

The Meaning and Suitability of Various Effect Sizes for Structured Rater \times Ratee Designs

Johannes Hönekopp
Technische Universität Chemnitz

Betsy Jane Becker
Florida State University

Frederick L. Oswald
Michigan State University

Four types of analysis are commonly applied to data from structured Rater \times Ratee designs. These types are characterized by the unit of analysis, which is either raters or ratees, and by the design used, which is either between-units or within-unit design. The 4 types of analysis are quite different, and therefore they give rise to effect sizes that differ in their substantive interpretations. In most cases, effect sizes based on between-ratee analysis have the least ambiguous meaning and will best serve the aims of meta-analysts and primary researchers. Effect sizes that arise from within-unit designs confound the strength of an effect with its homogeneity. Nonetheless, the authors identify how a range of effect-size types such as these serve the aims of meta-analysis appropriately.

Keywords: meta-analysis, effect size, Rater \times Ratee designs

Null-hypothesis significance testing has been an unparalleled success in the behavioral sciences—a success that has been bemoaned by its critics for several decades (e.g., Cohen, 1995; Meehl, 1967). Arguments against null-hypothesis significance testing claim that researchers do not understand its results, that it carries little information, and even that it forestalls scientific progress (Hunter, 1997; Loftus, 1996). In recognition of these problems, it has become more common in the last 20 years to look at the strength of effects (Wilkinson & Task Force on Statistical Inference, 1999)—that is, not to ask “how probable is it to obtain Result X or a more extreme result in the sample, given that no effect exists in the population?” but, instead, to ask “how strong do we believe Effect X to be in the

population?” This calls for the quantification of the strength of effects, and various ways to do so exist. For the sake of clarity, we want to distinguish the *strength of an effect* from an *effect size*, the latter being a given quantification of the former. If, for example, Treatment A tends to be more effective than Treatment B, A has a stronger effect than B, and this should be reflected in the respective effect sizes. When effect sizes properly reflect the strengths of effects, this allows us to compare effects across different studies (e.g., which type of therapy helps the most?) and to guide decision making (e.g., is this therapy worth its costs?); in both cases, the questions are beyond the simple reach of null-hypothesis significance testing.

As noted in the literature, differences in study design may give rise to incomparable effect sizes (e.g., Morris & DeShon, 2002). Furthermore, and more specific to the present article, different analytic approaches to the same data may also yield effect sizes that differ in meaning and interpretation. Our specific examples focus on studies involving a set of raters (e.g., teachers, supervisors, peers) who rate or score a set of ratees (e.g., students; exam answers; objects, such as pictures) on one or more characteristics. Studies of this type are common in psychological and educational research, in which even under the same research design, researchers may apply very different analyses of rating data. Specifically, in this article, we summarize four typical analytic approaches to rating data and their representations in the effect-size metric d , and we discuss the meanings of the respective results. The meanings of

Editor's Note. William R. Shadish served as action editor for this article.—SGW

Johannes Hönekopp, Institut für Psychologie, Technische Universität Chemnitz, Chemnitz, Germany; Betsy Jane Becker, Department of Educational Psychology and Learning Systems, Florida State University; Frederick L. Oswald, Department of Psychology, Michigan State University.

This work was supported by Deutsche Forschungsgemeinschaft Grant HO 2506/1-1 to Johannes Hönekopp. We thank Frank Renkewitz for useful comments.

Correspondence concerning this article should be addressed to Johannes Hönekopp, Technische Universität Chemnitz, Institut für Psychologie, Wilhelm-Raabe Strasse 43, D-09120, Chemnitz, Germany. E-mail: johannes.hoenekopp@phil.tu-chemnitz.de

effect sizes differ greatly across the four approaches; therefore, we address two key questions: (a) Retrospectively, how should meta-analysts deal with this diversity in effect sizes from past studies? (b) Prospectively, which effect size and related analytic approach should researchers adopt to best achieve their substantive goals?

The research examples we consider have two aspects in common. First, the Rater \times Ratee data are in some way aggregated within a study; second, each study examines differences in the rating data for some dichotomous grouping of the ratees. Although the examples are simple, their important aspects generalize to more complex rating situations. The article proceeds as follows: First, we outline four common analyses of Rater \times Ratee data, drawing on three example studies that are simple and fictitious in nature (the sample sizes, e.g., are much smaller than would be recommended for adequate statistical power, but this allows us to table the data and to guide the reader through the analyses). Second, we illustrate the great diversity of statistical approaches found in primary research by briefly surveying research dealing with the substantive topic of facial attractiveness, specifically whether men or women tend to have more attractive faces. Third, we discuss the meaning of the effect sizes that result from the various statistical approaches. Fourth, we look at several meta-analyses of Rater \times Ratee studies to see how meta-analysts have dealt with the diversity of effect-size types found in primary studies. Fifth, we discuss which properties are desirable for an effect size. Sixth, on the basis of these arguments, we address the aforementioned considerations of meta-analysts and researchers vis-à-vis different analytic approaches.

Four Typical Analytic Approaches to Rater \times Ratee Data

We begin by describing four typical analytic approaches to Rater \times Ratee data from facial-attractiveness research, in which the ratees are faces. In all examples, each of r raters rates several faces, and each of f faces is rated by several raters (thus, in the language of generalizability theory, we are dealing with *one-facet* designs). The four analytic approaches differ in two ways: (a) in their central unit of analysis, which is either raters or faces, and (b) in the nature of the design, which is either within-unit or between-units in relation to the central unit of analysis (rater or face). Table 1 presents all four possible approaches.

Before discussing each approach in detail, we note that some research questions may be less suitable or even impossible to address using any one particular approach described here. Among the characteristics of the research question that play a role in the choice of the subsequent design are the nature of the dichotomous grouping factor (e.g., whether it represents a manipulated or treatment factor or a preexisting status variable, such as gender), the nature

Table 1
Four Basic Approaches to Analysis of Rating Data

Design	Unit of analysis	
	Face	Rater
Between-units	Approach 1	Approach 2
Within-unit	Approach 3	Approach 4

of the ratees (e.g., whether they are real entities or artificial cases), and the nature of the ratee stimuli (e.g., visual stimuli or verbal descriptions). Also, issues of statistical power play a role.

As an obvious example, the question of whether female or male faces are more attractive cannot be tackled using a within-ratee analysis because the same face cannot be presented in both conditions (i.e., as both a female face and a male face). However, for a different outcome (e.g., a rating based on a verbal description), the gender of the ratee could be a within-subject factor, because the ratees could be hypothetical (e.g., Murphy, Herr, Lockhart, & Maguire's, 1986, "paper people"), and a verbal description could be attributed to both male and female ratees. Similarly, it may not be possible to vary certain status variables, such as race, as within-ratee factors, whereas manipulated (treatment) factors may be used in this way. Indeed, treatment factors are often designed to be within-study factors to capitalize on the increase in statistical power that comes with the blocking or pairing of similar cases.

Approach 1: Between-Ratees Analysis

Approach 1 centers on the faces being rated. Our fictitious example study, Study 1, focuses on the question of whether male or female faces tend to be rated as more attractive. A typical facial-attractiveness study may use a convenience sample of female and male faces, typically from the college-student population, to be evaluated by several raters (e.g., Langlois & Roggman, 1990). Each face may be rated by the same rater or by different raters; in the present example, different sets of raters rate the faces.

Table 2 provides the data for Study 1. Here, three female and three male faces were each evaluated by three raters on an 11-point scale ranging from 0 to 10, where higher ratings indicate a higher level of attractiveness. Faces are the central unit of analysis in Approach 1: First, we determine the average attractiveness rating for each face, which we call its *face score* or, more generally, its *ratee score* (see bottom section of Table 2). In this example, $n = f = 6$, and the average face scores for male (2.667) and female faces (5.333) are compared using an independent-groups t test, which results in $t(4) = 1.57, p = .19$, a nonsignificant mean difference favoring female faces. The standard deviation for each group depends only on variability between the face scores, and the analysis does not consider the variability of

Table 2
Approach 1: Data From Study 1

Rater	Faces					
	Female			Male		
	f_{f1}	f_{f2}	f_{f3}	f_{m1}	f_{m2}	f_{m3}
r_1	10	7	4			
r_2	6	6	3			
r_3	5	5	2			
r_4				7	4	1
r_5				5	2	2
r_6				3	0	0
Face score	7	6	3	5	2	1
M	5.33			2.67		
SD	2.08			2.08		

ratings within each face (i.e., the fact that different raters give different ratings for the same face). Approach 1, in principle, does not depend on the design of the study; that is, it could also be used if the same raters rated both groups of faces.

Approach 2: Between-Raters Analysis

Let us also use Study 1 to consider Approach 2, which uses raters as the central unit of analysis. Rather than computing the average attractiveness rating for each face, as we did in Approach 1, we now compute the average attractiveness judgment for each rater, which we call the *rater score* (see the last two columns of Table 3). Approach 2 compares the rater scores between the two groups of raters who rated male and female faces, with $n = r = 6$ using a t test for independent samples. This yields $t(4) = 2.14$, $p = .10$, a mean difference favoring raters of female faces. Here, the standard deviation for each group depends only on variability in the rater scores, and the analysis does not consider the variability of ratings within each rater (i.e., the fact that each rater gave some faces higher ratings and other faces lower ratings).

Comparing Approaches 1 and 2 to analyzing the data from Study 1, we necessarily find the same raw mean difference between sets of faces as between sets of raters (2.67 points), but the approaches produce different t values and statistical significance levels because of differences in the variability against which each raw mean difference was compared (i.e., variability in face scores in Approach 1 and variability in rater scores in Approach 2).

Approach 3: Within-Ratee Analysis

Approach 3 uses a repeated measures design, using relevant pairs of ratees (here, face stimuli) as the central unit of analysis. As noted above, Approach 3 is not suitable for addressing the question of whether female or male faces are

more attractive, because the same face cannot be presented as a female face and a male face. Therefore, we consider a research question different from the previous one: Does makeup affect female facial attractiveness? Study 2 addresses this question (see Table 4). Here, three raters rated three women, each woman having two pictures—one with makeup and one without. We analyze these data by computing the face score for each picture (see the left part of Table 4, third row from the bottom). We then compare the faces with and without makeup using a repeated measures t test, which is based on the face-score differences within the pairs of face stimuli (i.e., face score for face i with makeup minus face score for i with no makeup). The mean difference is 2.67 ($SD = 1.15$). This results in $t(2) = 4.00$, $p = .06$, a mean difference favoring female faces with makeup. The mean of the face-pair differences within raters is equal to the mean face-score difference across raters, which is always the case. Also, the standard deviation of the mean pairwise face-score difference depends only on the variability of the pairwise face-score differences. Like Approach 1, the analysis does not consider variability of the ratings within each face (i.e., independent of the face score, some raters gave higher scores to a face than did other raters). In addition, it ignores variability in facial attractiveness between the three rated faces (i.e., each rated face may have had relatively higher or lower levels of facial attractiveness than the average face).

Approach 4: Within-Rater Analysis

Approach 4 also uses a repeated measures design, using raters as the central unit of analysis rather than faces. Let us reconsider Study 2, this time computing two rater scores for each rater—one the average of ratings for faces with makeup and one the average of ratings of faces without makeup (see Table 5). We then compare these averages using a repeated measures t test, which is based on the differences of rater scores within each rater (i.e., rater score for faces with makeup minus rater score for faces with no

Table 3
Approach 2: Data From Study 1

Rater	Faces						Rater score	
	Female			Male			Female faces	Male faces
	f_{f1}	f_{f2}	f_{f3}	f_{m1}	f_{m2}	f_{m3}		
r_1	10	7	4				7	
r_2	6	6	3				5	
r_3	5	5	2				4	
r_4				7	4	1		4
r_5				5	2	2		3
r_6				3	0	0		1
M							5.33	2.67
SD							1.53	1.53

Table 4
Approach 3: Within-Ratee Data From Study 2

Rater	Faces						Pairwise difference			
	Makeup			No makeup						
	f_1	f_2	f_3	f_1'	f_2'	f_3'	f_1-f_1'	f_2-f_2'	f_3-f_3'	
r_1	10	7	4	7	4	1	3	3	3	
r_2	6	6	3	5	2	2	1	4	1	
r_3	5	5	2	3	0	0	2	5	2	
Face score	7	6	3	5	2	1	2	4	2	
M								2.67		
SD								1.15		

makeup). The mean difference is 2.67 points, which results in $t(2) = 8.00, p = .02$, a mean difference favoring female faces with makeup. The standard deviation of the mean pairwise rater-score differences (0.58) depends only on the variability of the pairwise rater-score differences. As in Approach 2, the analysis ignores the variability of each rater’s ratings within the two conditions. In addition, it ignores the variability of rater scores between raters (i.e., the fact that some raters tend to give high ratings, whereas others tend to give low ratings).

Paralleling Study 1, Approaches 3 and 4 to Study 2 involve the same raw mean difference of 2.67 points, yet they produce very different t -test results based on the variability against which the raw mean difference was compared. In principle, it would be also possible to apply Approach 1 and Approach 2 to the data of Study 2. This would result in the same t -test results that were obtained for Study 1, because Studies 1 and 2 have the same raw scores; we just changed the experimental design from a between-raters/between-ratees to a within-rater/within-ratee design and gave it a different research context. However, if Approaches 1 and 2 were applied to data from a true within-study design, a loss of power would be typical, because, as mentioned above, pairing is often used to reduce variation, and the analyses used in Approaches 1 and 2 would ignore this pairing.

Our examples show how identical data yield four different t tests and p values, depending on the experimental design and on the analysis chosen. As noted above, the samples for the two studies here are admittedly small (for didactic purposes). Therefore, the p values tend to be larger and, for the most part, not near conventional levels of statistical significance. However, that is not particularly relevant to our argument. Our point is that the analytic approach and corresponding t values are different both conceptually and empirically. We next take a look at actual research practice and consider these differences more closely.

How the Four Approaches Are Used in Research

The previous examples suggest that research practice could be very diverse when it comes to the analysis and interpretation of Rater × Ratee data. We have illustrated this by considering research on facial attractiveness, but we later demonstrate that practice in other research domains is probably also diverse. First, we look at actual research that informs the question of whether female or male faces tend to be rated as more attractive. Searches in PsycINFO and MEDLINE databases as well as searches of citations in recent facial-attractiveness studies revealed 55 documents comparing female and male facial attractiveness (excluding

Table 5
Approach 4: Within-Rater Data From Study 2

Rater	Faces						Rater score			
	Makeup			No makeup			Makeup	No makeup	Pairwise difference	
	f_1	f_2	f_3	f_1'	f_2'	f_3'				
r_1	10	7	4	7	4	1	7	4	3	
r_2	6	6	3	5	2	2	5	3	2	
r_3	5	5	2	3	0	0	4	1	3	
M								2.67		
SD								0.58		

those that used highly artificial stimuli such as line drawings). Six had insufficient data for computing any effect size. The 49 remaining documents (marked with asterisks in the References) contained 79 samples. Seven studies presented two analyses for the same sample data, yielding a total of 86 analyses. One study used an approach not captured by our classification scheme. Most common was Approach 1, with 38 occurrences (44%).¹ We found 24 instances of Approach 2 (28%) and 23 instances of Approach 4 (27%). Approach 3 does not apply to this particular research question, because it would require that the same face be presented once as female and once as male. Researchers can freely choose between Approaches 1 and 2 or Approach 4 if faces of both sexes are shown to each rater.

It is therefore instructive to look at the subsample of studies that presented faces of both sexes to each rater. This category included 45 documents, with 69 samples and 75 analyses. Approach 1 occurred 28 times (37%), and Approaches 2 and 4 occurred 23 times each (31%). Thus, at least in this domain of research (and likely in many others), meta-analysts are roughly equally likely to encounter each of these three types of information in Rater \times Ratee designs.

What gives rise to the diversity of approaches we found within this domain? Authors usually did not discuss their specific choice of an approach, and we cannot discern any pattern that explains why researchers chose one approach and not another. For example, we did not find that researchers based their analysis on ratees when the number of ratees was larger than the number of raters, or vice versa. However, studies that used Approach 4 tended to have more complex designs than the other studies. Cross and Cross (1971), for example, used Approach 4 and looked not only at the effect of ratees' sex but also at the effects of ratees' age and race and raters' age, sex, and race.

The Meaning of Effect Sizes Within the Four Common Approaches

We next look at the meaning of d within the four approaches. However, before examining the d values that result from the four approaches, we briefly recapitulate the purpose of d by looking again at the results of Study 1. We find that female faces ($M = 5.33$) are, on average, judged to be more attractive than male faces ($M = 2.67$). The raw difference of 2.67 points is the same, whether we look at the face scores (see Table 2) or at the rater scores (see Table 3). What does that mean? We are not dealing with a rating on a ratio scale, so we cannot conclude that women are, on average, twice as attractive as men. It would be rather misleading, too, to divide the raw difference of 2.67 scale points by the scale range of 10 points and conclude that women exceed men by 27% of the scale range.² This is because the average rated difference between women and

men—and sometimes its associated variability—could change substantially if we used a scale that ranged from, say, -5 to 5 instead of 0 to 10 (e.g., Schwarz, 1999). A useful effect-size statistic gives meaning to the obtained raw difference by comparing it with a yardstick based on the variability in the data.

To that end, an effect-size measure commonly applied to group differences is d , which expresses the between-groups difference in the unit of the standard deviation computed from the weighted average within-group variance (Cohen, 1969; Hedges, 1981). Specifically,

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}} = \frac{\bar{X}_1 - \bar{X}_2}{s_{\text{pooled}}}$$

Here \bar{X}_1 and \bar{X}_2 are the means for the samples of n_1 and n_2 cases in the two groups of the dichotomy of interest, and s_1^2 and s_2^2 are the sample variances in the groups.

Below, we discuss the meaning of d in light of Approaches 1–4. In our examples, we compute d from the raw data because this makes it easier to grasp its meaning. Meta-analysts, of course, rarely have this opportunity and must rely on the descriptive statistics or test statistics provided (e.g., Rosenthal, 1994; Rosenthal, Rosnow, & Rubin, 2000). For convenience, we use d_1 , d_2 , d_3 , and d_4 to represent the effect sizes for Approaches 1–4, respectively.

The Meaning of d_1 for Approach 1

In all four approaches, the effect size is based on the mean difference between two groups relative to a measure of variability within groups. In the first two examples (for Study 1), the mean difference is 2.67. What changes between the examples is the variability that qualifies this difference. With Approach 1, it is the variability of face scores that matters. The standard deviation is 2.08 for both groups in Study 1, which means that $s_{\text{pooled}} = 2.08$ as well, therefore $d_1 = 2.67/2.08 = 1.28$ (Table 2 shows the means

¹ Often no tests were performed, but means and standard deviations were given. Where standard deviations were based on face scores, we regard this as Approach 1; where standard deviations were based on rater scores, we regard this as Approach 2.

² Some authors have argued that when a scale with intrinsic meaning is used, the raw-score scale is more appropriate for reporting the size of effects than any standardized metric (see, e.g., Bond, Wiitala, & Richard, 2003). We concur; however, we suspect that when ratings are of interest, it is rare that the outcome scales have clear intrinsic meanings.

Table 6
Approach 1: Data From Study 3

Outcome	Face score			r	SD		d
					Treatment	Observed	
Control group	5	3	1			2.00	
Alternative 1	5	3	1	0	0.00	2.00	1.00
	+2	+2	+2				
	7	5	3				
Alternative 2	5	3	1	1.0	1.00	3.00	0.78
	+3	+2	+1				
	8	5	2				
Alternative 3	5	3	1	.5	1.00	2.65	0.85
	+3	+1	+2				
	8	4	3				
Alternative 4	5	3	1	−1.0	1.00	1.00	1.26
	+1	+2	+3				
	6	5	4				
Alternative 5	5	3	1	−1.0	2.00	0.00	1.41
	+0	+2	+4				
	5	5	5				

and standard deviations for the study).³ Thus, Study 1 indicates that on average, female faces look about 1.3 face-score standard deviations better than male faces. Although this is a large effect size, because of our small samples, it is not significant, $t(4) = 1.57$, $p = .19$ (as noted above). Although the meaning of d_1 appears straightforward, three aspects may complicate matters: the variability of ratee scores, the heterogeneity of treatment effects, and the meaning of ratee-score differences. We discuss these aspects in turn.

First, the size of d_1 hinges on the variability of ratee scores. Consequently, the sampling of ratees is crucial for the size of d_1 . If, for example, the face samples consist of faces whose attractiveness is close to the mean of their respective sex, d_1 would be large because of the low variability in face scores. Conversely, if researchers picked for each sex one face from the 1st, 50th, and 99th percentiles of the attractiveness continuum, d_1 would be small because of the high variability in face scores. Consequently, to yield a nonarbitrary and meaningful d_1 , the variability of attractiveness in the face sample should match the variability of attractiveness in the population of interest.

Second, if some treatment is evaluated, the homogeneity of the treatment effect affects d_1 . In our example, both groups show the same sample variability in face scores. Of course, it is not necessary—though it is typical—to assume that $\sigma_1^2 = \sigma_2^2$ in the population. Moreover, differences in the variances between the groups studied provide important information if some treatment is evaluated: If the group variances differ, this may suggest that the treatment has differential effects on the scores of ratees within the groups (Bryk & Raudenbush, 1988), which in turn affects d_1 .

To see why this is so, we consider several sets of fictitious results for Study 3, the evaluation of makeup on ratings of attractiveness in a between-groups design (the results of which are again analyzed following Approach 1). Three women are randomly assigned to the control group and photographed without makeup; three other women are assigned to the treatment group and photographed with makeup. Table 6 shows the face scores for the three women in the control group and the results for five alternative treatment outcomes. In each scenario, we represent the observed scores in the treatment group as a sum of two components that, in an actual study, cannot be observed or deduced: The first unobservable component represents the face score a specific face would have received without treatment; these values are identical to the scores of the control participants. The second unobservable component indicates a hypothetical treatment effect (i.e., how much

³ Owing to differences among raters and measurement error, the variance of the face scores is smaller than the variance of the individual ratings ($SD = 2.35$ for both groups). Thus, the ratio of the two variances carries information about the reliability of the face scores. Generalizability theory (e.g., Shavelson & Webb, 1991) addresses the reliability of aggregated scores and quantifies errors from various sources (e.g., sampling of ratees, sampling of raters, time of testing). Researchers may wish to adjust the obtained effect size for the unreliability of measurements (e.g., Rosenthal, 1994). However, we do not dwell on reliability here because the meaning of various types of effect sizes is independent from the reliability of the data, and differences in effect sizes in our fictitious examples do not arise from differences in reliability.

each particular face gained in overall attractiveness through the application of makeup).⁴

All alternative treatment outcomes are identical with respect to the first component—the unobservable no-treatment face scores (the average rating across all no-treatment faces). Moreover, for all the hypothetical outcomes, the treatment has the same strength: It elevates on average each face's face score by 2 points. However, the alternative outcomes differ with respect to the heterogeneity of the treatment effect (represented by the standard deviations listed in Table 6). Also, the alternatives differ with respect to the correlation between the treatment effect and the no-treatment face scores (see the column labeled r in Table 6). As can be seen, a positive correlation between the treatment effect and the no-treatment face scores (i.e., a heterogeneity-inducing treatment) results in the observed variance in the treatment group being larger than the variance in the control group (i.e., compare the control-group data with data for Alternative 2); conversely, a negative correlation between the treatment effect and the no-treatment face scores (i.e., a homogenizing treatment) prompts the observed variance in the treatment group to become smaller than the variance in the control group (e.g., data for the control group vs. that for Alternative 4). As a consequence, Alternatives 2 and 3 yield smaller d s than Alternative 1, whereas Alternatives 4 and 5 produce larger d s than Alternative 1 (see last column in Table 6).

If the variances in naturally occurring groups differ, this may be informative as well. It suggests either that the subjects in different groups are affected by different forces or that the same forces act on them in different ways. For example, men have only one X chromosome. Women, in contrast, have two different X chromosomes, and in each cell, a chance process determines which one of the two is switched off. Some have argued that this difference between the sexes and the fact that many genes affecting intelligence are located on the X chromosome give rise to greater variance in intelligence in men than in women (Check, 2005). Greater variance in male scores has been reported for a variety of cognitive outcome variables (e.g., Feingold, 1992; Hedges & Friedman, 1992).

Third, the meaning of rater differences is not always straightforward. For example, standards of attractiveness may depend on the age of the ratee. Consequently, variance across a mixed-age sample of pictures would partly reflect true differences and partly reflect differing standards, which makes the meaning of d_1 ambiguous.

In sum, if d_1 pertains to the effect of a treatment, it does not reflect only the average strength of an effect. Given that the variances of treatment and control groups differ, d_1 additionally reflects the heterogeneity of the treatment effect across treated entities and the correlation of the treatment effect with the no-treatment values. Use of the control-group standard deviation instead of the pooled standard

deviation avoids this ambiguity by dividing the between-groups difference by the standard deviation of the control group (Glass, McGaw, & Smith, 1981), which has prompted some to argue for this metric (which we denote as g). However, there are good reasons to use d , the most prominent being that one is more likely to find the data needed to compute d than one is to find the data needed to compute g and that most studies assume homogeneity of variance and typically do not provide a way to assess whether the homogeneity-of-variance assumption is not met.

The Meaning of d_2 for Approach 2

For Approach 2, the group mean difference is standardized on the basis of the variance of rater scores. For Study 1, we have the same mean difference of 2.67 as before, but this time the standard deviation of interest is 1.53, as shown in Table 3. Consequently, $d_2 = 2.67/1.53 = 1.75$. Thus, Study 1 also indicates that female faces look almost 2 rater score standard deviations better than male faces (though the difference is again not significant, $t[4] = 2.14, p = .10$). In this study, rater scores show less variability than face scores, therefore $d_1 < d_2$. Of course, it is not necessary that this be the case. The opposite would occur if the variability of face scores were smaller than the variability of rater scores, in which case $d_1 > d_2$.

Regarding the meaning of d_2 , three aspects are important. First, the size of d_2 hinges on the variability of rater scores. Consequently, the sampling of raters should guarantee that the variability of rater scores in the sample matches the variability in the population of interest. Otherwise, the size of d_2 may be misleading.

Second, if d_2 reflects the effect of some treatment, the homogeneity or heterogeneity with which raters respond to that treatment affects the size of d_2 in the same way as was discussed for d_1 . For example, raters may respond more or less homogeneously to facial makeup. Assuming that Table 6 showed rater scores instead of face scores, this provides a suitable example.

Third, the meaning of rater-score differences tends to be somewhat ambiguous. Consider our examples of facial attractiveness: Do rater-score differences reflect the fact that some raters generally find all faces more attractive than other raters do? Or do they indicate that raters systematically differ in the way in which they transform the same internal judgment into a response? Perhaps it is some mixture of both, and therefore the variance in rater scores—and, consequently, d_2 —may be hard to interpret.

⁴ Although we use the term *treatment* here, the ideas also apply to studies of group differences on status variables such as race and gender. This terminology usage appears in many discussions of analysis-of-variance “treatment effects.”

Table 7
Approach 3: Alternative Outcomes for Study 2 Data

Outcome	Face score									M	SD	d
	Makeup			No makeup			Pairwise difference					
	f ₁	f ₂	f ₃	f ₁ '	f ₂ '	f ₃ '	f ₁ -f ₁ '	f ₂ -f ₂ '	f ₃ -f ₃ '			
Original outcome	7	6	3	5	2	1	2	4	2	2.67	1.15	2.32
Alternative 1	8	7	4	5	2	1	3	5	3	3.67	1.15	3.19
Alternative 2	7	7	3	5	2	1	2	5	2	3.00	1.73	1.73
Alternative 3	5.3	2.2	1.2	5.0	2.0	1.0	0.3	0.2	0.2	0.27	0.06	4.50

The Meaning of d₃ for Approach 3

Approach 3 is based on pairwise face-score differences. Because there is only one group of change scores, Equation 1 cannot be applied to Approach 3 unless the pairing of scores is ignored and differences are not computed. Instead, d₃ is typically computed as the mean of the pairwise face-score differences relative to the difference-score standard deviation (see Table 4).⁵ That is,

$$d_3 = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{s_1^2 + s_2^2 - 2r_{12}s_1s_2}}$$

where r₁₂ is the correlation between the paired scores. Consequently, for Study 2, d₃ = 2.67/1.15 = 2.32. The meaning of d₃ is highly ambiguous in the sense that its magnitude confounds the strength of the manipulation (here, the effect of the use of makeup) with the homogeneity of the treatment effect across faces. This becomes clear when we look at three hypothetical alternative outcomes for Study 2, shown in Table 7 (see also Dunlap, Cortina, Vaslow, & Burke, 1996). In these three cases, the original outcome and all three alternative outcomes are identical with respect to the face scores in the no-makeup control condition; differences only occur with respect to the treatment (i.e., the effects of makeup). Alternative 1 reflects a higher d value than the original outcome, which was created by making all of the ratings for faces with makeup 1 point higher than in the original data (d₃ = 3.67/1.15 = 3.19 vs. the original d₃ of 2.32). Here, the values of s₁ and s₂ do not change, and the correlation between paired scores (r₁₂) is .85. But now consider Alternative 2. Here, one face with makeup is rated higher than in the original data, and the other ratings remain the same. This results in less homogeneity between ratings. The standard deviation of the pairwise face-score differences rises from 1.15 to 1.73 because not only is there greater variation in the makeup ratings but the correlation between paired scores is also lower (r₁₂ = .69). As a consequence, d₃ declines from 2.32 to 1.73, although for one face the makeup manipulation was more effective! Alternative 3 demonstrates that the opposite effect can occur too. Compared with the original outcome, the effectiveness of the makeup manipulation is much smaller

in Alternative 3. In the original data, if only one woman applies makeup, her rank within her group can change considerably. For instance, Woman 3 would change her rank from 3 to 2. For Alternative 3, however, this never happens. The treatment effect is small and nearly identical for all faces in Alternative 3, but the value of d₃ soars to 4.5 standard-deviation units (see Table 7), largely because of a reduction in variability (SD = 0.06). This reduction occurs because even though there is slightly more variance in the makeup face scores than in the original data, the correlation between the scores for the paired faces in Alternative 3 is 1.0.

These examples reveal that d₃ confounds the effectiveness of the treatment with the homogeneity of its effect across the treated entities. This ambiguity is not restricted to any particular domain studied (e.g., facial attractiveness); rather, this property is inherent in the nature of d₃.

As discussed earlier, d₁ may confound the strength of the treatment effect and its homogeneity. However, homogeneity affects d₁ and d₃ in ways that are both different and important. For d₃, greater homogeneity of the treatment effect always yields a larger effect size. In contrast, homogeneity in the treatment component may either increase or decrease d₁ (cf. Alternative 1 with Alternatives 2 and 4 in Table 6: Alternatives 2 and 4 show equal variances for the treatment component, but Alternative 2 shows a smaller d₁ value than Alternative 1, whereas Alternative 4 shows a larger d₁). More important, the impact of the heterogeneity of the treatment effect on d₁ is rather limited, but conversely, treatment homogeneity can inflate d₃ to an extreme degree. This will be the case if other factors producing change unrelated to treatment (e.g., maturation rates) end up producing little variance in change scores. In sum, we can expect d₁ to track the strength of a treatment effect more accurately than does d₃.

⁵ Some meta-analyses with a mix of independent-groups and paired designs ignore the pairing and apply Equation 1 to the paired data. This can be done for instance if the means and standard deviations are given for each of the paired scores.

The Meaning of d_4 for Approach 4

Approach 4 qualifies the between-groups difference (which is the same value shown for Approach 3) by the standard deviation of the pairwise rater-score differences (consequently, Equation 1, again, cannot be applied when only one set of change scores exists). Therefore, $d_4 = 2.67/0.58 = 4.60$ in Study 2 (see Table 5). Approaches 3 and 4, although based on the same mean difference from Study 2, can yield very different d values. The reason is that the effect of some grouping factor can be more homogeneous across raters than across faces, and vice versa. Approaches 2 and 4 both focus on rater scores. Nonetheless, unlike d_2 , d_4 is not plagued by the ambiguities that stem from differences among rater scores across raters. Because Approach 4 looks at differences between rater scores within the same rater, the overall differences between raters do not affect d_4 , yet the meaning of d_4 remains highly ambiguous. Just as with d_3 , d_4 confounds the effectiveness of the treatment with the homogeneity of the effect across raters. Thus, when we obtain, for example, a large d_4 , we cannot untangle the extent to which the makeup manipulation led to differences and the extent to which raters agree on the effect of the treatment. Again, this property of d_4 is independent of the substantive phenomenon studied.

Finally, comparisons can be made between d_2 and d_4 . The issues parallel those previously discussed in the comparison of d_1 and d_3 exactly.

What Do Meta-Analysts Do? Two Examples

So far, we have demonstrated for one domain (attractiveness differences between two sets of faces) that various approaches to Rater \times Ratee data are prevalent in primary studies, and we have shown that (independent from the domain studied) the effect sizes that come along with various approaches have different meanings. Are there actual meta-analyses that deal with various approaches to Rater \times Ratee data? And if so, how do meta-analysts treat these differences? To address these questions, we look in detail at two meta-analyses on physical attractiveness (Eagly, Ashmore, Makhijani, & Longo, 1991; Langlois et al., 2000), and we briefly consider other syntheses with different outcomes and effect-size metrics.

Eagly et al. (1991)

Eagly et al. (1991) looked at the question of which qualities are attributed to attractive versus unattractive strangers. When we inspected 17 primary studies included in their meta-analysis that were immediately available to us, we found studies using Approach 1 (e.g., Barnes & Rosenthal, 1985), Approach 2 (e.g., Bassili, 1981), and Approach 4 (e.g., Alicke, Smith, & Klotz, 1986), whereas we found no instances of Approach 3 (though some studies

inaccessible to us may have used this approach). How did the authors deal with this diversity? They computed several kinds of effect sizes—for between-groups studies they computed d_1 and d_2 , and for within-subject designs they used d_4 . The authors computed a single averaged effect size across all studies (i.e., across approaches) and an average effect size for each of several attribute types (e.g., social competence, adjustment). All of these analyses yielded heterogeneous effect-size distributions. Consequently, the authors examined whether various study characteristics significantly affected effect sizes. One of these characteristics was within-subject versus between-subjects designs in attractiveness research, and they found that “consistent with the greater control of extraneous variables and the more precise error term in within-subjects designs, within-subjects variations of attractiveness produced a stronger stereotype than between-subjects variations” (Eagly et al., 1991, p. 119). In sum, Eagly et al. (1991) acknowledged that between-subjects designs and within-subject designs tend to yield effect sizes of different magnitudes; they appeared to prefer the within-subject design, but they did not explicitly acknowledge that choosing raters or ratees as the unit of analysis has consequences for the effect size. In the end, they readily averaged across different types of effect size.

Langlois et al. (2000)

In a series of meta-analyses, Langlois et al. (2000) investigated substantively interesting covariates of physical attractiveness. Here, we examine two of these meta-analyses, one focusing on rater judgment and one focusing on the treatment of attractive versus unattractive adults. Again, we find three of the above-mentioned approaches to data analysis in the primary studies from both meta-analyses (Approach 1: e.g., Raza & Carpenter, 1987; Approach 2: e.g., Langlois, Roggman, & Rieser-Danner, 1990; Approach 4: e.g., Langlois, Ritter, Roggman, & Vaughn, 1991); but again, Approach 3 is either rare or missing in these domains. Langlois et al. (2000) did not address the issue of diverse effect-size types having different meanings, but in their analyses, they seemed to prefer d_4 over d_2 whenever primary studies allowed computation of both effect-size types (see their treatment of Marlowe, Schneider, & Nelson, 1996). In both meta-analyses, the authors readily averaged effect sizes across different effect-size types. For both meta-analyses, Langlois et al. (2000) also computed an average effect size for each of several subsets of results (e.g., judgments of occupational competence, judgments of social appeal, judgments of adjustment). On this basis, they concluded that higher ratings tend to advantage attractive adults most for the domain of occupational competence. However, this conclusion seems doubtful: The average effect size for occupational competence rests on five studies, two of which were accessible to us (Dipboye, Fromkin, & Wiback, 1975;

Marlowe et al., 1996). These two studies, which together contributed more than half of the total sample size in the analysis of occupational competence, provide d_4 as an effect size, whereas the analyses of judgment of social appeal and adjustment are largely based on d_1 .

Meta-Analyses With Other Metrics and Outcomes

Many other meta-analyses have examined rater effects for outcomes other than attractiveness, using metrics other than the standardized mean difference. A search of PsycINFO for meta-analyses (i.e., via Florida State University's online version of Cambridge Abstracts, with *meta-analysis* specified as the methodology) with the terms *rater* or *ratee* anywhere in the record yielded 79 records. (Other syntheses on Rater × Ratee data that do not use these terms likely also can be found.) These studies examined a variety of outcomes (e.g., aspects of leadership, job performance), and several grouping variables (e.g., age, gender, and race) are prominent. In one example, Eagly, Makhijani, and Klonsky (1992) examined gender effects in the evaluation of leaders (the ratees). Eagly et al. (1992) coded the nature of the factor for leader gender—reporting designs with within-subject, between-subjects, or “other, mixed, or unclear” (p. 9) gender factors. Most of the designs had gender as a between-subjects factor. Different methods were used to compute standardized-mean-difference effect sizes for between-subjects and within-subject designs, yet in subsequent analyses, the two kinds of effects were not distinguished when considered together.

Which Properties Are Desirable for an Effect Size?

So far, we have identified the meanings of effect sizes that arise from four common approaches to analyzing Rater × Ratee data; we have also shown that meta-analysts have generally failed to consider the different meanings of various effect-size types and, further, that such failures may give rise to questionable conclusions. Thus, two important questions remain: Given the many types of effect-size computations that primary studies have produced, what should a meta-analyst do? And which of these effect sizes are recommended to primary researchers? To address these questions in full, it is helpful to determine what characteristics are desirable for an effect size. We propose that a good effect size should suit two types of “consumers”—people who want to gain insight into a substantive topic (usually fellow scientists) and decision makers who seek aid in deciding about some practical course of action. We argue that both parties' interests are best served by effect sizes that (a) are unambiguous, (b) reflect primarily the strength of an effect, and as a side benefit (c) have a meaning that is easy to grasp. When we take the *scientist-as-consumer* perspective, the desirability of these properties seems self-evident

for us, because science strives for clarity and simplicity and because it is the very idea of an effect size to reflect the strength of an effect (and not its heterogeneity).

From the *decision-maker-as-consumer* perspective, the picture is a bit more complicated. Let us assume that the decision pertains to the application or rejection of some treatment, the cost of which is known and the benefit of which is described as an effect size. If we assume that the decision maker tries to maximize expected utility (Dawes, 1988; Savage, 1954), the solution to the decision problem is simple: If the expected net utility of the treatment is positive (i.e., the utility of its expected benefit is greater than the disutility of its costs), then apply the treatment; otherwise, do not. In this case, an effect size that correctly reflects the expected value of the treatment benefit contains all the information desired.

However, human decision makers rarely seek to maximize their expected utility: Most people, for example, would prefer to receive \$95 for sure over a gamble that, with equal probability, either yields \$200 or nothing (Kahneman & Tversky, 1979). They would thus sacrifice some expected utility to gain a reduction of outcome variance, which shows that outcome variability has some utility in itself. In a case in which the decision pertains not to a single application of the treatment but to its permanent implementation, the heterogeneity of the treatment effect can have some utility in itself too, but its desirability may be less clear. Given the choice between treatments that differ only with respect to the heterogeneity of their effects, policymakers may favor the treatment that shows the more homogeneous or consistent effect (e.g., an educational intervention from which all students receive the same benefit), the one that homogenizes (e.g., an educational treatment from which low-aptitude students benefit most), or the one that dehomogenizes (e.g., an educational treatment targeted toward high-aptitude students, who tend to benefit most).

Therefore, it will be valuable to some decision makers to know how a treatment affects variation and whether it tends to be homogenizing or variance inducing. Are effect sizes that reflect these treatment characteristics especially valuable for decision makers? Without knowing exactly the substantive causes for variability, the answer is clearly no.

To see why, consider first a policymaker who prefers to implement a treatment that reduces outcome differences. He or she would be willing to sacrifice a certain amount of treatment effectiveness for an increase in the homogenizing capacity of the treatment. Thus, many different combinations of treatment effectiveness and treatment capacity for homogenization exist, all having the same utility for the policymaker (one treatment may be highly effective but dehomogenizing, another treatment may be somewhat less effective but homogenizing, etc.). Now consider another decision maker who favors a dehomogenizing treatment (he or she may, e.g., strive for an elite group and thus advocate an educational treatment from which high-aptitude pupils

get the greatest benefit). For this person, the very same combinations of treatment effectiveness and capacity for homogenization will differ very much in their utility. Thus, the way in which a treatment's effectiveness, the heterogeneity of its effect, and its capacity for homogenization or increasing heterogeneity blend into its utility differs from decision maker to decision maker. But the way these three aspects blend into the sizes of d_1 and d_2 , on the one hand, and d_3 and d_4 , on the other, is fixed and, thus, cannot be homologous to the way in which these three aspects are linked to the preferences of diverse decision makers. For this reason, decision makers as a whole are best served by an effect size that reflects the effectiveness of the treatment and nothing else. Other information is valuable, too, but for the reasons given above, it is of no use to mix this information with the effectiveness of the treatment.

From our discussion, it follows that d_1 and d_2 serve the needs of researchers and decision makers better than do d_3 and d_4 , because they more accurately reflect the effectiveness of the treatment. However, a prerequisite for d_1 and d_2 to be useful is that the variability in rater scores (d_1 s) or in rater scores (d_2 s) be meaningful, which may not be the case when sampling is not representative of the population of interest. The decision remains whether to base the effect size on the variability in rater scores or in rater scores. The suitability of d_1 and d_2 may vary across domains, and it is not reasonable to prefer one of these effect sizes over the other under all circumstances. Nonetheless, for two reasons, it will be more appropriate in most cases to base the analysis on rater scores. First, although the meaning of rater-score differences may be ambiguous, for reasons already offered, it usually is not. Conversely, the meaning of rater-score differences is often less clear because these differences may reflect differences in perception, differences in scale use, or both (see *The Meaning of d_2 for Approach 2* above). Second, in many domains, people may be more familiar with differences between ratees (on which d_1 is based) than with differences between raters judging ratees (on which d_2 is based). To take facial attractiveness as an example, everyone has abundant opportunities to observe that faces differ in attractiveness; however, we have many fewer opportunities to observe how people differ in their general esteem for faces, and we can hardly observe at all how people differ in the way they transform internal judgments into a response. For these reasons, readers will probably have a more intuitive understanding of an effect size if it is based on between-rater variability and not on between-rater variability.

What Should Meta-Analysts Do?

A meta-analyst dealing with a question that is typically studied with a Rater \times Ratee design can easily end up with a set of effects containing the four types of d we have described, each having key conceptual and analytic differ-

ences. How should meta-analysts deal with this diversity? First, it might be possible to reduce the diversity. The d values obtained from repeated measures studies (i.e., d_3 and d_4) can be transformed into analogues for their independent-group counterparts (i.e., d_1 and d_2 , respectively) so long as the study reports separate standard deviations or the correlation between face scores (for d_3) or rater scores (for d_4 ; see Morris & DeShon, 2002, for a detailed account). Unfortunately, meta-analysts will not always have this opportunity. In our set of studies on female–male facial attractiveness differences, for instance, Approach 4 was used 23 times, but the correlation of the pairs of rater scores was never reported. Alternatively, it is possible to compute an independent-group effect size (i.e., d_1 or d_2) if the respective means and standard deviations are reported. However, this opportunity appears to occur rarely (in our sample, 5 times out of 23). One way to work around this is to focus only on p values, which can be mechanically combined notwithstanding differences in experimental design or type of analysis. However, the result of such an analysis is not very meaningful (Becker, 1994). Another approach might be to make a reasonable estimate of the face-score or rater-score correlation as needed and to impute that value, then proceed using Morris and Deshon's (2002) approach. A sensitivity analysis could try out a range of reasonable estimates of this correlation to see how much the d values would change.

In the end, the most viable way to deal with diverse d values may be to meta-analyze them separately by the approach used in the studies. Below, we distinguish three potential goals of meta-analysis.

Estimate the "True" Strength of an Effect

When the meta-analyst is simply interested in the best estimate of the size of some effect or treatment, it will often be best to base the estimate on d_1 . This is because d_1 , for reasons already given, is in most cases the least ambiguous effect size. Morris and DeShon (2002) suggested the use of repeated measures d values (d_3 and d_4 , in our notation) if the focus of the research is on individual change and not on differences across alternate treatments. They argued that "effectiveness could be defined as the amount of change produced as a result of training, suggesting the change-score metric" (Morris & DeShon, 2002, p. 111). We do not agree with their view entirely, because, as we have illustrated, the d based on change (or difference) scores is strongly affected by the homogeneity of the treatment effect.⁶ Thus, d_3 and d_4 values will lead to the erroneous belief that the effect of some treatment is substantial when, in fact, the effect is only

⁶ One other alternative is to compute Becker's (1988) standardized-mean-change measure for each group. This effect size measures the amount of change for each subsample, standardized using the standard deviation of the pretest (or posttest) scores.

marginal but homogeneous (see Table 7, Alternative 3). We argue that judgments about the practical significance of individual change require that one look at d_1 but not at d_3 and d_4 . Of course, the question of whether some treatment operates homogeneously across the treated individuals has merit, in which case one could compare d_1 with d_3 . The more d_3 exceeds d_1 , the more homogeneous the treatment effect.

Determine Which of Two Effects Is Larger

Now consider the case in which we want to determine which of two or more effects is larger (or largest). For example, we could be interested in whether makeup or smiling have a stronger effect on female facial attractiveness. Again, d_1 is an obvious and suitable statistic for this purpose. Because it adequately and meaningfully reflects the strength of an effect, it can also be used to compare different effects. However, we think that the d_2 metric is also suitable for a comparison of effects. Although the absolute magnitude of d_2 is not very meaningful in itself, a result that one treatment exceeds another in d_2 is straightforward and meaningful. This comparison would be problematic if the variability of the rater scores depended on the type of treatment. However, we see no reason to expect this. Naturally, any comparison of effects will benefit from larger numbers of primary studies being included. Thus, a meta-analysis that is not restricted to d_1 but finds the effect of A superior to the effect of B in separate analyses of both d_1 and d_2 would be more compelling.

How about d_3 and d_4 , then? We feel reluctant to recommend these two indices for comparisons between effects. The reason is that we cannot be sure that the effects compared are similar with respect to their homogeneity across ratees (d_3) or raters (d_4). Take the comparison between the effects of makeup and smiling as an example. If we found that makeup yields a higher d_3 than smiling, it might be premature to conclude that the former is more effective than the latter. It could easily be the case that makeup has a small but consistent effect on all studied faces, whereas the effect of smiling is (on average) strong but very heterogeneous (caused by the fact that some people can smile disarmingly but others cannot). Thus, d_3 would wrongly indicate that makeup is a more potent measure to promote facial attractiveness than smiling. Obviously, the same error can occur in other research domains.

Test a Theoretical Model of the Effects of Interest

Finally, we consider the case in which a meta-analyst seeks to test a theoretical model using a moderator analysis of the effects of interest. For example, consider sex differences underlying variation in facial attractiveness. Drawing on evolutionary theory, Buss and Schmitt (1993) argued that physical attractiveness is more important to male than

to female mate preferences. Consequently, this might suggest that women would have more attractive faces than men, especially at the age at which mate choice occurred in ancestral times. Accordingly, the age of an individual's face could moderate the effect of greater female facial attractiveness. Such a claim is best substantiated when it is based on separate analyses of diverse d types. If several differing sources of information converge, our confidence in the finding is strengthened. In this way, more studies can be included in the synthesis as well.

Which d types are suitable for a moderator analysis of a single effect? We argue that d_1 and d_2 clearly qualify, for the same reasons that make them suitable for comparing different effects. What about the use of d_3 and d_4 in moderator analyses? If we can assume that the homogeneity of the studied effect is the same across different values of the moderator, then d_3 and d_4 are suitable. Let us consider our particular research question of whether the difference in attractiveness between the sexes depends on ratees' age. Here, the use of d_4 is appropriate if the homogeneity of the female-male difference in attractiveness across raters is independent of the value of the moderator. This is the case when the agreement between raters about female-male attractiveness differences is comparable across ratees of different ages. However, whether the assumption of effect homogeneity across moderator values is plausible will always depend on the nature of the effect studied.

What Should Primary Researchers Do?

When we ask which of the four approaches to data analysis is most suitable for individual research studies, we must distinguish between inferential and descriptive statistics. Inferential statistics help researchers to attribute their obtained effects to "real" effects in the population of interest and not to random fluctuations that are due to sampling. Descriptive statistics refer more directly to the size of the obtained effect in the sample. The question for description is not whether the data reflect a real population difference but, rather, how large an effect is obtained in the sample at hand, without asking questions about the nature of the sampling or about generalizability to a larger population.

When it comes to inferential statistics, authors usually want to argue that an effect of some size obtained in the sample (e.g., a difference or a correlation) also exists in the population. Sometimes they wish to argue the converse: Some effect did not appear in their data because it does not exist in a population, not because it was missed owing to small sample size or random fluctuation. Both arguments are more convincing when the applied statistical test is of high power (i.e., there is a "large enough" sample size). In this respect, Approaches 3 and 4 are generally superior to Approaches 1 and 2, their independent-samples counter-

parts. Consequently, researchers should consider Approaches 3 or 4 if a within-rater (or within-ratee) effect can represent the phenomenon of interest. Is one of these analyses more suitable, then? When the focus is on differences between ratees (in our examples, the faces), as we have assumed throughout this article, Approach 3 may be more useful, because it seeks to generalize across ratees. The research question of interest helps the researcher decide whether it is more important to generalize across ratees or across raters. However, if generalization across ratees seems more important, that does not imply that generalization across raters does not matter! Therefore, it may be useful to test across raters as well, as we did in Study 2. Interested readers should consult the generalizability theory literature, which thoroughly discusses generalization across all facets in complex designs (e.g., Brennan, 2001; Shavelson & Webb, 1991).

As mentioned in the introduction, null-hypothesis significance testing provides only limited information. For a substantive and practical description and interpretation of results, an effect-size analysis is more suitable. The results of Approaches 3 and 4 allow computation of d_3 and d_4 . However, for the reasons given, these cannot inform us sufficiently about the strength of the effect. We think that in most cases, d_1 will be the most appropriate effect size. This may even be true if the question of interest concerns differences between raters. Consider as an example the question of whether alcohol consumption increases the attractiveness of faces of the opposite sex (Jones, Jones, Thomas, & Piper, 2003). Even in this case, we think it is most adequate to base the effect size on ratee variability and not on rater variability, for the reasons already discussed. First, the meaning of the former is less ambiguous than the meaning of the latter. Second, readers likely have a more intuitive understanding of attractiveness differences between faces than of differences regarding the evaluation of faces by others. But independent of whether d_1 or d_2 is chosen, a representative sampling of ratees (d_1) or raters (d_2) is indispensable to arriving at a meaningful effect size. Only then can d_1 or d_2 measure up to the two promises of effect sizes—to allow the comparison of the strength of effect across studies (and even across domains; e.g., Anderson, Lindsay, & Bushman, 1999) and to assist decision makers effectively.

References

References marked with an asterisk indicate studies included in the review of facial-attractiveness studies.

Alicke, M. D., Smith, R. H., & Klotz, M. L. (1986). Judgments of physical attractiveness: The role of faces and bodies. *Personality and Social Psychology Bulletin*, *12*, 381–389.

*Alley, T. R. (1993). The developmental stability of facial attractiveness: New longitudinal data and a review. *Merrill-Palmer Quarterly*, *39*, 265–278.

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999).

Research in the laboratory: Truth or triviality? *Current Directions in Psychological Science*, *8*, 3–9.

Barnes, M. L., & Rosenthal, R. (1985). Interpersonal effects of experimenter attractiveness, attire, and gender. *Journal of Personality and Social Psychology*, *48*, 435–446.

Bassili, J. N. (1981). The attractiveness stereotype: Goodness or glamour? *Basic and Applied Social Psychology*, *2*, 235–252.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278.

Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

*Berman, P. W., O’Nan, B. A., & Floyd, W. (1981). The double standard of aging and the social situation: Judgments of attractiveness of the middle-aged woman. *Sex Roles*, *7*, 87–96.

*Bernstein, I. H., Lin, T.-D., & McClellan, P. (1982). Cross- vs. within-racial judgments of attractiveness. *Perception & Psychophysics*, *32*, 495–503.

*Berry, D. S. (1991). Attractive faces are not all created equal: Joint effects of babyishness and attractiveness on social perception. *Personality and Social Psychology Bulletin*, *17*, 523–531.

*Berry, D. S., & Landry, J. C. (1997). Facial maturity and daily social interaction. *Journal of Personality and Social Psychology*, *72*, 570–580.

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*, 406–418.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*, 396–404.

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review*, *100*, 204–232.

Check, E. (2005, March 17). Genetics: The X factor. *Nature*, *434*, 266–267.

*Chen, A. C., German, C., & Zaidel, D. W. (1997). Brain asymmetry and facial attractiveness: Facial beauty is not simply in the eye of the beholder. *Neuropsychologia*, *35*, 471–476.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1995). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

*Cross, J. F., & Cross, J. (1971). Age, sex, race, and the perception of facial beauty. *Developmental Psychology*, *5*, 433–439.

Dawes, R. M. (1988). *Rational choice in an uncertain world*. San Diego, CA: Harcourt Brace Jovanovich.

Dipboye, R. L., Fromkin, H. L., & Wiback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant résumés. *Journal of Applied Psychology*, *60*, 39–43.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but . . . : A meta-analytic review

- of research on the physical attractiveness stereotype. *Psychological Bulletin*, 110, 109–128.
- Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin*, 111, 3–22.
- *Edwards, K. (1987). Effects of sex and glasses on attitudes toward intelligence and attractiveness. *Psychological Reports*, 60, 590.
- Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, 62, 61–84.
- *Ferrario, V. F., Sforza, C., Poggio, C. E., Colombo, A., & Tartaglia, G. (1997). The relationship between facial 3-D morphometry and the perception of attractiveness in children. *International Journal of Adult Orthodontic and Orthognathic Surgery*, 12, 145–152.
- *Friedenberg, J. (2001). Lateral feature displacement and perceived facial attractiveness. *Psychological Reports*, 88, 295–305.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- *Hassebrauck, M. (1983). Die Beurteilung der physischen Attraktivität: Konsens unter Urteilern? [The evaluation of physical attractiveness: Consent among raters?]. *Zeitschrift für Sozialpsychologie*, 14, 152–161.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128.
- Hedges, L. V., & Friedman, L. (1992). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research*, 63, 61–84.
- *Henss, R. (1987). Zur Beurteilerübereinstimmung bei der Einschätzung der physischen Attraktivität junger und alter Menschen [Consensus among raters of old people's physical attractiveness]. *Zeitschrift für Sozialpsychologie*, 18, 118–130.
- *Henss, R. (1989). "Schönheit liegt im Auge des Betrachters" (?): Zur Beurteilerübereinstimmung bei der Einschätzung der physischen Attraktivität junger Männer und Frauen ["Beauty lies in the eye of the beholder"?: On the consensus among raters of young men's and women's physical attractiveness]. Saarbrücken, Germany: Universität des Saarlandes.
- *Hildebrandt, K. A., & Fitzgerald, H. E. (1979). Adults' perception of infant sex and cuteness. *Sex Roles*, 5, 471–481.
- *Hume, D. K., & Montgomerie, R. (2001). Facial attractiveness signals different aspects of "quality" in women and men. *Evolution and Human Behavior*, 22, 93–112.
- Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8, 3–7.
- *Johnson, D. F., & Pittenger, J. B. (1984). Attribution, the attractiveness stereotype, and the elderly. *Developmental Psychology*, 20, 1168–1172.
- *Johnston, V. S., & Oliver-Rodriguez, J. C. (1997). Facial beauty and the late positive component of event-related potentials. *Journal of Sex Research*, 34, 188–198.
- *Jones, B. C., Little, A. C., Penton-Voak, I. S., Tiddeman, B. P., Burt, D. M., & Perrett, D. I. (2001). Facial symmetry and judgments of apparent health: Support for a "good genes" explanation of the attractiveness–symmetry relationship. *Evolution and Human Behavior*, 22, 417–429.
- *Jones, B. T., Jones, B. C., Thomas, A. P., & Piper, J. (2003). Alcohol consumption increases attractiveness ratings of opposite-sex faces: A possible third route to risky sex. *Addiction*, 98, 1069–1075.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263–291.
- *Koehler, N., Rhodes, G., & Simmons, L. W. (2002). Are human female preferences for symmetrical male faces enhanced when conception is likely? *Animal Behaviour*, 64, 233–238.
- *Kowner, R., & Ogawa, T. (1995). The role of raters' sex, personality, and appearance in judgments of facial beauty. *Perceptual and Motor Skills*, 81, 339–349.
- Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, 126, 390–423.
- Langlois, J. H., Ritter, J. M., Roggman, L. A., & Vaughn, L. S. (1991). Facial diversity and infant preferences for attractive faces. *Developmental Psychology*, 27, 79–84.
- *Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, 1, 115–121.
- Langlois, J. H., Roggman, L. A., & Rieser-Danner, L. A. (1990). Infants' differential social responses to attractive and unattractive faces. *Developmental Psychology*, 26, 153–159.
- Loftus, G. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5, 161–171.
- *Lundström, A., Woodside, D. G., & Popovich, F. (1987). Panel assessment of facial profile related to mandibular growth direction. *European Journal of Orthodontics*, 9, 271–278.
- *Maret, S. M. (1983). Attractiveness ratings of photographs of Blacks by Cruzans and Americans. *Journal of Psychology*, 115, 113–116.
- *Maret, S. M., & Harling, C. A. (1985). Cross-cultural perceptions of physical attractiveness: Ratings of photographs of Whites by Cruzans and Americans. *Perceptual and Motor Skills*, 60, 163–166.
- Marlowe, C. M., Schneider, S. L., & Nelson, C. E. (1996). Gender and attractiveness biases in hiring decisions: Are most experienced managers less biased? *Journal of Applied Psychology*, 81, 11–21.
- *McClellan, B., & McKelvie, S. J. (1993). Effects of age and gender on perceived facial attractiveness. *Canadian Journal of Behavioural Science*, 25, 135–142.
- *McKelvie, S. J. (1981). Sex differences in memory for faces. *Journal of Psychology*, 107, 109–125.
- *Mealey, L., Bridgstock, R., & Townsend, G. C. (1999). Symmetry and perceived facial attractiveness: A monozygotic co-twin comparison. *Journal of Personality and Social Psychology*, 76, 151–158.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103–115.
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-group designs. *Psychological Methods*, 7, 105–125.
- Murphy, K. R., Herr, B. M., Lockhart, M. C., & Maguire, E. (1986). Evaluating the performance of paper people. *Journal of Applied Psychology*, 71, 654–661.
- *O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: The role of medial

orbitofrontal cortex in facial attractiveness. *Neuropsychologia*, *41*, 147–155.

*O'Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & Cognition*, *26*, 146–160.

*Okkerse, J. M. E., Beemer, F. A., Cordia-de Haan, M., Heineman-de Boer, J. A., Mellenbergh, G. J., & Wolters, W. H. G. (1998). Facial attractiveness and facial impairment ratings in children with craniofacial malformations. *Cleft Palate Craniofacial Journal*, *38*, 386–392.

*Perkins, D. F., & Lerner, R. M. (1995). Single and multiple indicators of physical attractiveness and psychosocial behaviors among young adults. *Journal of Early Adolescence*, *15*, 269–298.

*Pittenger, J. B., Mark, L. S., & Johnson, D. F. (1989). Longitudinal stability of facial attractiveness. *Bulletin of the Psychonomic Society*, *27*, 171–174.

*Pollard, J., Shepherd, J., & Shepherd, J. (1999). Average faces are average faces. *Current Psychology: Developmental, Learning, Personality, Social*, *18*, 98–103.

*Raines, R. S., Hechtman, S. B., & Rosenthal, R. (1990). Physical attractiveness of face and voice: Effects of positivity, dominance, and sex. *Journal of Applied Social Psychology*, *20*, 1558–1578.

Raza, S. M., & Carpenter, B. N. (1987). A model of hiring decisions in real employment interviews. *Journal of Applied Psychology*, *72*, 596–603.

*Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition*, *19*, 57–70.

*Rhodes, G., Proffitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review*, *5*, 659–669.

*Rhodes, G., Sumich, A., & Byatt, G. (1999). Are average facial configurations attractive only because of their symmetry? *Psychological Science*, *10*, 52–58.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, England: Cambridge University Press.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

*Schulman, G. I., & Hoskins, M. (1986). Perceiving the male versus the female face. *Psychology of Women Quarterly*, *10*, 141–154.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93–105.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

*Sparacino, J., & Hansell, S. (1979). Physical attractiveness and academic performance: Beauty is not always talent. *Journal of Personality*, *47*, 449–469.

*Suman, H. C. (1988). Alienation, physical attractiveness and self-perception. *Journal of Psychological Researches*, *32*, 45–51.

*Swaddle, J. P., & Cuthill, I. C. (1995). Asymmetry and human facial attractiveness: Symmetry may not always be beautiful. *Proceedings of the Royal Society of London, Series B*, *261*, 111–116.

*Terry, R. L. (1993). How wearing eyeglasses affects facial recognition. *Current Psychology: Developmental, Learning, Personality, Social*, *12*, 151–162.

*Tobiasen, J. M. (1987). Social judgments of facial deformity. *Cleft Palate Craniofacial Journal*, *24*, 323–327.

*Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, *20*, 291–302.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

*Zebrowitz, L. A., & Montepare, J. M. (1992). Impressions of babyfaced individuals across the life span. *Developmental Psychology*, *28*, 1143–1152.

*Zebrowitz, L. A., Olson, K., & Hoffman, K. (1993). Stability of babyfacedness and attractiveness across the life span. *Journal of Personality and Social Psychology*, *64*, 453–466.

*Zuckerman, M., & Driver, R. E. (1989). What sounds beautiful is good: The vocal attractiveness stereotype. *Journal of Nonverbal Behavior*, *13*, 67–82.

*Zuckerman, M., Hodgins, H., & Miyake, K. (1990). The vocal attractiveness stereotype: Replication and elaboration. *Journal of Nonverbal Behavior*, *14*, 97–112.

Received September 22, 2004

Revision received August 8, 2005

Accepted September 12, 2005 ■

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://watson.apa.org/notify/> and you will be notified by e-mail when issues of interest to you become available!