

## Gender differences in empathic accuracy: Differential ability or differential motivation?

WILLIAM ICKES,<sup>a</sup> PAUL R. GESN,<sup>b</sup> AND TIFFANY GRAHAM<sup>c</sup>

<sup>a</sup>University of Texas at Arlington, <sup>b</sup>Texas Education Agency, and <sup>c</sup>J.C. Penney Corp.

### Abstract

Following their qualitative review of the findings from 10 relevant studies, Graham and Ickes (1997) speculated that reliable gender-of-perceiver differences in empathic accuracy (a) were limited to studies in which the empathic inference form made empathic accuracy salient as the dimension of interest, and (b) therefore reflected the differential motivation, rather than the differential ability, of female versus male perceivers. These speculations were tested more rigorously in the present study, which examined a larger set of 15 empathic accuracy studies and applied the techniques of quantitative meta-analysis to test Graham and Ickes' (1997) moderating variable hypothesis. The hypothesis was strongly supported, consistent with a motivational interpretation previously proposed by Berman (1980) and by Eisenberg and Lennon (1983), which argues that reliable gender differences in empathy-related measures are found only in situations in which (a) subjects are aware that they are being evaluated on an empathy-relevant dimension, and/or (b) empathy-relevant gender-role expectations or obligations are made salient.

According to the cultural stereotype of "women's intuition," women are assumed to display greater interpersonal sensitivity than men—at least at the level of group differences. Evidence for this stereotype within American culture can be found in a number of studies cited by Manstead (1992), in which women were specifically rated as being more empathic than men. For example, when responding to items on the "Beliefs about Women Scale" (Belk & Snell, 1986), both male and female respondents agreed with the generalizations that women have more emotional insight than do men. In addition, research by Broverman, Vogel, Broverman, Clarkson, and

Rosenkrantz (1994) revealed a general belief that the ideal woman is more "sensitive to the feelings of others," whereas the ideal man is perceived as "less aware of the feelings of others."

Such beliefs imply that the assumed gender difference in interpersonal sensitivity is primarily one of *differential ability*. In other words, at the level of group differences, women are assumed to possess greater empathic ability than men. Interestingly, however, when Graham and Ickes (1997) conducted a qualitative review of the gender-of-perceiver differences reported in 10 studies of empathic accuracy, they concluded that reliable differences favoring female perceivers are the exception—rather than the rule—in empathic accuracy research, occurring in only 3 of the 10 available studies. Even more important, Graham and Ickes concluded that this gender-of-perceiver difference, when it does occur, is primarily one of *differential motivation* rather than differential ability.

The authors would like to thank Greg Pool and three anonymous reviewers for their comments on a previous draft of this article.

Requests for reprints should be sent to William Ickes, Department of Psychology, University of Texas at Arlington, Arlington, TX 76019-0528. E-mail: ickes@uta.edu.

To understand these somewhat surprising conclusions, it is necessary to review the evidence on which they were based. Before doing so, however, it will be useful to examine the conceptual meaning of empathic accuracy and to consider how this construct was measured in the studies included in the qualitative synthesis reported by Graham and Ickes (1997).

### **Empathic Accuracy**

*Empathic inference* is the "everyday mind reading" that people do whenever they attempt to infer other people's thoughts and feelings. *Empathic accuracy* is the extent to which such "mind reading" attempts are successful (Ickes, 1997). According to Goleman (1995), the ability to accurately "read" other people's thoughts and feelings is an important skill that affects people's social adjustment in all phases of their life: as students in the classroom, as playmates and platonic friends, as dating and marriage partners, as parents, as members of the work force, and as members of the larger community. Indeed, this ability may be the quintessential aspect of what is commonly termed "social intelligence." All else being equal, it is this ability that distinguishes "the most tactful advisors, the most diplomatic officials, the most effective negotiators, the most electable politicians, the most productive salespersons, the most successful teachers, and the most insightful therapists" (Ickes, 1997, p. 2).

#### *Measuring empathic accuracy*

A general procedure for measuring empathic accuracy was developed by the first author and his colleagues (Ickes, Bissonnette, Garcia, & Stinson, 1990a; Ickes, Stinson, Bissonnette, & Garcia, 1990b). There are two major research paradigms in which this procedure has been applied. The first is the *unstructured dyadic interaction paradigm*, in which dyad members attempt to infer each other's thoughts and feelings from a videotaped record of their spontaneous interaction during a brief period in

which the experimenter has left them alone together (Ickes & Tooke, 1988; Ickes et al., 1990a; Stinson & Ickes, 1992). The second is the *standard stimulus paradigm*, in which individual participants each view the same standard set of videotaped interactions and attempt to infer the thoughts and feelings of the same set of "target" persons (Gesn & Ickes, 1999; Marangoni, Garcia, Ickes, & Teng, 1995).

*The unstructured dyadic interaction paradigm.* In the first research paradigm, pairs of subjects are unobtrusively audiotaped and videotaped during an initial, unstructured interaction (Ickes et al., 1990a). At the end of a brief (6-to-10-minute) observation period, the dyad members are seated in separate cubicles and asked to view a videotape of the interaction that has just occurred and to provide data about the thoughts and feelings they experienced during the interaction. The subjects are each given a start/pause control to use while viewing their respective copies of the videotape, along with a supply of standardized thought/feeling coding forms. At each point a subject remembers having had a specific thought or feeling during the interaction, the subject's task is to pause the tape and write down (a) the exact time the thought or feeling occurred (available from a running timer that is superimposed on the video image at the top of the screen), and (b) the specific content of that thought or feeling (expressed in the form of a sentence beginning with one of two sentence stems: "I was thinking" or "I was feeling").

After making a complete log of their own thoughts and feelings during the interaction, the subjects are asked to view the videotape a second time. This time the tape is stopped for them at each of those points at which the subject's *partner* reported having had a thought or feeling. It is now the task of each subject to infer the specific content of each of his or her partner's thoughts and feelings. Following the collection of all of the thought/feeling data, independent raters are asked to judge the similarity between each of the actual thoughts

and feelings reported by one partner and the corresponding thought/feeling inferences reported by the other partner. These similarity ratings are then aggregated to create a measure of empathic accuracy that is scaled—like a percentage measure—to range from 0 (no accuracy) to 100 (perfect accuracy). For a more detailed description of these procedures, see Ickes et al. (1990a, 1990b).

*The standard stimulus paradigm.* In the second research paradigm, individual participants are asked to view the same standard set of videotaped interactions and attempt to infer the thoughts and feelings of the same “target” persons. The prototype for these studies was the study by Marangoni et al. (1995) of empathic accuracy in a clinically relevant setting. In this study, individual participants were asked to view three videotaped interactions. Each interaction depicted a female client discussing a genuine personal problem with a client-centered male therapist.

Each participant independently viewed all three stimulus tapes, and in each case attempted to infer the content of the thought or feeling the client had reported at each of 30 tape stops. Using a written log of the times at which each client’s tape stops had occurred, either the experimenter or a research assistant paused the tape at each stop. The participant then wrote down her or his thought/feeling inference on the empathic inference form, and restarted the tape by means of the remote control. Empathic accuracy scores were later derived in the manner described above. Highly edited versions of the stimulus tapes developed by Marangoni et al. (1995) were later used in a study by Gesn and Ickes (1999).

In the most recent application of the standard stimulus paradigm, Kelleher (1998) used a set of four standard stimulus tapes to study perceivers’ ability to infer the thoughts and feelings of individuals who were trying to influence their partner’s behavior by carrying out an assigned “hidden agenda” during their initial, mixed-sex dy-

adic interaction. For a more complete description of this study, see Kelleher (1998).

*Reliability and validity of the empathic accuracy measure.* Recall that independent raters are asked to judge the similarity between each of the actual thoughts and feelings reported by one partner and the corresponding thought/feeling inferences made by the other partner. Interrater reliability in the empathic accuracy studies conducted at the University of Texas at Arlington has consistently been quite high, ranging from a low of .85 in a study in which only four raters were used to a high of .98 in two studies in which either seven or eight raters were used. Across all of the studies we have conducted to date, the average interrater reliability has been about .90.

The predictive validity of the empathic accuracy measure has been established in several studies. One of our first predictions was that if our procedure for assessing empathic accuracy was indeed valid, close friends should display higher levels of accuracy than should strangers when inferring the content of each other’s thoughts and feelings. This prediction was confirmed in studies by Stinson and Ickes (1992) and Graham (1994). In the clinically relevant study conducted by Marangoni et al. (1995), the predictive validity of our empathic accuracy measure was further tested with respect to the hypotheses that (a) perceivers’ empathic scores should be significantly greater at the end of the psychotherapy tapes than at the beginning, reflecting their greater acquaintance with the clients and their problems, and (b) perceivers who receive immediate feedback about clients’ actual thoughts and feelings during the middle portion of each tape should subsequently achieve better empathic accuracy scores than perceivers who do not receive such feedback. Statistically significant support for both of these hypotheses was obtained. Further evidence for the predictive validity of the empathic accuracy measure is available in studies by Simpson, Ickes, and Blackstone (1995), Kelleher (1998), and Gesn and Ickes (1999).

### Graham and Ickes' (1997) Qualitative Review

Given this background on the empathic accuracy construct and how it is measured, we are now in a better position to evaluate the somewhat unexpected conclusions reached by Graham and Ickes (1997). Following their qualitative review of the gender-of-perceiver differences reported in 10 studies of empathic accuracy, Graham and Ickes (1997) reached two conclusions that were somewhat counterintuitive with respect to the widely held belief that, at the level of group differences, women are better able than men to accurately infer the content of other people's thoughts and feelings. First, they concluded that reliable differences favoring female perceivers are the exception—rather than the rule—in empathic accuracy research. Second, they concluded that this gender-of-perceiver difference, when it does occur, is primarily one of differential motivation rather than differential ability.

How did Graham and Ickes (1997) arrive at these conclusions? They were attempting to understand why, after seeing no evidence of a reliable gender-of-perceiver difference in empathic accuracy in the first seven studies using the procedure developed by Ickes and his colleagues (Ickes et al., 1990a, 1990b), reliable gender differences favoring female over male perceivers were evident in the next three studies using this procedure. As Graham and Ickes (1997, pp. 128–132) noted:

*Results of the first seven studies.* . . . Taken together, the results of [the] first seven studies revealed no evidence that the average levels of empathic accuracy are reliably greater for female than for male perceivers. In virtually all of these studies, the relevant *F* statistic for the gender-of-perceiver difference had a value of less than one. Furthermore, the only exceptions—the nonsignificant trends in Replication 2 of Hancock and Ickes' (1996) study of same-sex groups and in Thomas and Fletcher's (1996) study of married couples—were split in terms of favoring female versus male perceivers.

These null results were found in studies of both mixed-sex and same-sex dyads: in studies of strangers, dating partners, and married couples; and in studies conducted in Texas, North Carolina, and New Zealand. Moreover, these null results cannot readily be attributed to either floor or ceiling effects, because the mean empathic accuracy scores tended to vary within the general range of .15 to .30 on an index having a theoretical range of .00 to 1.00, and an average standard deviation of about .11.

*Results of the next three studies.* In sharp contrast to the results of the first seven studies, the next three studies (Gesn, 1995; Ickes, Hancock, Gesn, Graham, & Mortimer, 1995; Graham, 1996) all yielded significant gender differences indicating greater empathic accuracy for female than for male perceivers. Although these three studies were also somewhat diverse in their designs and procedures (the first was a dyadic interaction study; the other two used standard tapes from the Marangoni et al., 1995, study), all three of them were conducted in our lab after September of 1994, and they all used a new version of the empathic inference form that had not been used in any of the seven previous studies. With regard to the studies conducted in our lab at UTA, the new form differed from the old one in one—and only one—respect [see Figure 1]. In the last column of this form, subjects were now required to estimate their accuracy in inferring each and every one of the target person's thoughts and feelings, instead of inferring whether the general emotional tone of each specific thought or feeling was positive (+), neutral (0), or negative (-).

Why should changing the empathic inference form in this way change the results so dramatically, in effect creating significant gender differences that were never evident before? To answer this question, Graham and Ickes (1997) turned to an important theoretical precedent established by Eisenberg and Lennon (1983; Lennon & Eisenberg, 1987). Eisenberg and Lennon conducted broad-scope reviews of gender differences in several empathy-relevant measures obtained in a large number of studies that all predated the first published study using the empathic accuracy measure (Ickes et al., 1990b). The empathy-relevant

measures reviewed by Eisenberg and Lennon (1983, pp. 102–103) included:

(a) infants' crying in response to another's distress; (b) individual reports to the experimenter of emotional responsiveness after [exposure] to stories or . . . pictures containing information regarding a hypothetical other's affective state; (c) self-report (via pencil-and-paper measures) of emotional responsiveness in simulated distress situations; (d) observers' ratings of individuals' facial, gestural, and/or vocal reactions to another's emotional state; (e) subjects' physiological responses to another's predicament; (f) individuals' responses on self-report scales specifically designed to measure empathy; and (g) report by others of individuals' empathy.

In both their original 1983 review and in their updated 1987 review, Eisenberg and Lennon concluded that the data revealed a pattern of gender differences that was highly inconsistent across the set of measures they examined. They further concluded that this pattern could not be satisfactorily explained either in terms of the age of the subjects or in terms of the argument of Buck et al. (Buck, 1981; Buck, Savin, Miller, & Caul, 1972; Buck, Miller, & Caul, 1974) that women tend to be externalizers of their emotional responses whereas men tend to be internalizers. Instead, Eisenberg and Lennon concluded—and provided compelling evidence to support their claim—that this pattern could be satisfactorily explained in terms of the hypothesis that

[T]he inconsistent pattern of results is due to differences in the various methods. The sex difference in empathy is most evident when it is obvious what behavior or trait is being assessed. Thus, when individuals have been asked to rate themselves on behaviors or reactions clearly related to the concept of empathy, females have scored much higher than males. When the demand characteristics inherent in a methodology have been a bit more subtle, that is, when the purpose of the assessment situation was not clear, females have still tended to score higher on empathy, but the sex difference is much smaller. Finally, when the measure of empathy has been even more unobtrusive, for example,

when physiological measures or facial/gestural measures have been the indices of empathy, females have not exhibited more empathy than males. (Eisenberg & Lennon, 1983, p. 124)

Eisenberg and Lennon (1983, p. 125) further noted that there is direct empirical support for the idea “that self-report of empathy may be influenced by demand characteristics,” citing the results of a study by Wispe, Kiecolt, and Long (1977), who “found a strong relation between the outcome for the story protagonist in the film and self-report of affect, but only when there were clear demand characteristics in the experimental situation.” Finally, they noted that Berman (1980) reached a very similar conclusion to their own in her review of the literature on sex differences in responsiveness to the young. “In brief, Berman's findings can be interpreted as indicating that . . . [such differences are found] only in situations in which (a) it is clear that subjects are being evaluated on that dimension, or (b) subjects are in circumstances in which role expectations or obligations are salient” (Eisenberg & Lennon, 1983, p. 125, brackets ours).

Following the theoretical precedents established by Berman (1980), Eisenberg and Lennon (1983) and Lennon and Eisenberg (1987), Graham and Ickes argued that the same interpretation could be applied to the data from the 10 empathic accuracy studies that were included in their 1997 qualitative review. Specifically, they argued that the empathic inference form used in the first seven studies did not make salient to subjects the fact that their empathic accuracy was being assessed. Only in the last three studies, when the new empathic inference form required the perceivers to think about and rate how empathically accurate they were with respect to each and every thought/feeling inference they made, was there a salient demand characteristic of the type emphasized by Berman (1980) and by Eisenberg and Lennon (1983). Graham and Ickes (1997) therefore endorsed Eisenberg and Lennon's (1983) conclusion that gender differences in empathic accuracy are

DATE \_\_\_\_\_

NUMBER \_\_\_\_\_

M F

TIME	THOUGHT OR FEELING	+, 0, -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	+ 0 -	

Figure 1. Old (this page) and new (facing page) empathic inference forms

DATE \_\_\_\_\_

NUMBER \_\_\_\_\_

M F

TIME	THOUGHT OR FEELING	How accurate do you think you were?
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very
	<input type="checkbox"/> He/she was thinking: <input type="checkbox"/> He/she was feeling:	0 - not at all 1 - slightly 2 - moderately 3 - very

Figure 1. (continued)

evident only in circumstances that (a) make it clear that subjects are being evaluated on that dimension, or (b) make salient the gender-role expectation that females are supposed to display greater empathy than males.

According to this interpretation, our new empathic inference form evoked real gender differences that we had not observed before, but differences that should be attributed more to differential motivation than to differential ability. In other words, although the average difference in the empathic ability of women versus men may be slight, or limited more specifically to women's greater ability to decode emotions from visual cues (Eisenberg & Lennon, 1983, pp. 119–123), female perceivers may (a) want to appear more empathic than men and/or (b) feel a greater obligation to do so. If female perceivers have been socialized to believe that women are, and should be, more empathic than men (see Cross & Madson, 1997), then the deceptively simple change we made in our response form may have been all that was needed to engage the female perceivers' greater motivation to appear highly empathic. And this enhanced motivation, rather than greater empathic ability per se, might account for the significant gender differences that were evident in the last three empathic accuracy studies reviewed by Graham and Ickes (1997) but not in the seven previous ones.

In the present study, we sought to provide a more rigorous test of Graham and Ickes' (1997) hypothesis that the magnitude of gender differences in empathic accuracy was moderated by the type of empathic inference form used in the studies reviewed by Graham and Ickes (1997). This study provided a more rigorous test in two respects. First, we analyzed the results of 15 empathic accuracy studies, in contrast to the smaller set of 10 studies included in Graham and Ickes' (1997) review. Second, we subjected these data to a quantitative meta-analysis that enabled us to assess the statistical significance of the hypothesized moderating variable, a feature that was

lacking in Graham and Ickes' previous qualitative review.

If Graham and Ickes' (1997) conclusions are upheld by the results of this more rigorous test, there will be a stronger evidentiary basis for arguing that differential motivation, rather than differential ability, contributes more to the advantage of female over male perceivers in accurately "reading" other people's thoughts and feelings. In addition, there will be further evidence for the generality of Berman's (1980) and Eisenberg and Lennon's (1983) argument that reliable gender differences in empathy-related measures are found only in situations in which (a) subjects are aware that they are being evaluated on an empathy-relevant dimension, and/or (b) empathy-relevant gender-role expectations or obligations are made salient.

## Method

Results of Graham and Ickes' (1997) qualitative review suggested that the change we made in our empathic inference coding form might account for the presence or absence of significant gender-of-perceiver differences in the set of 10 available empathic accuracy studies in which such differences could be tested. The results of our current quantitative meta-analysis include the 10 studies considered in Graham and Ickes' (1997) review, plus five more recent studies for which gender-of-perceiver differences could also be tested. Our goal was to determine whether Graham and Ickes' (1997) moderating variable hypothesis would be supported by the results of a quantitative meta-analysis that included the data for all 15 studies.

### *The 15 studies*

The 15 studies we examined all used Ickes, et al.'s (1990b) procedure for assessing empathic accuracy. Twelve of the studies were conducted in the Social Interaction Lab at the University of Texas at Arlington. Of the other three studies, one was conducted at Texas A & M University (Simpson et al.,

1995), another at the University of North Carolina (Bissonnette et al., 1997), and the last at the University of Canterbury in Christchurch, New Zealand (Thomas, Fletcher, & Lange, 1997). A search of PsychInfo and other electronic databases did not reveal any additional studies that were relevant to the present meta-analytic review.

*Standard stimulus paradigm.* Eight of the 15 studies employed the *standard stimulus paradigm* (see Ickes, in press), in which individual participants each view the same standard set of videotaped interactions and attempt to infer the thoughts and feelings of the same set of "target" persons (Gesn & Ickes, 1999; Graham, 1996; Ickes et al., 1995; Ickes et al., 1996; Kelleher, 1998; Marangoni et al., 1995; Mortimer, 1996; Renshaw, 1998). Of these eight studies, seven used either two or all three of the stimulus tapes of simulated psychotherapy sessions that were originally developed for use in the study by Marangoni et al. (1995). (Highly edited versions of these three tapes were used in the study by Gesn and Ickes, 1999.) The only exception was the study by Kelleher (1998), in which the perceivers inferred the thoughts and feelings of the male and female members of four mixed-sex dyads whose initial interactions were unobtrusively recorded and used as the standard stimulus tapes.

The perceivers in all eight standard stimulus paradigm studies were college students who were currently enrolled in introductory psychology at the University of Texas at Arlington. In all but the Kelleher (1998) study, the target persons whose thoughts and feelings they attempted to infer were three Caucasian women in their mid-20s who discussed their actual relationship problems with the same male, Rogerian-trained therapist. The edited tapes of the psychotherapy sessions used in these studies ranged from 26 to 36 minutes in length. In the Kelleher (1998) study, the target persons were eight students (four male, four female) at the University of Texas at Arlington whose initial interactions in op-

posite-sex dyads were unobtrusively videotaped during their participation in a study for credit in their introductory psychology course. The tapes of each of the four interactions in the Kelleher (1998) study were each shown through the first six minutes of the dyad members' interaction.

*Dyadic interaction paradigm.* Seven of the 15 studies employed the *unstructured dyadic interaction paradigm*, in which dyad members attempt to infer each other's thoughts and feelings from a videotaped record of their spontaneous interaction during a brief period in which they were left alone together (Bissonnette et al., 1997; Thomas et al., 1997; Gesn, 1995; Graham, 1994; Hancock & Ickes, 1996; Ickes et al., 1990b; Simpson et al., 1995). The dyad members were married couples in the studies by Bissonnette et al. (1997) and Thomas et al. (1997); dating couples in the study by Simpson et al. (1995); a mix of same-sex friends and same-sex strangers in the studies by Gesn (1995), Graham (1994), and Hancock and Ickes (1996); and opposite-sex strangers in the study by Ickes et al. (1990b). Because the participants in the dyadic interactions recorded in these studies served as both perceivers and as targets, the perceiver characteristics and the target characteristics were essentially the same.

#### *Measure of effect size*

Pearson correlation coefficients for the gender-of-perceiver effect were computed for each study as a measure of effect size using the method suggested by Rosenthal (1991) for calculating Pearson  $r$  directly from  $t$  values. These  $r$  values were then used to calculate Cohen's effect size  $d$  for each study. Effect size  $d$  can be interpreted as a standardized mean difference that indexes the magnitude of a given effect (Cohen, 1988).

## **Results**

The values of  $r$  and  $d$  for each study, along with their associated  $p$ -values, are pre-

sented in Table 1. Consistent with the Graham and Ickes (1997) hypothesis, the studies in Table 1 are grouped *not* according to the type of paradigm used (standard stimulus paradigm vs. dyadic interaction paradigm) but rather according to the type of empathic inference coding form used (self-ratings of accuracy required vs. no self-ratings of accuracy required).

The results reveal that the effect sizes for the gender-of-perceiver differences in the first set of studies (those in which the participants did not rate the accuracy of their empathic inferences) were consistently low and were associated with correlations that were all nonsignificant. In contrast, the effect sizes of the gender-of-perceiver differences in the second set of studies (those in which the participants did rate the accuracy of their empathic inferences) were consistently higher and were associated with correlations that were significant in every case but one. The single exception was the study by Thomas et al. (1997), which Graham and

Ickes (1997) had originally categorized as one in which no self-ratings of accuracy were made. This categorization was changed in the present study, however, following the recommendation of Geoff Thomas (the first author of Thomas et al., 1997), who argued that the confidence ratings the perceivers made following each of their empathic inferences were similar enough to our accuracy ratings to justify moving the Thomas et al. (1997) study from the first subset of studies to the second. In the interest of ensuring a conservative test of our moderating variable hypothesis, we were happy to accept this recommendation.

Mean effect sizes and the  $p$ -values for their associated  $t$  or  $F$  tests are provided in Table 2. The first thing to note in Table 2 is that across all 15 studies there is a statistically significant gender-of-perceiver difference in our performance measure of empathic accuracy (i.e., women obtained higher empathic accuracy scores than did men,  $p < .001$ ). The effect size, as indexed by

**Table 1.** Significance tests and gender-of-perceiver effect sizes

Study	$t$ ( $df$ )	$p$ value	Pearson $r$	Cohen's $d$
<i>No Self-Ratings of Accuracy</i>				
Ickes, Stinson, Bissonnette, & Garcia (1990b)	1.05 (37)	.15	.17	.35
Graham (1994)	-.66 (46)	.26	-.10	-.19
Marangoni, Garcia, Ickes, & Teng (1995)	-.14 (78)	.45	-.02	-.03
Simpson, Ickes, & Blackstone (1995)	-.24 (78)	.41	-.03	-.05
Hancock & Ickes (1996)	-.56 (28)	.29	-.11	-.21
Bissonnette, Rusbult, & Kilpatrick (1997)	.38 (23)	.35	.08	.16
Kelleher (1998)	.79 (41)	.43	.12	.24
Renshaw (1998)	.97 (71)	.33	.11	.22
Gesn & Ickes (1999)	.87 (71)	.20	.10	.21
<i>Self-Ratings of Accuracy</i>				
Gesn (1995)	2.83 (62)	.003	.34	.72
Ickes, Hancock, Gesn, Graham, & Mortimer (1995)	2.68 (126)	.004	.23	.48
Graham (1996)	4.31 (94)	.001	.41	.89
Ickes, Gesn, Dugosh, Pham, & Renshaw (1996)	2.22 (49)	.02	.30	.63
Mortimer (1996)	1.66 (70)	.05	.19	.40
Thomas, Fletcher, & Lange (1997)	1.21 (73)	.12	.14	.28

\*All  $p$  values are one-tailed, consistent with the hypothesis based on the "women's intuition" stereotype.

**Table 2.** Combined effect sizes and significance levels

	<i>p</i> value*	Pearson <i>r</i>	Cohen's <i>d</i>
Studies with no self-ratings of accuracy	.41	.02	.04
Studies with self-ratings of accuracy	< .001	.27	.56
All studies	< .001	.13	.26

\*All *p* values are one-tailed.

Cohen's *d*, is a moderate .26, with a 95% confidence interval (CI) of .10 to .43.

Of much greater importance, however, is the evidence in Table 2 that the change in our empathic inference coding form reliably moderated the presence or absence of significant gender-of-perceiver differences in the 15 studies we examined. The mean effect size *d* for the nine studies in which self-ratings of accuracy were not obtained (top of Table 1) was small (.04), with a 95% CI of  $-.03$  to .10. The combined probability level for these nine studies was not significant (.41). In contrast, the mean effect size for the six studies in which self-ratings of accuracy were obtained (bottom of Table 1) was large (.56), with a 95% CI of .38 to .74. The combined probability level for these six studies was  $p < .0001$ . Moreover, the difference between the two mean effect sizes (.10 and .62) was clearly significant,  $z = 3.85, p < .0001$ .

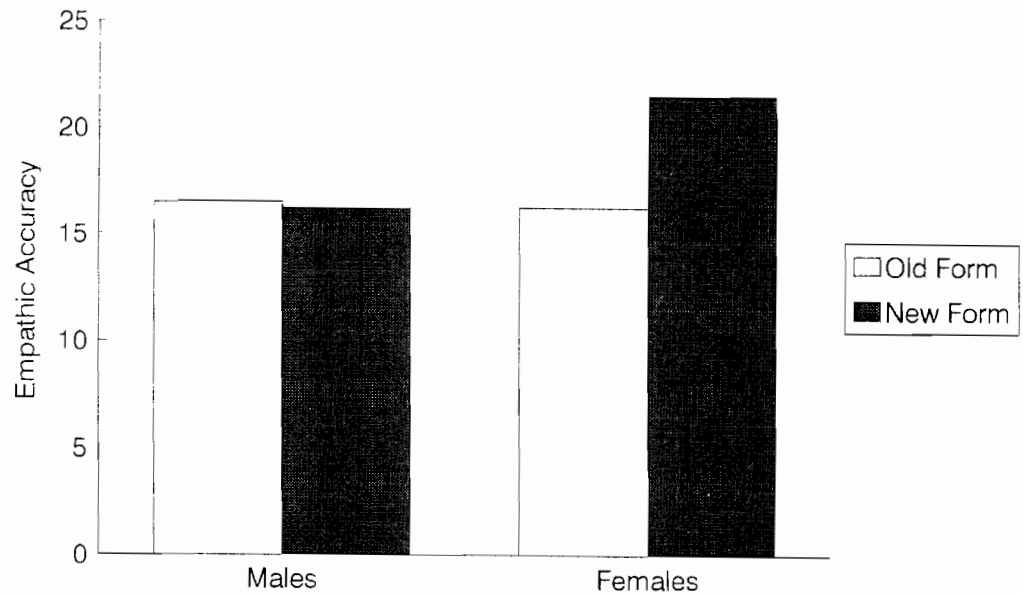
The interpretation of this significant moderating effect will be considered at length below. For the present, an issue that is highly relevant to this interpretation is whether the effect of using the new, as opposed to the old, empathic inference form was to (a) increase the empathic accuracy of the female, but not the male, perceivers, (b) decrease the accuracy of the male, but not the female, perceivers, or (c) increase the females' accuracy slightly but decrease the males' accuracy slightly. Using the data from her dissertation study in combination with data from the Marangoni et al. (1995) study, Graham (1996) was able to provide a preliminary answer to this question. Graham's (1996) study was designed to replicate and extend the findings obtained by

Marangoni et al. (1995). Because some of the cells in her design corresponded to ones in the design used by Marangoni et al. (1995), she was able to apply *t*-tests to make a number of relevant cross-study comparisons.

After transforming the data from each study to *z*-scores to remove any main effect differences between the two studies, Graham (1996) compared the mean empathic accuracy scores obtained by male and female perceivers in the Marangoni et al. (1995) study (in which the old inference form was used) with those obtained by the male and female perceivers who were run in the corresponding cells of her own study (in which the new inference form was used). These comparisons revealed that (a) the female perceivers in Graham's study performed significantly better than did the female perceivers in the corresponding cells of the Marangoni et al. (1995) study, whereas (b) the male perceivers in Graham's study performed exactly the same as did the male perceivers in the Marangoni et al. (1995) study (i.e., the means for the male perceivers in the two studies were virtually identical). These results, depicted in Figure 2, confirmed Graham and Ickes' (1997) assumption that changing the empathic inference form affected the performance of the female, but not the male, perceivers.

## Discussion

The results of the present quantitative meta-analysis of the data from 15 empathic accuracy studies offer compelling statistical support for Graham and Ickes' (1997) more qualitative conclusion that the presence or



**Figure 2.** Cross-study comparison of gender differences in empathic accuracy using the old and new empathic inference forms (Graham, 1996).

absence of significant gender-of-perceiver effects is moderated by the type of empathic inference coding form the perceivers in these studies used. When the coding form required the perceivers to provide self-estimates of their empathic accuracy for each of the thought/feeling inferences they made, women achieved significantly higher empathic accuracy scores than did men in five of the six relevant studies. However, when the coding form did not require the perceivers to provide such self-estimates, the women did not score significantly higher than the men in any of the nine relevant studies.

The moderating effect of the empathic inference form was, in statistical terms, a strong one. There was minimal overlap between the effect sizes found in the subset of studies in which the perceivers did not provide self-estimates of their empathic accuracy and the effect sizes found in the subset of studies in which they did. Moreover, the 95% CI (-.03 to .10) for the mean effect size ( $d = .04$ ) of the nine studies in the first subset included zero, suggesting that, for these studies, the null hypothesis of no gen-

der differences cannot be rejected. In contrast, the 95% CI (.38 to .74) for the mean effect size ( $d = .56$ ) of the six studies in the second subset did not include zero, suggesting that the significant gender differences obtained in these studies were real, robust, and replicable. Finally, the difference between the mean effect sizes (.10 and .62) for the two subsets of studies was clearly significant,  $p < .0001$ .

In summary, when the empathic inference coding form does not require the perceivers to estimate the accuracy of each of their inferences, no reliable gender differences in empathic accuracy are found. It is only when the empathic inference coding form requires the perceivers to make such estimates that reliable gender differences favoring female perceivers are evident.

#### *Graham and Ickes' (1997) motivational interpretation*

How, exactly, should these results be interpreted? The interpretation proposed by Graham and Ickes (1997) is that the significant gender-of-perceiver effects found in

the studies using the newer empathic inference form reflect differences in motivation rather than differences in ability. This interpretation was suggested by Eisenberg and Lennon's earlier reviews of the literature on several other empathy-relevant measures (Eisenberg & Lennon, 1983; Lennon & Eisenberg, 1987). In their reviews, Eisenberg and Lennon concluded that female perceivers score higher than did male perceivers on certain empathy-related measures, but that this effect appears to be limited to studies in which situational cues were available that reminded the participants that some aspect of empathy was being measured, thereby activating the gender-role stereotype that women are supposed to be more empathic than men. The women's higher scores on these tasks might therefore have reflected a higher level of motivation to present themselves as being highly empathic, consistent with the gender-role stereotype, rather than a higher level of empathic ability *per se*.

Graham and Ickes (1997) proposed that the same motivational interpretation could be applied to the gender-of-perceiver data from the 10 empathic accuracy studies that were included in their qualitative review. Specifically, they argued that the effect of the new empathic inference form was to require the perceivers to think about and rate how empathically accurate they were with respect to each thought/feeling inference they made. Thus, in studies in which the new empathic inference form was used, participants should have been more aware that their empathic accuracy was being measured—an awareness that potentially activated the gender-role stereotype that women are supposed to be more empathic than men and thereby motivated the female perceivers to display average levels of empathic accuracy that were reliably greater than those of the male perceivers.

According to this interpretation, differential motivation rather than differential ability accounts for the significant gender differences that were evident in the subset of studies in which the perceivers were asked to provide estimates of their em-

pathic accuracy following each inference they made. The plausibility of this interpretation is further supported by Graham's (1996) finding that changing the empathic inference form seemed to enhance the performance of the female perceivers, but apparently had no effect on male perceivers' performance.

#### *An alternative interpretation*

Can an alternative interpretation be offered to account for the present findings? Could one propose, for example, an interpretation based on the assumption that the significant gender-of-perceiver differences in empathic accuracy found in studies using the newer empathic inference form reflect differences in ability rather than differences in motivation? If so, what supporting assumptions would such an interpretation require?

One might argue that, as a group, women really do have more empathic ability than men do, but that women are only motivated to display their greater ability when situational cues remind them that, according to their gender-role stereotype, they are supposed to excel in this domain. The plausibility of this alternative interpretation can be questioned, however, because it would require us to assume that whenever such situational cues are not available, women either (a) work less hard than men do on empathic accuracy tasks, thus performing no better than men despite their superior ability, or (b) work below their superior potential for some other unspecified reason(s).

For a number of reasons, we find this alternative interpretation to be less compelling than the motivational interpretation proposed by Eisenberg and Lennon (1983; Lennon & Eisenberg, 1987) and by Graham and Ickes (1997). First, it is clearly less parsimonious from a theoretical standpoint, requiring that an additional assumption be made (i.e., that women, more often than men, perform below their potential on empathic accuracy tasks). Second, it is also less parsimonious in the broader sense that it cannot readily account for the patterns of

findings reviewed by Berman (1980), Eisenberg and Lennon (1983; Lennon & Eisenberg, 1987), and the present authors, whereas the motivational interpretation can. Third, the alternative explanation requires its advocates to provide a plausible rationale to account for what would appear to be an unexpected and rather puzzling phenomenon (i.e., why *should* female perceivers so often perform below their potential on empathic accuracy tasks?). Fourth, the alternative explanation itself runs counter to another cultural stereotype (i.e., one that holds that it is men, rather than women, who generally perform below their potential on empathic accuracy tasks). Fifth, the alternative explanation, though seemingly based on the assumption of differential ability, must itself invoke the assumption of differential motivation in order to explain why female perceivers should not always perform better than male perceivers, given their presumably superior level of ability.

Given these conceptual difficulties, we are inclined to accept Graham and Ickes' (1997) conclusion, based on the earlier conclusions of Eisenberg and Lennon (1983; Lennon & Eisenberg, 1987), that the empathic advantage of female perceivers is most likely the product of differential motivation rather than differential ability. Perhaps a viable alternative explanation can be proposed that would deal satisfactorily with each of the five objections we have noted above. Any theorists who would like to propose such an explanation are encouraged to so do. Until they do, however, we must regard the kind of ability-based alternative explanation we have just considered as both unparsimonious and theoretically implausible.

#### Limitations of the Present Study

As intriguing as the present findings are from a theoretical standpoint, there are at least two reasons to question their generality. First, only one "manipulation"—a variation in the empathic inference form—was used to elicit the observed gender differ-

ences in empathic accuracy. Thus, our effects at present are confined to only one "manipulation." Second, all but one of the standard stimulus paradigm studies used the same set of videotapes (the client-therapist interactions) as the stimuli. Hence, there is reason to question whether the effects in these studies would generalize to other standard stimulus tapes as well.

With regard to the first limitation, Klein and Hodges (1999) have recently reported two studies in which a female advantage was evident only when the perceivers rated their sympathy for the target prior to performing the empathic accuracy task. If the sympathy assessment can be interpreted as another way of manipulating the salience of empathy-relevant gender-role expectations, then Klein and Hodges' findings may complement—and extend the generality of—the ones we have reported here. With regard to the second limitation, Klein and Hodges also used a standard stimulus paradigm, but with tapes of their own creation in which college student targets were interviewed about a recent academic problem that each had experienced. Thus, Klein and Hodges' (1999) data may also extend the generality of the present findings with respect to the type of standard stimulus tapes that are used.

#### Implications for Future Research

Although we think the differential-ability alternative to our differential-motivation explanation is implausible on conceptual grounds, we would encourage proponents of this alternative (or a more creative one) to explore its viability in future research. For the present, the contribution of our current meta-analysis has been to demonstrate that the basis of the advantage in interpersonal sensitivity that is commonly referred to as "women's intuition" must be regarded as an open rather than a closed question. Is this advantage the product of differential ability, as is commonly assumed, or is it the product of differential motivation, as Eisenberg and Lennon's broad-scope reviews of the empathy litera-

ture have suggested, and as our current quantitative meta-analysis of the data from 15 empathic accuracy studies also sug-

gests? We pose this question as an important challenge for future research.

## References

- Belk, S.S., & Snell, W.G. (1986). Beliefs about women: Components and correlates. *Personality and Social Psychology Bulletin*, *12*, 403-413.
- Berman, P.W. (1980). Are women more responsive than men to the young? A review of developmental and situational variables. *Psychological Bulletin*, *88*, 668-695.
- Bissonnette, V.L., Rusbult, C.E., & Kilpatrick, S.D. (1996). Empathic accuracy and marital conflict resolution. In W. Ickes (Ed.), *Empathic accuracy* (pp. 251-281). New York: Guilford Press.
- Broverman, I.K., Vogel, S.R., Broverman, D.M., Clarkson, F.E., & Rosenkrantz, P. (1994). Sex-role stereotypes: A current reappraisal. In B. Puka et al. (Eds.), *Caring voices and women's moral frames: Gilligan's view. Moral development: A compendium* (Vol. 6, pp. 191-210). New York: Garland.
- Buck, R. (1981). The evolution and development of emotion expression and communication. In S.S. Brehm, S.M. Kassir, & F.X. Gibbons (Eds.), *Developmental social psychology*. New York: Oxford University Press.
- Buck, R., Miller, R.E., & Caul, W.F. (1974). Sex, personality, and physiological variables in the communication of affect via facial expression. *Journal of Personality and Social Psychology*, *30*, 587-596.
- Buck, R.W., Savin, V.J., Miller, R.E., & Caul, W.F. (1972). Communication of affect through facial expression in humans. *Journal of Personality and Social Psychology*, *23*, 362-371.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cross, S.E., & Madson, L. (1997). Models of the self: Self-construals and gender. *Psychological Bulletin*, *122*, 5-37.
- Eisenberg, N., & Lennon, R. (1983). Sex differences in empathy and related capacities. *Psychological Bulletin*, *94*, 100-131.
- Gesn, P.R. (1995). *Shared knowledge between same-sex friends: Measurement and validation*. Unpublished master's thesis, University of Texas at Arlington.
- 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*, 746-761.
- Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.
- Graham, T. (1994). *Gender, relationship, and target differences in empathic accuracy*. Unpublished masters thesis, University of Texas at Arlington.
- Graham, T. (1996). *Effects of feedback and frames on empathic accuracy*. Unpublished doctoral thesis, University of Texas at Arlington.
- Graham, T., & Ickes, W. (1997). When women's intuition isn't greater than men's. In W. Ickes (Ed.), *Empathic accuracy* (pp. 117-143). New York: Guilford Press.
- Hancock, M., & Ickes, W. (1996). Empathic accuracy: When does the perceiver-target relationship make a difference? *Journal of Social and Personal Relationships*, *13*, 179-199.
- Ickes, W. (in press). Measuring empathic accuracy. In J. Hall & F. Bernieri (Eds.), *Interpersonal sensitivity: Theory, measurement, and applications*. Hillsdale, NJ: Erlbaum.
- Ickes, W. (1997). *Empathic accuracy*. New York: Guilford Press.
- Ickes, W., Bissonnette, V., Garcia, S., & Stinson, L. (1990a). Implementing and using the dyadic interaction paradigm. In C. Hendrick & M. Clark (Eds.), *Review of personality and social psychology: Vol. 11. Research methods in personality and social psychology* (pp. 16-44). Newbury Park, CA: Sage.
- Ickes, W., Gesn, R., Dugosh, J., Pham, H., & Renshaw, K. (1996). More nonsignificant personality correlates of empathic accuracy. Unpublished data, University of Texas at Arlington.
- Ickes, W., Hancock, M., Gesn, P.R., Graham, T., & Mortimer, C. (1995). Nonsignificant personality correlates of empathic accuracy. Unpublished data, University of Texas at Arlington.
- Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990b). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology*, *59*, 730-742.
- Ickes, W., & Tooke, W. (1988). The observational method: Studying the interactions of minds and bodies. In S. Duck, D.F. Hay, S.E. Hobfoll, W. Ickes, & B. Montgomery (Eds.), *Handbook of personal relationships: Theory, research, and interventions* (pp. 79-97). Chichester, UK: Wiley.
- Kelleher, J. (1998). *The effects of frames of reference on empathic accuracy*. Unpublished master's thesis, University of Texas at Arlington.
- Klein, K.J.K., & Hodges, S. (1999). *Gender differences, motivation, and empathic accuracy: When it pays to understand*. Manuscript submitted for publication, University of Oregon.
- Lennon, R., & Eisenberg, N. (1987). Gender and age differences in empathy and sympathy. In N. Eisenberg & J. Strayer (Eds.), *Empathy and its development: Cambridge studies in social and emotional development* (pp. 195-217). New York: New York University Press.
- Manstead, A. (1992). Gender differences in emotion. In A. Gale & M. W. Eysenck (Eds.), *Handbook of individual differences: Biological perspectives* (pp. 355-387). Chichester, UK: Wiley.
- Marangoni, C., Garcia, S., Ickes, W., & Teng, G. (1995). Empathic accuracy in a clinically relevant setting. *Journal of Personality and Social Psychology*, *68*, 854-869.
- Mortimer, D.C. (1996). "Reading" ourselves "reading" others: Actual versus self-estimated empathic accuracy. Unpublished master's thesis, University of Texas at Arlington.
- Renshaw, K. (1998). *Empathic accuracy: The impact of family openness and expressiveness*. Unpublished master's thesis, University of Texas at Arlington.

- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.
- Simpson, J.A., Ickes, W., & Blackstone, T. (1995). When the head protects the heart: Empathic accuracy in dating relationships. *Journal of Personality and Social Psychology*, 69, 629-641.
- Stinson, L., & Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of Personality and Social Psychology*, 62, 787-797.
- Thomas, G., Fletcher, G.J.O., & Lange, C. (1997). Online empathic accuracy and projection in marital interaction. *Journal of Personality and Social Psychology*, 72, 839-850.
- Wispe, L., Kiecolt, J., & Long, R.E. (1977). Demand characteristics, moods, and helping. *Social Behavior and Personality*, 5, 249-255.