**, 1407–18.

Green, D.M. & Swets, J.A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Guilford, J.P. (1954). *Psychometric methods*, 2nd edn. New York: McGraw–Hill.

Halberstadt, A.G. & Hall, J.A. (1980). Who's getting the message? Children's nonverbal skill and their evaluation by teachers. *Developmental Psychology,* **16**, 564–73.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 276 17.6.2005 1:59am

Hall, J.A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin,* **85**, 845–57.

Hall, J.A. (1984). *Nonverbal sex differences: communication accuracy and expressive style.* Baltimore: The Johns Hopkins University Press.

Hall, J.A. (1998). How big are nonverbal sex differences? The case of smiling and sensitivity to nonverbal cues. In *Sex differences and similarities in communication: critical essays and empirical investigations of sex and gender in interaction* (ed. D.J. Canary & K. Dindia), pp.155–77. Mahwah, NJ: Erlbaum.

Hall, J.A. (2001). The PONS test and the psychometric approach to measuring interpersonal sensitivity. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp.143–60. Mahwah, NJ: Erlbaum.

Hall, J.A. & Bernieri, F.J. (ed.) (2001). *Interpersonal sensitivity: theory and measurement.* Mahwah, NJ: Erlbaum.

Hall, J.A. & Braunwald, K.G. (1981). Gender cues in conversations. *Journal of Personality and Social Psychology,* **40**, 99–110.

Hall, J.A. & Carter, J.D. (1999). Gender-stereotype accuracy as an individual difference. *Journal of Personality and Social Psychology,* **77**, 350–9.

Hall, J.A. & Matsumoto, D. (2004). Sex differences in judgments of multiple emotions from facial expressions. *Emotion,* **4**, 201–6.

Hall, J.A. & Murphy, N.A. (2004). *Recall of nonverbal and verbal cues: exploring a new definition of nonverbal sensitivity.* Manuscript submitted for publication.

Hall, J.A., Carter, J.D., & Horgan, T.G. (2000). Gender differences in the nonverbal communication of emotion. In *Gender and emotion: social psychological perspectives* (ed. A.H. Fischer), pp. 97–117. Paris: Cambridge University Press.

Hall, J.A., Carter, J.D., & Horgan, T.G. (2001). Status roles and recall of nonverbal cues. *Journal of Nonverbal Behavior,* **25**, 79–100.

Hall, J.A., Coats, E.J., & Smith LeBeau, L. (2004). *Nonverbal behavior and the vertical dimension of social relations: a meta-analysis.* Manuscript submitted for publication.

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

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

Heider, F. (1958). *The psychology of interpersonal relations.* Hillsdale, NJ: Erlbaum.

Holder, H.B. & Kirkpatrick, S.W. (1991). Interpretation of emotion from facial expressions in children with and without learning disabilities. *Journal of Learning Disabilities,* **24**, 170–7.

Horgan, T.G., Schmid Mast, M., Hall, J.A., & Carter, J.D. (2004). Gender differences in memory for the appearance of others. *Personality and Social Psychology Bulletin,* **30**, 185–96.

Ickes, W. (ed.). (1997). *Empathic accuracy.* New York: Guilford.

Ickes, W. (2001). Measuring empathic accuracy. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 219–41. Mahwah, NJ: Erlbaum.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 277 17.6.2005 1:59am

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology,* **59**, 730–42.

Judd, C.M. & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review,* **100**, 109–28.

Kenny, D.A. (1994). *Interpersonal perception: a social relations analysis.* New York: Guilford.

Kenny, D.A. & Acitelli, L.K. (2001). Accuracy and bias in the perception of the partner in a close relationship. *Journal of Personality and Social Psychology*, **80**, 439–48.

Kenny, D.A. & Albright, L. (1987). Accuracy in interpersonal perception: a social relations analysis. *Psychological Bulletin,* **102**, 390–402.

Kenny, D.A. & Winquist, L. (2001). The measurement of interpersonal sensitivity: consideration of design, components, and unit of analysis. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 265–93. Mahwah, NJ: Erlbaum.

Kenny, D.A., Kashy, D.A., & Bolger, N. (1998). Data analysis in social psychology. In *The handbook of social psychology* (4th edn) (ed. D.T. Gilbert, S.T. Fiske, & G. Lindzey), pp. 233–65. Boston: McGraw–Hill.

Kenny, D.A., Kieffer, S.C., Smith, J.A., Ceplenski, P., & Kulo, J. (1996). Circumscribed accuracy among well-acquainted individuals. *Journal of Experimental Social Psychology*, **32**, 1–12.

Kline, P. (2000*a*). *A psychometrics primer.* London: Free Association Books.

Kline, P. (2000*b*). *A handbook of psychological testing.* London: Routledge.

Knapp, M.L. & Hall, J.A. (2002). *Nonverbal communication in human interaction* (5th edn). Belmont, CA: Wadsworth/Thomson Learning.

Koning, C. & Magill–Evans, J. (2001). Validation of the Child and Adolescent Social Perception Measure. *Occupational Therapy Journal of Research*, **21**, 49–67.

Kunda, Z. (1999). *Social cognition: making sense of people.* Cambridge, MA: MIT Press.

LaFrance, M. & Henley, N.M. (1994). On oppressing hypotheses: or differences in nonverbal sensitivity revisited. In *Power/gender: social relations in theory and practice* (ed. L. Radtke & H. Stam). London: Sage.

Landis, C. (1924). Studies of emotional reactions: 2. General behavior and facial expression. *Journal of Comparative Psychology*, **4**, 447–509.

Larkin, K.T., Martin, R.R., & McClain, S.E. (2002). Cynical hostility and the accuracy of decoding facial expressions of emotions. *Journal of Behavioral Medicine,* **25**, 285–92.

Lenti, C., Lenti–Boero, D., & Giacobbe, A. (1999). Decoding of emotional expressions in children and adolescents. *Perceptual and Motor Skills,* **89**, 808–14.

Lippa, R. (1998). The nonverbal display and judgment of extraversion, masculinity, femininity, and gender diagnosticity: a lens model analysis. *Journal of Research in Personality,* **32**, 80–107.

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.

Lopes, P.N., Salovey, P., & Straus, R. (2003). Emotional intelligence, personality, and the perceived quality of social relationships. *Personality and Individual Differences,* **35**, 641–58.

Loewenthal, K. (2001). *An introduction to psychological tests and scales* (2nd edn). Hove, UK: The Psychology Press.

Magill–Evans, J., Koning, C., Cameron–Sadava, A., & Manyk, K. (1995). The Child and Adolescent Social Perception Measure. *Journal of Nonverbal Behavior,* **19**, 151–69.

Malone, B.E. & DePaulo, B.M. (2001). Measuring sensitivity to deception. In Interpersonal sensitivity: Theory and measurement (ed. J.A. Hall & F. J. Bernieri), pp. 103–24. Mahwah, NJ: Erlbaum.

Marangoni, C., Garcia, S., Ickes, W., & Teng, G. (1995). Empathic accuracy in a clinically relevant setting. *Journal of Personality and Social Psychology,* **68**, 854–69.

Marcus, D.K. & Leatherwood, J.C. (1998). The interpersonal circle at zero acquaintance: a social relations analysis. *Journal of Research in Personality,* **32**, 297–313.

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

Mayer, J.D. & Salovey, P. (1997). What is emotional intelligence? In *Emotional development and emotional intelligence: implications for educators* (ed. P. Salovey & D. Sluyter), pp. 3–31. New York: Basic Books.

Mayer, J.D., Salovey, P., Caruso, D.R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion,* **3**, 97–105.

McAndrew, F.T. (1986). A cross-cultural study of recognition thresholds for facial expression of emotion. *Journal of Cross-Cultural Psychology,* **17**, 211–24.

McArthur, L.Z. & Baron, R.M. (1983). Toward an ecological theory of social perception. *Psychological Review*, **90**, 215–38.

McClure, E.B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological Bulletin,* **126**, 424–53.

McLarney–Vesotski, A. (2003). *Trait predictors of the good judge: three proposed underlying mediators.* Unpublished doctoral dissertation, University of Toledo, Toledo, OH.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review,* **92**, 350–71.

Meltzoff, A.N. & Moore, M.K. (1977). Imitation of facial and manual gestures by human neonates. *Science,* **198**, 75–8.

Murphy, N.A. (2003). *Intelligence in social interaction.* Unpublished doctoral dissertation, Northeastern University, Boston, MA.

Murphy, N.A., Hall, J.A., & Colvin, C.R. (2003). Accurate intelligence assessments in social interaction: mediators and gender effects. *Journal of Personality,* **71**, 465–93.

Murphy, S.T. & Zajonc, R.B. (1993). Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, **64,** 723–39.

Niit, T. & Valsiner, J. (1977). Recognition of facial expressions: an experimental investigation of Ekman's model. *Acta et Commentationes Universitatis Tarvensis,* **429**, 85–107.

Noller, P. (1980). Misunderstandings in marital communication: a study of couples' nonverbal communication. *Journal of Personality and Social Psychology*, **39**, 1135–48.

Noller, P. (2001). Using standard content methodology to assess nonverbal sensitivity in dyads. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 243–64. Mahwah, NJ: Erlbaum.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 279 17.6.2005 1:59am

Nowicki, S., Jr. & Duke, M.P. (1994). Individual differences in the nonverbal communication of affect: the Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior*, **18**, 9–35.

Nowicki, S., Jr. & Duke, M.P. (2001). Nonverbal receptivity: the Diagnostic Analysis of Nonverbal Accuracy (DANVA). In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 183–98). Mahwah, NJ:Erlbaum.

O'Connor, R.M. & Little, I.S. (2003). Revisiting the predictive validity of emotional intelligence: self-report versus ability-based measures. *Personality and Individual Differences,* **35**, 1893–902.

Patterson, M.L. (1995). A parallel process model of nonverbal communication. *Journal of Nonverbal Behavior*, **19**, 3–29.

Phillips, R.D., Wagner, S.H., Fells, C.A., & Lynch, M. (1990). Do infants recognize emotion in facial expressions? Categorical and 'metaphorical' evidence. *Infant Behavior and Development,* **13**, 71–84.

Reynolds, D.J. & Gifford, R. (2001). The sounds and sights of intelligence: a lens model channel analysis. *Personality and Social Psychology Bulletin,* **27**, 187–200.

Riggio, R. E., & Riggio, H. R. (2001). Self-report measurement of interpersonal sensitivity. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 127–42. Mahwah, NJ: Erlbaum.

Roberts, R.D., Zeidner, M., & Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? Some new data and conclusions. *Emotion,* **1**, 196–231.

Rojahn, J., Singh, N.N., Singh, S.D., Baker, J.A., Lawrence, M.A., & Davis, C.M. (2002). Concurrent validity studies of the Facial Discrimination Task. *Journal of Child and Family Studies,* **11**, 203–15.

Rollman, S.A. (1978). The sensitivity of Black and White Americans to nonverbal cues of prejudice. *Journal of Social Psychology,* **105,** 73–7.

Rosenthal, R. (1976). *Experimenter effects in behavioral research* (enlarged edn). New York: Irvington.

Rosenthal, R. & Rubin, D.B. (1978). Interpersonal expectancy effects: the first 345 studies. *Behavioral and Brain Sciences*, **3**, 377–86.

Rosenthal, R. & Rubin, D.B. (1989). Effect size estimation for one-sample multiple-choice-type data: design, analysis, and meta-analysis. *Psychological Bulletin,* **106**, 332–7.

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.

Rosip, J.C. & Hall, J.A. (2004). Knowledge of nonverbal cues, gender, and nonverbal decoding accuracy. *Journal of Nonverbal Behavior,* **28**, 267–86.

Russell, J.A. (1994). Is there universal recognition of emotion from facial expressions? A review of cross-cultural studies. *Psychological Bulletin,* **115**, 102–41.

Sapir, E.A. (1949). Communication. In *Selected writings of Edward Sapir in language, culture, and personality* (ed. D.G. Mandelbaum). Berkeley and Los Angeles: University of California Press.

Scherer, K.R. (1978). Inference rules in personality attribution from voice quality: the loud voice of extraversion. *European Journal of Social Psychology,* **8**, 467–87.

Scherer, K.R. (1982). Methods of research on vocal communication: paradigms and parameters. In *Handbook of methods in nonverbal behavior research* (ed. K.R. Scherer & P. Ekman), pp. 136–98. Cambridge, UK: Cambridge University Press.

Scherer, K.R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication,* **40**, 227–56.

Scherer, K.R., Feldstein, S., Bond, R.N., & Rosenthal, R. (1985). Vocal cues to deception: a comparative channel approach. *Journal of Psycholinguistic Research,* **14**, 409–25.

Scherer, K.R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In *Handbook of the affective sciences* (ed. R.J. Davidson, H. Goldsmith, & K.R. Scherer), pp. 433–56. New York: Oxford University Press.

Scherer, K.R., Scherer, U., Hall, J.A., & Rosenthal, R. (1977). Differential attribution of personality based on multi-channel presentation of verbal and nonverbal cues. *Psychological Research,* **39**, 221–47.

Schmid Mast, M., & Hall, J.A. (2004). Who is the boss and who is not? Accuracy of judging status. *Journal of Nonverbal Behavior,* **28**, 145–66.

Smith, H.J., Archer, D., & Costanzo, M. (1991). 'Just a hunch': accuracy and awareness in person perception. *Journal of Nonverbal Behavior,* **15**, 3–18.

Snodgrass, S.E. (1985). Women's intuition: the effect of subordinate role on interpersonal sensitivity. *Journal of Personality and Social Psychology,* **49**, 146–55.

Snodgrass, S.E. (1992). Further effects of role versus gender on interpersonal sensitivity. *Journal of Personality and Social Psychology,* **62**, 154–8.

Snodgrass, S.E. (2001). Correlational method for assessing interpersonal sensitivity within dyadic interaction. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 201–18. Mahwah, NJ: Erlbaum.

Snodgrass, S.E., Hecht, M.A., & Ploutz–Snyder, R. (1998). Interpersonal sensitivity: expressivity or perceptivity? *Journal of Personality and Social Psychology,* **74**, 238–49.

Snyder, M., Tanke, E.D., & Berscheid, E. (1977). Social perception and interpersonal behavior: on the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology,* **35**, 656–66.

Spitz, H.H. (1997). *Nonconscious movements: from mystical messages to facilitated communication.* Mahwah, NJ: Erlbaum.

Stinson, L. & Ickes, W. (1992). Empathic accuracy in the interaction of male friends versus male strangers. *Journal of Personality and Social Psychology,* **62**, 787–97.

Swim, J.K. (1994). Perceived versus meta-analytic effect sizes: an assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology,* **66**, 21–36.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin,* **52**, 1–23.

Thomas, G. & Fletcher, G.O. (2003). Mind-reading accuracy in intimate relationships: assessing the roles of the relationship, the target, and the judge. *Journal of Personality and Social Psychology,* **85**, 1079–94.

Vogt, D.S. & Colvin, C.R. (2003). Interpersonal orientation and the accuracy of personality judgments. *Journal of Personality,* **71**, 267–95.

Vrij, A. & Semin, G.R. (1996). Lie experts' beliefs about nonverbal indicators of deception. *Journal of Nonverbal Behavior,* **20,** 65–80.

Wagner, H.L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior,* **17**, 3–28.

Harrigan / The new handbook of methods in nonverbal behaviour research 06-harrigan-chap06 Page Proof page 281 17.6.2005 1:59am

Wagner, H.L. (1997). Methods for the study of facial behavior. In *The psychology of facial expression* (ed. J.A. Russell & J.M. Fernández–Dols), pp. 31–54. New York: Cambridge University Press.

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**, 207–19.

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**, 120–8.

Zebrowitz, L.A. (2001). Groping for the elephant of interpersonal sensitivity. In *Interpersonal sensitivity: theory and measurement* (ed. J.A. Hall & F.J. Bernieri), pp. 333–50. Mahwah, NJ: Erlbaum.

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**, 238–49.

Zuckerman, M. & Larrance, D.T. (1979). Individual differences in perceived encoding and decoding abilities. In *Skill in nonverbal communication* (ed. R. Rosenthal), pp. 171–203. Cambridge, MA: Oelgeschlager, Gunn & Hain.

Zuckerman, M., Lipets, M.S., Koivumaki, J.H., & Rosenthal, R. (1975). Encoding and decoding nonverbal cues of emotion. *Journal of Personality and Social Psychology,* **32**, 1068–76.

Author Queries

[q1] There is Lippa 1998 in the reference list, and Lippa & Dietz 2000. Should this be either of these? If not, need to include details in the reference list.
[q2] Details of this Colvin et al. 1995 reference are missing from the reference list. Please add.
[q3] Likewise, details of Scherer & Ceschi 2000 are missing from reference list.
[q4] All these abbreviations should be spelt out in full here, at their first mention.
[q5] Details of Funder 1980 are missing from the reference list.
[q6] Details of Kenny 2001 are missing from the reference list.
[q7] What does CARAT stand for?
[q8] Presumably details for a Ekman & Friesen 1978 reference should be included in the reference list.
[q9] Cannot find Ambady & Rosenthal 1993 cited in text. Please locate.
[q10] Likewise this Brackett & Mayer 2003 reference.
[q11] How can this Carney et al. reference have a publication date of 2004 but also be described as submitted (i.e. awaiting) publication? Which is correct?
[q12] Likewise this Hall et al. 2004 reference.
[q13] Likewise this Hall & Murphy 2004 reference.
[q14] Cannot find Lopes et al. 2003 cited in text. Please locate.
[q15] Likewise this O'Connor & Little reference.
[q16] Likewise this Snodgrass 2001 reference.