Valid Theory-Testing Meta-Analyses Further Question the Negative State Relief Model of Helping

Norman Miller and Michael Carlson
University of Southern California

Cialdini and Fultz (1990) questioned the validity of our method (Carlson & Miller, 1987) and objected to three of our reported tests of their negative state relief (NSR) model of mood-induced increments to helpfulness. In response, we presented substantial evidence that the use of judges to define variables is a common tool in psychology and, when used within meta-analyses, consistently meets the relevant criteria of convergent and discriminant trait validity as well as construct validity. In addition to the 17 nonsupportive tests of the NSR model that we initially reported, multiple new tests based directly on the more objective criteria that Cialdini and Fultz stipulate for defining NSR variables also fail to support their model. New data they presented to challenge the discriminant validity of judges' ratings are shown to be based on both methodologically and conceptually flawed procedures. Finally, they reported a (nonsignificant) positive correlation between sadness-manipulation check effect sizes and helpfulness effect sizes, in support of the NSR model. When we computed this correlation in nine different ways on the basis of more extensive sets of studies, including in some instances new studies published since our original report, it was uniformly close to zero and nonsignificant, thereby supporting our prior meta-analytic outcome.

Although researchers now frequently use meta-analysis to summarize research literatures, we believe that its most important strength is its capacity to test theory. Toward this end, virtually all who use it have moved away from the single mean effect-size summary statement sought by the policy maker. Instead, investigators attempt to assess the extent to which given theoretical variables moderate or mediate variation in the dependent measure of interest. For this purpose, they use the method sections of studies in order to code the studies in terms of their levels on theoretically relevant predictors (e.g., Dillard, Hunter, & Burgoon, 1984). Sometimes the predictor is a noncontinuous categorical distinction that can be objectively determined on the basis of information in the original study (e.g., Isenberg, 1986; Signorella & Jamison, 1986; Steblay, 1987; Wood, 1987). Often, however, an underlying dimension is continuous and requires inferential judgment about subjects' subjective experience. Authors sometimes make these judgments themselves (e.g., Bowers & Clum, 1988; Eagly & Steffen, 1986; Hull & Bond, 1986; Mullen et al., 1985; Steele & Southwick, 1985), but often they use additional judges (e.g., Eagly & Carli, 1981; Eagly & Crowley, 1986; Eagly & Steffen, 1986; Steele & Southwick, 1985). In some instances, authors judge a new variable not previously noted (e.g., Bowers & Clum, 1988; Hull & Bond, 1986; Mullen et al., 1985; Steele & Southwick, 1985).

To assess several explanations of why negative moods increase helpfulness relative to neutral moods, we used such judgment procedures (Carlson & Miller, 1987). The judges read portions of the individual method sections of the experimental literature in order to make continuous ratings of the contextual levels of theoretically relevant variables. Additionally, we sought greater analytical power by using partial correlation to control statistically for the effects of other variables when assessing the contribution of each single variable. Our application of these procedures consistently supported two models of mood and helping behavior, but produced no support for the negative state relief (NSR) model (e.g., Cialdini, Darby, & Vincent, 1973). Although Cialdini and Fultz (1990) have accepted meta-analysis as a useful tool for summarizing research, they object to our own use of judges in conjunction with it, stating that this is a "novel methodology." Both of these procedures—the use of judges and partial correlation analysis—are, however, not new to psychology, and a sequential use of them does not markedly increase their uniqueness. Because judgments were elicited on more than a dozen variables, however, an extensive amount of new data was generated in addition to the effect-size estimates of helpfulness computed from the individual studies. Our use of the term meta-analysis to describe our research was intended to reflect this fact: that the quantity of new data brought to bear on the issues exceeded that of a standard literature-summary meta-analysis by a factor greater than 10:1.

In addition to questioning the general validity of judges' ratings, Cialdini and Fultz (1990) also challenged our repeated failures to confirm any aspect of the NSR model. In contesting our application of meta-analytic procedures to the NSR model, they rejected three of our tests of it, claiming that they are inappropriate. To support their contention that judges generate invalid information and that consequently our analyses should
not be viewed as disconfirming the NSR model, they invoke five subissues: (a) the divergent perspective effect seen in actor–observer attributions; (b) the literature on the replication of social psychological experiments by use of role-playing procedures; (c) our use of graduate versus undergraduate students as judges; and (d) their own computation of the correlation between sadness (as reported by subjects in experimental studies) and helpfulness.

Although their article appears to be a powerful indictment of our work, closer examination reveals its serious flaws. With respect to the general criticism regarding the use of judges to generate data, we argue that when held against the only reasonable scientific criteria for assessing the validity of judges' ratings, interpersonal agreement, and construct validity (Funder, 1987; Kenny & Albright, 1987), the use of judges is a valid defining operation and one that has been widely used and accepted in past research. Regarding our tests of the NSR model, we propose that Cialdini and Fultz's (1990) depiction of our work is incomplete and selective in that it omits discussion of numerous tests that incorporated the conceptual features they have proposed as necessary in a proper test. A number of new tests of their model, including those they specifically mandate in their criticism of our work, continue to highlight the essential inadequacy of the NSR model. Finally, with respect to other, more specific criticisms, we argue that the new data they reported are conceptually and methodologically flawed and inappropriate for their intended use; Cialdini and Fultz's invocation of the actor–observer and role-playing literatures reflects important misunderstandings of their nature; and our use of graduate-student as opposed to undergraduate-student judges does not explain our failure to find support for the NSR model.

Use of Judges to Generate Meta-Analytic Data

The use of judges to define variables in psychological research is so extensive that it hardly needs documentation. The entire field of decision analysis assumes a basic validity in human judgment (von Winterfeldt & Edwards, 1986). Among noteworthy citations, Block's (1971) application of judgment procedures to the longitudinal personality data of the Institute of Human Development files is particularly instructive, transcending divergence both in the specific dimensions studied across waves of measurement and the specific instruments used to assess those dimensions that were common across measurement waves by the simple expedient of having judges use Q sorts to transform individual measures into a common metric on a uniform set of dimensions. The extensive work of Rosenthal and his colleagues on nonverbal communication provides another well-implemented example (e.g., Rosenthal, 1987; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979; Scherer, Scherer, Hall, & Rosenthal, 1977). Other instances of the successful use of judges can be found in the prediction of behavior in industrial and organizational settings (e.g., Holland, 1965; Mayfield, 1972; Waters & Waters, 1970).

Within the context of meta-analysis, too, a striking feature is the substantial construct validity seen in judges' ratings. In an appendix concerned specifically with this issue, we presented evidence of criterion-related validity as well as construct validity (Carlson & Miller, 1987). With the exception of the variables of the NSR model, virtually all of the theoretically relevant constructs defined by our judges' ratings were significantly related to helpfulness in the expected, theoretically meaningful way (e.g., guilt, self-absorption, objective self-awareness, responsibility for the mood-lowering event, target of the negative event, legitimacy–urgency and salience of the helping request, and anger and frustration). This basic outcome was replicated in a further study concerned with the effects of positive mood on helpfulness (Carlson, Charlin, & Miller, 1988). Table 1 lists variables theoretically related to mood and helpfulness (other than the NSR variables in the case of negative mood, which in each instance were unrelated to helping), gives a citation for each variable that experimentally confirms its theoretical importance, and presents the correlation between judges' ratings of the variables and helpfulness effect sizes (evidence that the theoretical variable, as defined by judges' ratings, affects mood-biased helpfulness in a theoretically predicted manner). Note that guilt and self as target of the mood-inducing event interact with type of mood as theoretically expected and thus evidence the additional feature of discriminant construct validity.

Table 2 presents parallel evidence for ratings elicited by others. Within it, the majority of judgmentally defined variables (approximately 85%) show construct validity. In their critique, Cialdini and Fultz (1990) did not account for how such substantial convergent evidence for the construct validity of judges' ratings in meta-analytic studies could emerge if judges cannot validly rate variables.

It might be argued that although judges (persons) have normative or reliable views of the world, such views nevertheless reflect erroneous assumptions about the relation between moods and behavior. If so, ratings of mood should show consistent relations with judgments about expected behavior, but these relations should be incorrect. However, such correlations do not conceptually parallel our procedure, which instead shows that judgments of subjects' subjective experiences are related to the actual behavior of those subjects. The correlations presented in Table 1 confirm prior experimental work and, taken together, consistently support the basic validity of our judgment procedure. They provide an integration that exceeds that of the individual studies. Moreover, such validity could not have been obtained had the ratings lacked reliability. To account for this lack of support for their model, NSR theorists must argue that key NSR variables, sadness and distastefulness of the helping act, are a unique subset about which judges make useless ratings.²

---

1 At an earlier point, Cialdini (personal communication, 1984) raised the argument that lower reliabilities for the NSR model's variables (compared with those of other models) may have contributed to its lack of confirmation. Averaged, however, the reliabilities of the NSR variables match or exceed those of other tested models that were supported by our procedure.

2 Arguing from analogy, Cialdini (personal communication, 1984) implied that there might be differential validity for judgments on emotional dimensions as opposed to cognitive dimensions. In his view, the variables of the other models we tested are primarily cognitive, whereas sadness is emotional. He then evoked this distinction to account for
Construct Validity: Correlations Between Judgmentally Rated Theoretical Variables and Helping Effect-Size Estimates Within Two Meta-Analytic Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Variable</th>
<th>Example of experiment providing theoretical prediction</th>
<th>Predicted relation</th>
<th>Association with helping effect sizes(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative mood and helping</td>
<td>OSA (\times) salience of request</td>
<td>Rogers, Miller, Mayer, &amp; Duval (1982)</td>
<td>Positive</td>
<td>(0.50, p &lt; .001)</td>
</tr>
<tr>
<td>(Carlson &amp; Miller, 1987)</td>
<td>Responsibility for causing negative event</td>
<td>Gibbons &amp; Wicklund (1982)</td>
<td>Positive</td>
<td>(0.41, p &lt; .001)</td>
</tr>
<tr>
<td></td>
<td>Guilt</td>
<td>Rogers et al. (1982)</td>
<td>Positive</td>
<td>(0.50, p &lt; .001)</td>
</tr>
<tr>
<td></td>
<td>Other (vs. self) as target of negative event</td>
<td>Carlsmit &amp; Gross (1969); McMillen &amp; Austin (1971)</td>
<td>Positive</td>
<td>(0.30, p &lt; .01)</td>
</tr>
<tr>
<td></td>
<td>Self-absorption (given a nonsalient request for help)</td>
<td>Thompson, Cowan, &amp; Rosenhan (1980)</td>
<td>Positive</td>
<td>(0.27, p &lt; .05)</td>
</tr>
<tr>
<td>Positive mood and helping</td>
<td>Pleasantness of the helping task</td>
<td>Isen &amp; Simmonds (1978); Forest, Clark, Mills, &amp; Isen (1979)</td>
<td>Positive</td>
<td>(0.34, p &lt; .015)</td>
</tr>
<tr>
<td>(Carlson, Charlin, &amp; Miller, 1988)</td>
<td>Extent to which positive event enhances estimate of human nature or kindness</td>
<td>Holloway, Tucker, &amp; Hornstein (1977); Hornstein, Lakind, Frankel, &amp; Manne (1975)</td>
<td>Positive</td>
<td>(0.31, p &lt; .015)</td>
</tr>
<tr>
<td></td>
<td>Oneself (vs. another) as target of positive event</td>
<td>Rosenhan, Salowe, &amp; Hargis (1981)</td>
<td>Positive</td>
<td>(0.49, p &lt; .001)</td>
</tr>
<tr>
<td></td>
<td>Guilt</td>
<td>Cunningham, Steinberg, &amp; Grev (1980)</td>
<td>Negative</td>
<td>(-0.44, p &lt; .001)</td>
</tr>
</tbody>
</table>

\(^a\) Partial correlation (controlling other theoretical and control variables) of rated theoretical variable with helping effect sizes.

Problems in Deriving NSR Predictions

Cialdini and Fultz (1990) objected to the manner in which we tested the major variables of the NSR model (viz., subjects’ age, distastefulness of the helping act, and sadness). This objection centers in the first two instances on the proper content domain of the model, namely, whether it refers solely to the mood of sadness or to other negative moods as well. We confess our own confusion on this issue, but attribute it to the key presentations of the model. On the one hand, NSR theorists have stated repeatedly that their theory refers specifically to the single negative affect of sadness (e.g., Cialdini, Baumann, & Kenrick, 1981; Cialdini et al., 1973; Cialdini & Kenrick, 1976; Cialdini et al., 1987; Manucia, Baumann, & Cialdini, 1984). In stark contrast, but in harmony with the more global title negative state relief, the model is said to apply to a “general negative affective state” throughout a 1973 review (Cialdini et al., 1973, pp. 503, 505, 512, and 513). Moreover, numerous outcomes concerned with negative moods other than sadness are cited (Cialdini & Kenrick, 1976) as producing increments in helpfulness that are explainable by the NSR model, for example, embarrassment (Apsler, 1975), feelings of deviancy (Filter & Gross, 1975), guilt (Freedman, Wallington, & Bless, 1976), perpetration of another’s harm (McMillen, 1971), and cognitive dissonance (Kidd & Berkowitz, 1976).³

There are two possible reasons why these studies might be cited in support of the model despite the fact that they clearly do not contain manipulations of sadness. One is that although the specific emotion of sadness may have particular importance, the model does in fact apply to an array of other negative emotions as well. Alternatively, outcomes of negative mood manipulations other than sadness may have been cited as supporting the NSR model in the belief that sadness was a common, underlying conceptual component of these other diverse manipulations. In either case, if the studies that contain manipulations of the specific negative moods that the model claims will disrupt sadness-based helpfulness (e.g., frustration and anger) are excluded, a test of the model within the broad literature should be supportive. It is inconsistent, however, to contend that a broad set of studies supports a model and then reject a test of it that not only includes those studies but also, in further fairness why other models received support, whereas the NSR model did not. Although differential validity in the judgments of emotions versus cognitions is possible and worthy of further research, this cannot account for the negative outcomes for the NSR model. Two of its three major variables, subjects’ age and distastefulness of the helping task, are not primarily emotional. Age can be directly coded from studies. Distastefulness requires subjects (and judges) to make a cost-benefit assessment—a cognitive task. On the other hand, many other judged variables that enter into theoretical models for which we provide convergent, discriminant, and construct validity are clearly emotional, for example, anger and frustration, positive affect, and guilt.

³ Other citations of support for the model are also confusing. It is said that the Berkowitz and Connor (1966) studies, which failed to support the NSR model, do not appropriately apply to it because their manipulations featured failure (Cialdini & Kenrick, 1976, p. 912). Yet Weyant (1978) is cited as supporting the model’s prediction of an interaction between negative mood, task pleasantness, and helping despite the fact that this study manipulated mood via failure.
Construct Validity of Judged Variables in Psychological Meta-Analyses

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Judged by</th>
<th>Judged variable</th>
<th>Dependent variable</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowers &amp; Clum (1988)</td>
<td>Authors</td>
<td>Credibility of placebo control</td>
<td>Specific and nonspecific treatment effects in behavior therapy</td>
<td>( \delta = 0.35^a )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (high)</td>
<td></td>
<td>( \delta = 0.66 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 (medium)</td>
<td></td>
<td>( \delta = 0.65 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 (medium)</td>
<td></td>
<td>( \delta = 0.34 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 (medium)</td>
<td></td>
<td>( \delta = 0.66 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 (medium)</td>
<td></td>
<td>( \delta = 0.21 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 (medium)</td>
<td></td>
<td>( \delta = 0.45 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 (low)</td>
<td></td>
<td>( \delta = 0.00 )</td>
</tr>
<tr>
<td>Eagly &amp; Carli (1981)</td>
<td>Others</td>
<td>Masculine topic</td>
<td>Sex differences in susceptibility to persuasive influence</td>
<td>( r = .31, p &lt; .01 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interest in topic (Male-female)</td>
<td>( r = .04, ns )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Knowledge of topic (Male-female)</td>
<td>( r = .00 )</td>
<td></td>
</tr>
<tr>
<td>Eagly &amp; Crowley (1986)</td>
<td>Others</td>
<td>Competence</td>
<td>Gender and helping (Male-female)</td>
<td>( \beta = 0.58, p &lt; .001 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comfortable</td>
<td></td>
<td>( \beta = 0.27, p &lt; .001 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Danger</td>
<td></td>
<td>( \beta = 0.72, p &lt; .001 )</td>
</tr>
<tr>
<td>Hull &amp; Bond (1986)</td>
<td>Authors</td>
<td>Helping likeliness</td>
<td></td>
<td>( \beta = 0.49, p &lt; .001 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Own behavior</td>
<td></td>
<td>( \beta = 0.32, p &lt; .001 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stereotypic</td>
<td></td>
<td>( \beta = 0.00 )</td>
</tr>
<tr>
<td>Eagly &amp; Steffen (1986)</td>
<td>Authors</td>
<td>Amount of provocation</td>
<td>Gender and aggression (Male-female)</td>
<td>( Q_a = 35.4, p &lt; .001 )</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>Aggressive behavior</td>
<td></td>
<td>( Q_a = 35.4, p &lt; .001 )</td>
</tr>
<tr>
<td>Hull &amp; Bond (1986)</td>
<td>Authors</td>
<td>Quality of the expectancy manipulation: low (0)-high (3)</td>
<td>Consequences of alcohol consumption &amp; expectancy</td>
<td>( r = .07, ns )</td>
</tr>
<tr>
<td>Mullen et al. (1985)</td>
<td>Authors</td>
<td>Generality of the reference population</td>
<td>False consensus effect</td>
<td>( d = .0009, Z = 0.1515 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( p = .05 ) (one-tailed)</td>
</tr>
<tr>
<td>Steele &amp; Southwick (1985)</td>
<td>Authors</td>
<td>Degree of inhibitory conflict</td>
<td>Alcohol &amp; social behavior</td>
<td>( \delta = 0.14^b )</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>Low</td>
<td></td>
<td>( \delta = 1.063 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td></td>
<td>( \delta = 0.19 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( \delta = 0.94 )</td>
</tr>
</tbody>
</table>

Note. This table is based on a nonexhaustive search for meta-analyses that were reported in the Journal of Personality and Social Psychology and Psychological Bulletin from January 1981–July 1988. We thank Sharon R. Gross for preparing this table.

\( ^a \) Neither the comparison of the high versus pooled medium effect sizes nor the pooled medium versus low effect sizes is significant \( (Z = 0.45, p = .33, \text{ and } Z = 0.58, p = .28, \text{ respectively, using Rosenthal's [1984] Equation 4.27 with the modification that } w = n \text{ because } n = 2 \text{ for the low effect size}). \) The authors explain the low effect size in the high credible group as being due to the occurrence of less-powerful treatments within that set of studies.

\( ^b \) Both the author-rated and other-rated comparisons of the high versus low degree of inhibitory conflict are significantly different \( (Z = 2.37, p = .0089, \text{ and } Z = 1.94, p = .0265, \text{ respectively}). \)

Our response to this ambiguity, however, was not merely to present the three tests cited as inappropriate. Instead, we tested the NSR model in a variety of ways: using a full array of relevant negative mood cases in order to maximize power; testing the specific variable sadness as well as the more global variable negative affect; partialing out of sadness ratings the effects of other rated mood states; testing subsets of cases in which manipulations featuring high levels of either anger or frustration were omitted; testing the model within the subset of studies containing only adult subjects; assessing the mean effect size produced by explicit manipulations of sadness among adult subjects; and so forth. In all, we reported at least 17 tests of the NSR model and conducted many others. None of our tests supported the model.

In their discussion of the three tests they dispute, Cialdini and Fultz (1990) presented their own requirements for a proper assessment of the three key NSR variables: age, distastefulness of the helping task, and sadness. For the first two variables, they argued that analyses should be limited to cases containing experimental manipulations of sadness. For the third, they stated that the analysis should use a measure of sadness that is not based on judges' ratings. We have performed new analyses implied by these suggestions. One direct indicator of whether an experimenter manipulated sadness is whether the author explicitly describes the manipulation as one that features sadness or depression. It seems axiomatic that in such studies some degree
of sadness was indeed induced and that on the average its amount would exceed that in studies that manipulated variables other than sadness. This expectation is strongly supported by a mean effect size of approximately 1.5 for the sadness-manipulation checks in the subset of studies that explicitly manipulated and then measured sadness.

Table 3 presents the results of nine new analyses based directly on the criteria stipulated by Cialdini and Fultz (1990). In each instance as well as in other unreported new analyses based on the criteria they stipulate for a proper test, the NSR model fails to receive support. In no case does a p value approach significance for any predictions from the NSR model, and the direction of effect is often opposite to that specified by it.

In summary, with the exception of NSR variables, our results consistently show that judges’ ratings exhibit construct validity and that for a number of variables operationalized by such ratings (other than those of the NSR model) the direction of their correlation with mood-based helpfulness confirms prior theorizing and experimental work. When the new tests that Cialdini and Fultz (1990) stipulated are applied, the NSR model again receives no support. Moreover, this disconfirmation is maintained in analyses that do not even use judges’ ratings to define the comparisons, but instead use objective criteria.

Specific Criticisms

We now discuss some other, more specific issues raised by Cialdini and Fultz (1990).

Actor–Observer Differences

The first component of the more specific criticisms regarding the validity of judges’ ratings raises the divergent perspectives effect seen in the comparison of actor and observer ratings. To assess the relevance of this literature, it is important to reiterate that our judges’ task was to indicate the relative level of a given variable across a set of studies. The literature on the divergent perspectives of actors and observers is instead largely concerned with mean differences between their ratings on a common scale, not with an assessment of the correlation between their relative ordering of an array of conditions. In contrast to instances of divergent perspective effects that can be cited from individual studies, when the causal attribution literature is examined meta-analytically (Harrington & Miller, 1988), there is a strong between-studies positive correlation between actor and observer judgments (r = .68). As Funder (1987) cogently noted, the literature on bias as a function of perspective can only address the process of judgment, not its accuracy. Nevertheless, despite the disclaimers of some of the researchers involved in it (e.g., Tversky & Kahneman, 1983), many psychologists erroneously view this literature as challenging the correctness or documenting the ineptness of human judgment. Accuracy, however, cannot be inferred from bias (Kenny & Albright, 1987). It is thus mistaken to focus on processes that produce bias and, having documented them, to assume that human judgment is poor and does not improve much on chance. Lately, although acknowledging human susceptibility to bias, there is an emerging recognition that the accuracy of observers can nevertheless be quite high (Hastie & Raisinski, 1988; Kruglanski & Ajzen, 1983).

Role-Playing Replications of Social Psychology Experiments

Cialdini and Fultz (1990) discussed the literature on role playing as a second source of evidence against our reported outcomes. We argue that this literature, too, has little bearing on the validity of the ratings asked of our judges. Whereas our judges were asked merely to assess the degree to which the average subject would experience sadness, anger, and a number of other emotions and cognitions in an array of particular situations, judges in a typical role-playing study are asked a single, but far more complex question; namely, how would the subject behave? Presumably, to be accurate, such role players must first assess how they would feel in a given situation, and then how they would consequently behave. Yet the typical role-playing instructions do not even prime or elicit thought about the mediating cognitive-affective state, but instead simply elicit a behavioral prediction. Moreover, Funder (1987) and McArthur & Barron (1983) persuasively document how the verbal stimuli that constitute the task for such role players (as opposed to more ordinary judgment tasks) deviate from ordinary circumstances.

4 To provide as favorable a test as possible for the NSR model, our original data set (Carlson & Miller, 1987) on negative mood and helping was expanded by treating separate helping task conditions within a study, as well as within-studies age distinctions within the categories of children and adults, as independent observations in the analysis. In our original report, single observations were formed by collapsing over these distinctions in order to achieve a greater degree of statistical independence. This expansion of our original data set produced a total of 98 distinct negative versus neutral mood effect-size cases. In addition, we obtained new ratings of the variable sadness (reliability = .90) and included a new, potentially more sensitive helping-task pleasantness–distastefulness variable (viz., capacity of helping task to alter mood, defined as “the extent to which performing the particular helping tasks would tend to either lower or elevate the mood of the average subject” [scale range: −5, would lessen mood to an extreme extent, to 5, would elevate mood to an extreme extent; reliability = .80]). Thanks are due to Amy Marcus-Newhall and Lisa Maybrown (sociology psychology graduate students at the University of Southern California [USC]) for providing ratings on these two variables. Finally, all of the effect-size estimates were corrected for sample size bias by applying Hedges’ correction formula (Glass, McGraw, & Smith, 1981). None of these changes and additions significantly altered the outcomes originally reported.

5 For example, in a relative sense, they judge how much guilt subjects experienced when listening to noxious music as opposed to being told after they had knocked over a deck of IBM cards that as a consequence Jane will never complete her degree in time for graduation; how distasteful was a helping task that required picking up an armful of books versus answering hundreds of boring multiple choice questions; how sad would one feel when asked to imagine that a very close friend was dying of cancer as opposed to being required to multiply pairs of three-digit numbers mentally while a pneumatic jackhammer was in operation outside a nearby window.
<table>
<thead>
<tr>
<th>NSR variable</th>
<th>Cialdini and Fultz's requirement</th>
<th>Manner tested</th>
<th>NSR prediction</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (children vs. adults)</td>
<td>Test only among cases featuring (a) a sadness manipulation and (b) a private helping opportunity</td>
<td>Using only instances featuring an explicit sadness manipulation (and a private helping opportunity, given children as subjects), a comparison of the average effect-size values of cases featuring children (aged 12 and under) and adults (aged 13 years or older) was undertaken.</td>
<td>Larger effect sizes for adults</td>
<td>Children: $M = -0.14$ ($n = 8$) Adults: $M = 0.05$ ($n = 20$) $t(26) = 0.67$, ns</td>
</tr>
<tr>
<td>Hedonic value of helping task</td>
<td>Test only among sadness cases</td>
<td>1. Given the presence of an explicit sadness manipulation (i.e., as defined by the experimenter), zero-order and partial correlation coefficients between “capacity of helping task to alter mood” and effect size were calculated. 2. Same as 1 above, but analysis was limited to subjects aged 13 years or older. 3. Given the presence of a study cited in one of Cialdini’s own published reviews (Cialdini, Baumann, &amp; Kenrick, 1981) as entailing a manipulation that is capable of producing an NSR effect, and given subjects aged 13 years or older, zero-order and partial correlation coefficients between “capacity of helping task to alter mood” and effect size were calculated.</td>
<td>Positive correlation with effect size</td>
<td>Positive correlation with effect size</td>
</tr>
<tr>
<td>Sadness</td>
<td>Test with a valid indicator of sadness</td>
<td>A dichotomous variable corresponding to whether the experimenter explicitly manipulated sadness (coded value = 1, $n = 36$) or did not (coded value = 0, $n = 62$) was derived. This variable was then correlated with helping effect sizes among subjects aged 13 years or older.</td>
<td>Positive correlation with effect size</td>
<td>Positive correlation with effect size</td>
</tr>
</tbody>
</table>

Because Cialdini has stated that the public versus private dimension is critical only for certain childhood age groups (Cialdini et al., 1981), both public and private helping opportunities were allowed for adult cases. To avoid confounding, instances featuring high levels of guilt (a value above the median) were eliminated from the analysis, as guilt tends to augment helping and is largely absent among studies that used children as subjects. However, additional analytic strategies used on the sadness cases, such as including observations high in guilt, limiting the adult cases to private helping opportunities, or testing the dichotomous age variable of children versus adults in a partial correlation analysis that controls for other variables, all uniformly fail to support the NSR predictions. In conducting partial correlation analyses, the variables guilt, responsibility for the negative event, objective self-awareness, self versus other as target of the negative event, and salience of the helping request (Carlson & Miller, 1987) were statistically adjusted for.

to undermine the normal rules of social inference. Consequently, we view the various failures and successes of the role-playing procedure as having no bearing on the adequacy of our own methodology because the tasks are not comparable. Recent evidence from a validity study confirms this view. Judges' ratings of subjects' affect and cognition correlate well with manipulation checks on subjects' affect and cognition, but judges could not predict subjects' behavior (Miller, Lee, & Carlson, in press).

### Graduate Student Judges

Cialdini and Fultz (1990) criticize our use of graduate student judges, pointing out that they are less similar to the subjects who served in the original studies than are undergraduate judges, and therefore are presumably more prone to divergent perspective effects. We chose graduate students primarily because they tend to be more conscientious and knowledgeable than undergraduates. Our judgment task was of large magnitude, requiring for each of perhaps a dozen variables a separate rating of each of 85 method sections, one by one. This makes...
Data Purportedly Refuting the Discriminant Validity of Judges' Ratings of Sadness

Our original article reported a validity correlation of .47 between judges' ratings of sadness and sadness effect sizes, as computed from those studies that reported a manipulation check administered to the actual subjects (n = 14). Although Cialdini and Fultz (1990) reported data from their own judges that confirm this outcome, they disagree with the empirical evidence for discriminant validity, alleging that such results lack discriminant validity. Their basis for this contention is that other ratings they obtained on global negative affective dimensions such as general emotionality and strength of affect correlate as highly with manipulation checks of experimentally induced sadness as does judged sadness. We feel this argument is faulty for two reasons.

First, in their procedure Cialdini and Fultz (1990) strongly and artificially invited halo effects by instructing their judges to rate each study on all judgment dimensions before proceeding to the next study, rather than having judges rate all studies on each single dimension before proceeding to judgments on the next as did Carlson and Miller (1987). With each judge's rating of sadness clearly visible while rating other dimensions, consistency motives will induce spuriously high correlations among semantically similar dimensions. Second, even if this biased methodology were corrected, in this specific data set one would still fully expect a high correlation between judges' ratings of emotionality and sadness-manipulation check effect sizes for the following reason: The presence of a sadness-manipulation check was the criterion for selecting this subset of studies. Consequently, the preponderance of cases correspondingly featured an explicit manipulation of sadness. Within such a set, sadness is virtually the sole source of variance in emotionality, and therefore, discriminant validity between ratings of it and other variables directly related to the strength of its induction (such as general emotionality) cannot be expected. A proper assessment of the discriminant validity of sadness ratings relative to ratings of general emotionality requires a set of cases in which other negative affects (e.g., anger, jealousy, and boredom) were sometimes strongly present in the absence of sadness. Instead, Cialdini and Fultz assessed discriminant validity by analyzing a selected set of cases within which sadness and general emotionality are largely, by definition, synonymous, and then contended that the obtained high correlations between the two variables casts doubt on judges' ability to differentiate.

Valid Evidence of Discriminant and Convergent Validity

Although our initial report of the correlation between judges' ratings and manipulation check effect sizes of sadness inductions was only intended to provide evidence of convergent validity (i.e., agreement between judges' and subjects' ratings of sadness), further evidence indicates that our judges were able to discriminate sadness from other specific negative affects. For example, among our rated variables, guilt, anger, and frustration (all of which, like sadness, fall within the more general category of negative affect), the average correlation with the 14 sadness-manipulation check effect sizes was —.08. Evidence showing that judges validly differentiate conceptually related constructs such as objective self-awareness and self-absorption (Carlson & Miller, 1987) and anger and frustration (Carlson, 1988) further demonstrates the discriminant validity of their ratings. In the latter case, for example, manipulation check effect-size estimates of the degree of experienced anger (n = 33) correlated positively with judges' ratings of felt anger (r = .44, p < .01), but were unrelated to their ratings of frustration (r = -.16).

In addition, we have obtained positive convergent validity correlations between judges' ratings and manipulation check effect sizes of positive affect (Carlson et al., 1988), frustration, and negative affect (Carlson, 1988), as well as for a combination of four variables including objective self-awareness, guilt, self-absorption, and anger (Carlson & Miller, 1987). When these correlations are pooled via the Stouffer method of combining Z scores (Rosenthal, 1984), the null hypothesis of a zero correlation is untenable (Z = 4.32, p < .0001). Moreover, although these consistent positive correlations are substantial (range = .4-.7), they are obtained in the face of factors likely to depress their magnitude: (a) commonly, a restricted range is present on the dimension being assessed because of the fact that a manipulation check is more likely in cases where a high level of a variable was intentionally manipulated; (b) nonidentical manipulation checks, associated with different degrees of sensitivity and
unique method variance, are lumped together in computing the correlations; and (c) self-report and other measures imperfectly assess the construct in question, and may therefore produce systematic bias in certain cases.

**Correlation Between Sadness-Manipulation Check and Helping Effect Sizes**

To support a central theoretical contention of the NSR model, Cialdini and Fultz (1990) argued that, among adult subjects, sadness-manipulation check effect sizes correlate positively with helping effect sizes. We question this contention. First, the reported correlation is nonsignificant—$r(12) = .42$, $p > .05$, one-tailed—and therefore could be due to sampling error. Second, it was obtained only after excluding the Rogers, Miller, Mayer, and Duval (1982) data, whose inclusion would produce a negative correlation between sadness and helping effect sizes, $r(16) = -.17$, $p = .50$.

To assess more adequately the relation between sadness and helping effect sizes, we increased the number of sampled observations by including relevant studies published after 1982, the cutoff date for inclusion in our own meta-analysis (viz., Berkowitz, 1987; O'Malley & Andrews, 1983; Manucia et al., 1984; Shaffer & Smith, 1985). In all, eight new observations were added, producing a total of 26 cases. In harmony with our earlier outcome, the resulting correlation between sadness and helping effect sizes failed to support the NSR model, $r(24) = -.11$, $p = .58$. In sum, when we computed this correlation in at least nine different ways on the basis of more extensive sets of studies, including in some instances new studies published since our original report, it was uniformly close to zero and nonsignificant, thereby supporting our prior meta-analytic outcome.

**Summary and Conclusion**

To contest our application of meta-analysis to the NSR model, Cialdini and Fultz (1990) contended that ratings made by judges are invalid. There are two basic meanings to validity in social judgments: first, agreement among those who are judging the same thing; and second, the ability of those judgments to predict behavior. Campbell and Fiske (1959) referred to the first as *trait validity*, whereas the term *construct validity* (Cronbach & Meehl, 1955) is commonly used in reference to the second. With respect to the first of these two basic criteria, we not only show high interrater agreement among our judges, but more important, substantial evidence of convergence between judges' ratings and the ratings of subjects in their responses to manipulation checks. Moreover, we present ample evidence that judges' ratings evince Campbell and Fiske's companion criterion of discriminant validity. Regarding the second, construct validity, we show very substantial evidence that judges' ratings do predict the behavior of subjects in experiments. This construct validity not only appears consistently in our own work, but also in that of others.

In attempting to strengthen their argument concerning the invalidity of judges' ratings, Cialdini and Fultz (1990) cited nonrelevant literatures on the process of judgmental bias, not error, namely, the literatures on the actor–observer effect and role playing as a means of replicating experimental social psychological research—literatures in which the subjects' tasks deviate in important ways from those of our judges (Funder, 1987; McArthur & Barron, 1983). Moreover, Cialdini and Fultz provided no explanation for how the proposed error in judges' ratings only works against the NSR model, whereas outcomes based on judges' ratings of other variables consistently correspond to those obtained in experiments. In addition, they present no tests of effects for the study that they present to support their criticism that our judges' ratings lacked discriminant validity—a study that we argue is methodologically and conceptually flawed. They cited a nonsignificant correlation to argue that sadness and helping effect sizes are positively related. When we compute this correlation in nine different ways on the basis of less selective data sets, we find it uniformly close to zero. Finally, Cialdini and Fultz raised an objection regarding the outcomes of the 17 nonconfirming tests of the NSR model reported in our article, contending that we did not test it appropriately. Cialdini and Fultz (1990) presented their own specific criteria for proper tests. However, when we use such tests, they produce additional evidence that fails to support the NSR model.

Cialdini and Fultz (1990) argued against analyses in which we statistically controlled for levels on variables not manifestly part of the NSR model by using partial correlation procedures. However, comparisons across studies must control for potential confounding. In the helping literature, this confounding within paradigms is illustrated well in Eisenberg and Miller's (1987) meta-analytical review of empathy and helping. Picture-story measures, which fail to show empathy effects, are commonly used in studies with children, but are rarely used in studies on

---

10 The six studies that provided the 14 cases for the correlation computed by Cialdini and Fultz (1990) are Cunningham, Steinberg, and Grev (1980); Forest, Clark, Mills, and Isen (1979); Fried and Berkowitz (1979); Harvey and Enzle (1981); Kidd and Marshall (1982); and Thompson, Cowan, and Rosenhan (1980). Cialdini and Fultz discarded the Rogers et al. data because it yielded abnormally large sadness effect sizes. Alternate, more acceptable analytic strategies, such as including it and either using a nonparametric correlation method or normalizing the resulting distribution via a transformation, produce nonsignificant, near-zero correlations between sadness and helping effect sizes, for example, Spearman $r(16) = -.11$, $p = .66$; using log-transformed sadness effect sizes, Pearson $r(16) = -.07$, $p = .77$. In addition, the correlation between sadness and helping effect sizes remains nonsignificant if key rated theoretical variables are controlled for, for example, adjusting for the variable guilt, partial $r(15) = -.17$, $p = .52$; adjusting for the variables anger and frustration, partial $r(14) = -.20$, $p = .47$.

11 When included in this larger data set, Rogers et al. (1982) is no longer an outlier. The Z score pertaining to its largest sadness effect size is below 2.0. Nevertheless, within this expanded data set, numerous alternate analytic tactics, such as excluding the study by Rogers et al., $r(20) = .13$, $p = .57$; using a nonparametric correlation procedure, Spearman $r(24) = .05$, $p = .82$; or using log-transformed sadness effect sizes, $r(24) = -.03$, $p = .87$, uniformly failed to produce a significant correlation between experimentally manipulated sadness and subsequent helpfulness. In fairness to the NSR model, the fixed mood condition within Manucia et al. (1984), for which the NSR proponents explicitly predicted an absence of sadness-generated helpfulness, was excluded from all of these analyses.
adults. Therefore, a review of the research on children will for
this reason alone yield a negative outcome (e.g., Underwood & Moore, 1982), whereas other measures that show a positive
relation between empathy and prosocial behavior in adults (when they happen to be used in studies on children) will con-
firm empathy effects in children too. Between-studies compari-
sions of age effects are particularly vulnerable to such interac-
tions. When, for instance, we control for the effects of extrane-
ous variables in our analyses of the effect of age and fail to find
support for the NSR model, such outcomes question the validity
of the explanatory dynamic proposed by the model to ac-
count for mood-based interactions with age—that is, whether
the hypothesized timetable for levels of socialization (with re-
spect to self-reinforcement) does indeed interact with sadness
to carry the explanatory burden (Cialdini & Kenrick, 1976).

Finally, we argue that the theoretical importance of such
meta-analytic outcomes for the interpretation of individual
studies should not be overlooked. These outcomes suggest why,
in the face of broader, more general disconfirmation of the NSR
model, one can nevertheless obtain an outcome within an indi-
vidual study that seemingly supports it. Namely, we suggest that
in such instances researchers have (inadvertently) confounded
manipulations of NSR variables with levels of other (non-NSR)
variables that do in fact explain mood-based helpfulness, such as
objective self-awareness or attentional focus.

A promising theory should not be discarded on the basis of
a single disconfirmation or on the basis of a set of disconfirma-
tions that were all obtained within a single, narrow research
paradigm. With respect to the NSR model, however, the evi-
dence we have presented in this article, in combination with
our prior evidence (Carlson & Miller, 1987), integrates findings
from numerous individual research paradigms. Across an array
of standard procedures for quantitatively integrating individual
experimental findings, the outcomes are consistently noncon-
firming for the NSR model.12 In line with Cook and Campbell’s
(1979) discussion of the Quine-Duhem thesis, it is in principle
possible to explain away each new nonconfirming outcome by
continually advancing new ad hoc hypotheses. At some point,
however, it becomes more appropriate to question the adequacy
of the unconfirmed model, especially when the same research
integration procedures do support alternative models.

12 It is also worth noting that the results of a number of individual
studies, in agreement with our meta-analytic findings, contradict the
NSR model’s insistence on the importance of sadness (Berkowitz, 1987;
Rogers et al., 1982; Shroeder, Dowd, Sibicky, Matthews, & Allen,
1988; Thompson et al., 1980; Underwood et al., 1977) and pleasantness
of the helping task (Cunningham et al., 1980; Forest et al., 1979; Freed-

References
nosticity of cognitive/affective and behavioral data. Journal of Per-
Berkowitz, L., & Connor, W. H. (1966). Success, failure, and social re-
669.
and nonspecific treatment effects: Meta-analysis of placebo-cont-
trolled therapy research. Psychological Bulletin, 103, 315–
323.
validation by the multitrait-multimethod matrix. Psychological Bul-
letin, 56, 81–105.
Carlson, M. (1988). Toward understanding the relationship between un-
pleasant stimulation and aggressive behavior in humans. Unpub-
lished doctoral dissertation, University of Southern California, Los
Angeles.
behavior: A test of six hypotheses. Journal of Personality and Social
Psychology, 55, 211–229.
sadness: A three-step model of the development of altruism as hedo-
Cialdini, R. B., Darby, B. L., & Vincent, J. E. (1973). Transgression and
altruism: A case for hedonism. Journal of Experimental Social
Psychology, 9, 502–516.
Cialdini, R. B., & Fultz, J. (1990). Interpreting the negative mood/help-
ing literature via “mega”-analysis: A contrary view. Psychological
development perspective on the relationship of negative mood state
and helping. Journal of Personality and Social Psychology, 34, 907–
914.
Cialdini, R. B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., & Be-
758.
and analysis issues for field settings. Chicago, IL: Rand McNally.
cal tests. Psychological Bulletin, 52, 283–308.
Cunningham, M. R., Steinberg, J., & Grev, R. (1982). Wanting to and
having to help: Separate motivations for positive mood and guilt-in-
192.
persuasive strategies: Meta-analysis of foot-in-the-door and door-in-
the-face. Human Communication Research, 10, 461–488.
communications as determinants of sex difference in influenceability:
A meta-analysis of social influence studies. Psychological Bulle-
tin, 90, 1–20.
meta-analytic review of the social psychological literature. Psycholog-
ical Bulletin, 100, 283–308.
Eagly, A. H., & Steffen, V. J. (1986). Gender and aggressive behavior: A
meta-analytic review of the social psychological literature. Psycholog-


Received August 19, 1988
Revision received June 1, 1989
Accepted June 16, 1989

Instructions to Authors

Authors should prepare manuscripts according to the Publication Manual of the American Psychological Association (3rd ed.). All manuscripts must include an abstract containing a maximum of 1,000 characters and spaces (which is approximately 100–150 words) typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the Manual. Also, all manuscripts are subject to editing for sexist language.

Authors reporting original data will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. (A copy of the APA Ethical Principles may be obtained from the APA Ethics Office, 1200 17th Street, N.W., Washington, DC 20036.)

Anonymous review will be first an author's option. If anonymous review is not requested in a cover letter, it will become the prerogative of the processing editor. Authors requesting anonymous review are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the authors' names and institutional affiliations, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the authors' identity or affiliations should be on separate pages. Every effort should be made to see that the manuscript itself contains no clues to the authors' identity.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more journals. APA policy also prohibits duplicate publication, that is, publication of a manuscript that has already been published in whole or in substantial part in another journal. Prior and duplicate publication constitutes unethical behavior, and authors have an obligation to consult journal editors if there is any chance or question that the paper might not be suitable for publication in an APA journal. Also, authors of manuscripts submitted to APA journals are expected to have available their raw data throughout the editorial review process and for at least 5 years after the date of publication.

Manuscripts should be submitted in quadruplicate. All copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typeface is acceptable only if it is clear and legible. Authors should keep a copy of the manuscript to guard against loss. Mail manuscripts to the Editor, Mark I. Appelbaum, Psychological Bulletin, Department of Psychology and Human Development, Box 159 Peabody Station, Vanderbilt University, Nashville, TN 37203.