

## COMMENTS

# Tests of Multiplicative Models in Psychology: A Case Study Using the Unified Theory of Implicit Attitudes, Stereotypes, Self-Esteem, and Self-Concept

Hart Blanton

University of North Carolina at Chapel Hill

James Jaccard

University at Albany, State University of New York

Theories that posit multiplicative relationships between variables are common in psychology. A. G. Greenwald et al. (2002) recently presented a theory that explicated relationships between group identification, group attitudes, and self-esteem. Their theory posits a multiplicative relationship between concepts when predicting a criterion variable. Greenwald et al. suggested analytic strategies to test their multiplicative model that researchers might assume are appropriate for testing multiplicative models more generally. The theory and analytic strategies of Greenwald et al. are used as a case study to show the strong measurement assumptions that underlie certain tests of multiplicative models. It is shown that the approach used by Greenwald et al. can lead to declarations of theoretical support when the theory is wrong as well as rejection of the theory when the theory is correct. A simple strategy for testing multiplicative models that makes weaker measurement assumptions than the strategy proposed by Greenwald et al. is suggested and discussed.

*Keywords:* regression, interactions, cross-products, implicit association test

Numerous theories in psychology posit multiplicative relationships between variables. For example, in educational as well as organizational psychology, performance on a task is thought to be a function of ability times motivation (e.g., Anderson & Butzin, 1974; Gupta & Singh, 1981). In cognitive psychology, it has been suggested that decisions are a function of subjective probabilities times utilities (Edwards & Fasolo, 2000). In social psychology, attitudes are thought to be a function of expectancies times values (Ajzen & Fishbein, 1981). In psycholinguistics, the perceived truth value of a compound statement is said to be a multiplicative function of the truth value of the component parts of that statement (e.g., Oden, 1978). Given the presence of multiplicative models in diverse areas of psychology, it is useful to consider the challenges of testing such models, especially in the context of correlational designs that use multiple regression. These tests often are linked to the analysis of interactions. The present article considers the relationship between interaction analysis in multiple regression and the evaluation of models with simple multiplicative relationships.

Greenwald et al. (2002) recently presented a theory of attitudes, stereotypes, self-esteem, and self-concept that explicated dynamic relationships between group identification, group attitudes, and

self-esteem. Using fundamental concepts from connectionistic frameworks in cognitive psychology and consistency theories in social psychology, these authors derived a set of principles that led them to postulate relationships between the traditional psychological constructs of self-esteem (SE), group identity (GI) and attitudes toward a group (AG). Greenwald et al. argued that any one of these constructs should be a simple multiplicative function of the other two. They characterized this multiplicative function as an interaction and tested the model using the following equations:

$$SE = a_1 + b_1(GI)(AG) + e \quad (1)$$

$$GI = a_2 + b_2(SE)(AG) + e \quad (2)$$

$$AG = a_3 + b_3(SE)(GI) + e. \quad (3)$$

Greenwald et al. (2002) argued that the model is supported if regression analyses for the above equations yield strong correlations and statistically significant and positive regression coefficients. They suggested an additional set of statistical analyses for their theory based on augmenting the “interaction” model with the component parts of the product term, using equations of the following form:

$$SE = a_4 + b_4GI + b_5AG + b_6(GI)(AG) + e \quad (4)$$

$$GI = a_5 + b_7SE + b_8AG + b_9(SE)(AG) + e \quad (5)$$

$$AG = a_6 + b_{10}SE + b_{11}GI + b_{12}(SE)(GI) + e. \quad (6)$$

Greenwald et al. viewed the regression coefficients for the component parts as representing the “effects” of the two predictors on the criterion and the coefficient for the product term as represent-

---

Hart Blanton, Department of Psychology, University of North Carolina at Chapel Hill; James Jaccard, Department of Psychology, University at Albany, State University of New York.

James Jaccard is now at the Psychology Department, Florida International University.

Correspondence concerning this article should be addressed to Hart Blanton who is now at the Department of Psychology, 4235 TAMU, Texas A&M University, College Station, TX 77843-4235. E-mail: hblanton@gmail.com

ing the interaction effect for the two predictors (see p. 10). They argued that the two component parts should not account for criterion variability over and above the “interaction” term. Specifically, they proposed the following statistical predictions for their theory beyond those based on Equations 1–3: (a) The coefficient associated with the product term in Equations 4 through 6 should be positive in sign, (b) the coefficients for the component parts of the product terms should not be statistically significant, and (c) a hierarchical regression analysis should show that the addition of the two component parts does not add substantial amounts of explained variance over and above the interaction model that contains only the product term. Coupled with the aforementioned product-term only analysis, these four analyses represent a test collection Greenwald et al. suggested using to evaluate the multiplicative model (see p. 11).

The primary analytic strategy used by Greenwald et al. (2002) was framed in the language of interaction analysis, thereby linking the concept of interaction to the evaluation of multiplicative models. However, the Greenwald et al. procedure is not the method by which interaction effects traditionally are evaluated in psychological research that relies on multiple regression (see Cohen, Cohen, West, & Aiken, 2003; Jaccard & Turrisi, 2002). Researchers might infer that the procedure used by Greenwald et al. is an appropriate strategy for testing multiplicative models more generally. The purpose of the present article is to use the Greenwald et al. procedure as a case study for discussing important issues for testing multiplicative models. Greenwald et al.’s analysis is of substantive importance because it is embedded within a larger research program on the implicit measurement of attitudes that is receiving widespread attention in psychology. Research focusing on implicit attitudes and the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) used by Greenwald et al. has been reported in over 100 journal articles (e.g., Asendorpf, Banse, & Muecke, 2002; de Jong, van den Hout, Rietbroek, & Huijding, 2003; Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Gray, MacCulloch, Smith, Morris, & Snowden, 2003; Phelps, Cannistraci, & Cunningham, 2003; Teachman, Gapinski, Brownell, Rawlins, & Jeyaram, 2003; Wiers, VanWoerden, Smulders, & de Jong, 2002). Their statistical approach to testing a multiplicative model has limitations and the strategy could encourage psychologists to adopt suboptimal practices when analyzing theoretical models that have simple product terms in them or models that include interaction effects. Although we use the Greenwald et al. approach as a springboard for discussion, many of the comments we make are directed at research applications outside of their framework and are designed to flesh out relevant issues for psychological research more generally.

### Multiplicative Models, Regression Coefficients, and Measurement Assumptions

#### *Rational Zero-Points in Multiplicative Models*

At the conceptual level, multiplicative models typically focus on variables that have rational zero-points. These zero-points reflect either a property of complete absence of the variable in question (as is the case for ratio level measures) or points on a bipolar construct that represent indifference, neutrality, or a zero-based midpoint. For example, in the Greenwald et al. (2002) theory, implicit self-esteem is a theoretical construct that reflects an implicit evaluation of self that

ranges from very negative through neutral to very positive. The rational zero-point for the Greenwald et al. theory is that of neutral self-esteem (cf., Karpinski, 2004).

Psychological constructs like self-esteem are not directly observable, and an individual’s standing on such constructs must be inferred from overt responses. Typically, these overt responses are answers to questions on a survey, responses on a rating scale, or performance on a task in a laboratory. Based on these overt behaviors, individuals are assigned numbers that are thought to reflect their standing on the underlying dimension. These numbers are, of course, measures. A key question is how the numbers assigned to individuals map onto their standing on the true unobserved psychological construct.

For tests of multiplicative models, a crucial issue is the identification of the number on the *observed* measure that corresponds to the *true* rational zero-point on the underlying variable of interest. Unless this number can be isolated and the metric rescaled so that observed zero is at the true rational zero, tests of multiplicative models are not straightforward when strategies such as those used by Greenwald et al. (2002) are pursued. As shown later, these analytic strategies assume that the observed score of zero maps perfectly onto the theoretical zero-point. Psychologists typically are reluctant to make such strong assumptions, given the arbitrary metrics that they so often rely upon.

To illustrate why the assumption of rational zeros is crucial, consider one of the analyses that Greenwald et al. (2002) advocate as part of their strategy. This is the regression analysis where one predicts the criterion from just the product term while excluding the component parts of the product term (see Equations 1–3). The strategy of testing interaction effects while omitting the component parts of the product term has been criticized by numerous methodologists (e.g., Cohen, 1978; Cronbach, 1987), and the relevant issues are germane to the Greenwald et al. test as well. One problem with the approach is that the correlation between the product term and the criterion is influenced by the origin of the measures that serve as indicators of the predictor variables. This means that additive transformations of the measures impact the fit of the model. For example, one can obtain different results for the theory test if one chooses to score a predictor variable that has 11 categories from 0 to 10, or from 5 to 15, or from –5 to +5. Stated more generally, if  $X$  and  $Z$  are predictors, and  $Y$  is a criterion, then transformations of the form  $X + c$  and  $Z + k$  can affect model fit, where  $c$  and  $k$  are arbitrarily defined constants. Irwin and McClelland (2002) describe an example in which a correlation between a product term and a criterion that was 0.97 when analyzed with one set of origins changed to 0.00 when analyzed with another set of origins based on subtracting a constant from one of the two variables in the product term. Theory tests that are influenced by such trivial transformations are questionable, given the arbitrary nature of so many psychological measures. To apply this strategy, one must have confidence that the origins of the measures map onto the true rational zero-points of the underlying constructs. If the observed zero-point in the measured value does not map directly onto the true zero-point of the theoretical value, then parameter estimates can be biased.

Greenwald et al. (2002) seemed to adopt the logic that the simple act of differencing two unipolar measures creates a set of numbers where the observed 0 reflects the true zero-point for their bipolar theoretical constructs. For example, Greenwald et al. had respondents rate themselves on six pleasant meaning words and six unpleasant meaning words on 7-point scales with anchors 1 = *not*

at all characteristic of you and 7 = extremely characteristic of you. They averaged the responses to the positive items and also averaged the responses to the negative items and then subtracted the latter from the former. According to Greenwald et al., this measure had a “rational zero” because it was a difference score (p. 12). The implication is that positive scores on the metric reflect positive self-regard, negative scores reflect negative self-regard, and a score of zero reflects neutral self-regard. However, this may not be the case. Consider the case where the positive items are more positive in nature than the negative items are negative. Following Anderson (1981), suppose that each item has a true scale value that reflects its degree of positivity or negativity on the underlying dimension of self worth. If the average absolute scale value of the positive items is larger than the average absolute scale value of the negative items, then someone who equally endorses the positive items and the negative items might actually have positive self regard rather than neutral self regard. Despite the difference scoring, the zero-point obtained by taking a difference of two measures does not map onto the true zero-point because the scale values of the two sets of items are not equally polarized.

Even when the same single-item rating scale is used to make two judgments, it is not assured that a “rational zero” results from differencing. Consider a respondent who is asked to rate the height of different target people on a scale from 0 (*very short*) to 10 (*very tall*). If we average the ratings made for male targets and subtract this from the average ratings for female targets, it might be found that the “relative evaluation” for the two target groups on the rating scale is zero. It would be a mistake to infer from this that the person considers men and women to be, on average, of equal height. As Biernat, Manis, and Nelson (1991) have shown, a different psychometric standard links “true height” to how we use the rating scales for men as opposed to women. A 6-ft man and a 5-ft, 7-in. woman might both be given a rating of “somewhat tall” because people frame their ratings within gender. The psychometric dynamics of mapping observed zero-points onto rationale zeros for an underlying dimensions are complex, involving how people make a judgment in response to a question, the cognitive processes and strategies that individuals use to translate that judgment onto a rating scale and the extent to which the same response translation processes are invoked by all individuals. The above examples suggest that one cannot be certain that difference scores capture true relational zeros. Assertions that they do require assumptions about underlying scale values and how, for example, people frame responses to the scales being used.

In addition to rating scales, Greenwald et al. (2002) used implicit measures of self-esteem based on the IAT. The IAT index is based on a difference between two response latencies, one that reflects how long it takes a participant to classify stimuli shown on a computer screen into the categories “Self or Positive” versus “Other and Negative” and the other that reflects how long it takes a participant to classify stimuli into the categories “Self or Negative” versus “Other and Positive” (see Greenwald et al., 2002, for details). The guiding assumption is that a difference score of zero between the two latencies reflects neutral self-esteem. Some might think that because the observed metrics represent a time dimension, the metrics must have rational zeros. This is not the case. A metric such as milliseconds may indeed have a rational zero as an indicator of the dimension of time, but such properties do not automatically transfer to how such numbers map onto a completely different dimension, namely self-esteem. Even

something as simple as a count (e.g., the number of times a child hits another child on a playground) can have an arbitrary origin when it is mapped onto an underlying psychological dimension like “aggressive tendencies.”

Features of the IAT make it unusually bold to assume that the observed measure score of 0 reflects the true rational zero-point. For example, for self-esteem, the response latencies in the IAT measure supposedly reflect four association strengths: (a) the strength of the association between “self” and “pleasant” (SP), (b) the strength of the association between “self” and “unpleasant” (SU), (c) the strength of the association between “others” and “pleasant” (OP), and (d) the strength of the association between “others” and “unpleasant” (OU). One of the IAT tasks (sometimes called the *compatible* task) measures a response latency for categorizations involving the SP and OU links considered simultaneously (see Figure 1 of Greenwald et al., 2002). The other task (sometimes called the *incompatible* task) measures a response latency for categorizations involving the SU and OP links considered simultaneously. For either task, the precise function relating the two relevant association strengths to the response latency is unknown. We do not know, for example, how the association strengths for SP and OU affect the single response latency that is said to reflect the two links. We do not know whether the response latency is reflective of a simple average of the two association strengths, a weighted average, a sum, or an interaction of the two. We do not know whether the same rule applies to the compatible and the incompatible tasks. Additional ambiguities exist because a number of studies have suggested possible confounds and artifacts in the IAT (Karpinski, 2004; Karpinski & Hilton, 2001; McFarland & Crouch, 2002; Olson & Fazio, 2003; Rothermund & Wentura, 2004). It is beyond the scope of this article to evaluate the existing literature on this relatively new measure, but we think it is fair to say that any claims of lack of bias or the presence of meaningful zeros are premature.

It is important to note that Greenwald et al. (2002) explicitly recognized that metric assumptions come into play in their tests, stressing that the scale used to reflect a given construct must have a “rational zero-point” (p. 11). Despite this, Greenwald et al. are equivocal about the properties of the measures they analyzed. Sometimes they asserted that a score of zero on the measure reflects a true rational zero by virtue of the fact that it represents a difference score. Other times, when selected results did not conform to theoretical predictions, they suggested that this may be because of metric assumption violations (see, e.g., their comments on pp. 15, 17, and 23). In our opinion, a preferable strategy is one that does not require such strong measurement assumptions and that is less susceptible to such hedging when hypotheses are not confirmed. We describe such a strategy later.

### *Tests of the Product Term Only Model*

We already have identified one limitation of a regression analysis that examines the squared correlation and the significance of the regression coefficient when only the product term is used to predict a criterion (as per Equations 1–3). This limitation is the strong measurement assumption required for the origin of the metrics. There is a second limitation to this analysis even when the scaling assumption is met. This is the fact that the analysis can be insensitive to specification error, that is, the analysis can suggest that a model that does not reflect the true underlying generating process is indeed an appropriate model (Anderson, 1981).

Greenwald et al. (2002) suggest that the coefficients for the product terms in Equations 1 through 3 reflect “interaction effects.” Although product terms do indeed reflect interactions, they do not reflect interactions as traditionally parameterized in the psychological literature. A product term, by and of itself, is a complex amalgamation of main effects and interactions (Cohen, 1978; Cronbach, 1987). It is only when the component parts of the product term are included in the regression equation that the regression coefficient associated with the product term takes on interaction meaning that is typically embraced in psychology. This point has been made repeatedly by methodologists (e.g., Cohen, 1978; Cohen & Cohen, 1983). Because the product term reflects both main effects and interactions (i.e., it is confounded with main effects), it is possible to obtain significant prediction from a product term even when the underlying generating process is strictly additive and not multiplicative.

To illustrate this, we conducted a simple computer simulation where we created a set of population data in which a criterion variable,  $Y$ , was defined as being a simple linear function of two predictors,  $X$  and  $Z$ . For example, the self-esteem ( $Y$ ) of a group of African Americans might be influenced, in part, by how positively they evaluate European Americans relative to African Americans ( $X$ ), with higher levels of self-esteem resulting from embracing the values of the majority culture. Independent of this, their self-esteem also may be influenced by how much they identify with European Americans relative to African Americans ( $Z$ ), with self-esteem being higher as one assimilates to a majority culture. In the simulation,  $X$  and  $Z$  were defined to be normally distributed, each with a mean of 0.20 and a standard deviation of 0.15. These values are not atypical of IAT measures (see Table A1 in the Appendix). The true model in the population was

$$Y = \alpha + 0.50X + 0.50Z + \varepsilon.$$

The error variance was defined to be normally distributed with a mean of zero and of a magnitude that would produce a population multiple correlation of 0.40. The population correlation between  $X$  and  $Z$  was set to equal 0.10. The intercept was set equal to zero. In this model, there is no interaction effect, and the generating mechanisms clearly are at odds with a multiplicative model.

We selected a random sample of 95 cases from the population (based on the sample sizes used by Greenwald et al., 2002) and applied the analysis suggested by Greenwald et al. (2002). Specifically, we regressed the measure of  $Y$  onto the product term between  $X$  and  $Z$  and observed a correlation that was positive in sign, statistically significant, and equal to 0.47 (95% confidence interval = 0.30 to 0.61). The significant coefficient and the sizable correlation is not unexpected even though the underlying model is additive because, as noted, product terms are amalgamations of main effects and interactions. Thus, using the first analysis suggested by Greenwald et al., one would conclude that the data at this stage of the analysis are consistent with a multiplicative model when, in fact, an additive model underlies the data.<sup>1</sup>

We stated in the previous section that the results of the Greenwald et al. (2002) regression analysis are not invariant to shifts in origins of the observed measures. This was evident in the simulation. For example, using mean-centered predictors, the correlation for predicting self-esteem from the product term alone was 0.13 and was statistically nonsignificant (95% confidence interval =

−0.07 to 0.32). These results led to a different set of conclusions than the original analysis.

In sum, the first subtest suggested by Greenwald et al. (2002) is problematic because of the strong measurement assumptions it makes and because it is susceptible to specification error. More fine-grained analyses of the regression residuals should help to flag misspecification in such analyses, but Greenwald et al. report no such analyses, and too often, residual analyses are not pursued in the literature more generally (see Cohen et al., 2003, for a discussion of residual analysis). We discuss later an analytic strategy that does not require the stringent measurement assumption and that is not subject to the form of specification error that creates problems for the Greenwald et al. approach.

### *Coefficients for Component Terms*

A second set of analyses that Greenwald et al. (2002) suggested for their multiplicative model is one that includes both the product term and its component parts in the regression equation (as per Equations 4–6). According to Greenwald et al., a multiplicative model is supported if the coefficients associated with the component parts are statistically nonsignificant. Greenwald et al.’s discussion of this subtest might lead readers to infer that the coefficients for the component parts of the product term reflect the “effects” of the two predictors on the criterion in a “main effect” sense (see Greenwald et al., 2002, p. 10). This is not the case. It is well known that these are conditional coefficients that represent what are commonly called “simple main effects” in the interaction literature (Cohen, 1978; Cohen et al., 2003). Specifically, they represent the slope of the criterion on one of the predictors when the other predictor equals zero. Because the coefficients associated with the component parts are conditional effects, it follows that their size and significance tests also are impacted by shifts in the origins of the measures. Simple transformations of the form  $X + c$  and  $Z + k$  affect these significance tests as well. Thus, like the product-term only model, this step in the Greenwald et al. strategy also requires that the observed scale origins be nonarbitrary. A failure of the data to conform to theory predictions permits the theorist to hedge by blaming countertheoretical results on faulty scaling assumptions. Although measurement assumptions are inherent to all statistical tests, the Greenwald et al. strategy is unique because its assumptions are much more stringent than is typically the case.

Another complication with this analysis is that support for the theory is based on statistical nonsignificance. This embraces the logic of “accepting” the null hypothesis, a practice that is widely recognized as problematic. In the studies reported by Greenwald et al. (2002; see the Appendix), the statistical power for tests of the component coefficients was low, making a theoretically consistent result likely. For example, if one assumes a sample size of 95 (which is comparable to Greenwald et al.’s largest sample size in the studies they report), a two-tailed alpha of 0.05, a population multiple correlation of 0.40 (which is near the median multiple correlation in the Greenwald et al., 2002, studies), a population

<sup>1</sup> A large-scale simulation of this matter could be pursued, although our purposes are purely pedagogical. For the relevant equations that underlie this phenomenon, see Aiken and West (1991).

predictor correlation between  $X$  and  $Z$  of 0.20, an interaction coefficient of 1.00 (which is near the median value in the Greenwald et al., 2002, studies), and a true regression coefficient for a given component part that is 0.50 (predictor variable) standard deviations from the value of 0, the statistical power for the component test is approximately 0.34.<sup>2</sup> For a sample size of 55 (near Greenwald et al.'s, 2002, smallest sample size), the approximate statistical power is 0.20.<sup>3</sup> A more effective statistical strategy for evaluating lack of an effect is that based on statistical equivalence testing as developed in the field of biostatistics and introduced to psychologists by Rogers, Howard, and Vessey (1993). At the very least, if one is going to make important theoretical inferences based on statistical nonsignificance, then the statistical power that is operative should be reported.

In the preceding simulation example, neither of the coefficients for the component parts of the product term were statistically significant when examined in a three-term regression equation. Thus, even though the generating model was additive, the data passed the criterion of Greenwald et al. (2002) at this stage.

*The Greenwald et al. (2002) Hierarchical Test*

The traditional procedure for calculating a standardized effect size for an interaction is to conduct a hierarchical analysis contrasting the sample squared multiple correlations for the following two equations:

$$Y = \alpha + \beta_1 X + \beta_2 Z + \epsilon \tag{7}$$

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon. \tag{8}$$

The squared multiple correlation for Equation 7 is subtracted from that of Equation 8. This indicates the proportion of variance in  $Y$  that can be “explained” by the interaction and is a popular index of effect size for interactions (although it is not without controversy; see Prentice & Miller, 1992; Jaccard, 1998). The statistical test of significance of the change in the squared correlation yields a result that is equivalent to testing bilinear interaction in a single equation that includes the product term and its component parts, that is, the  $p$  value for the hierarchical  $F$  test is identical to the  $p$  value for the regression coefficient for the product term.

By contrast, Greenwald et al. (2002) suggest a hierarchical strategy that calculates the squared multiple correlations for the following two equations:

$$Y = \alpha + \beta_1 XZ + \epsilon, \text{ and} \tag{9}$$

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon. \tag{10}$$

They then subtract the squared multiple correlation for Equation 9 from that of Equation 10. This approach evaluates how much the “interaction” term can account for variance in the outcome variable as contrasted with a model that includes both the “interaction” term and the component parts (or what is thought to be the additive “effects of the predictors”). The theoretical prediction is one of a nonsignificant change in the squared  $R$ s. Although this approach is at odds with traditional interaction analysis, we could discern one plausible rationale for the approach, not considered by Greenwald et al.

Suppose that the generating process for  $Y$  reflects a simple multiplicative function between two variables in accord with

Equation 9, that is, the multiplicative model is precisely true. Now suppose that the zero-point on the measure of  $X$  deviates from the true zero-point on the underlying construct by a constant of  $c_1$  and the zero-point for the observed measure of  $Z$  deviates from its construct’s true zero-point by a constant of  $c_2$ . This yields the equation

$$Y = \alpha + \beta_1(X - c_1)(Z - c_2) + \epsilon.$$

Some algebraic manipulation yields the equation

$$Y = (\alpha + \beta_1 c_1 c_2) - \beta_1 c_2 X - \beta_1 c_1 Z + \beta_1 XZ + \epsilon.$$

If we let  $\alpha'$  equal  $(\alpha + \beta_1 c_1 c_2)$ ,  $\beta'_1 = -\beta_1 c_2$ ,  $\beta'_2 = -\beta_1 c_1$ , and  $\beta'_3 = \beta_1$ , then the above becomes

$$Y = \alpha + \beta'_1 X + \beta'_2 Z + \beta'_3 XZ + \epsilon.$$

The Greenwald et al. (2002) strategy of comparing the two models thus becomes an evaluation of the relative fit of a model that assumes a simple multiplicative function with correctly specified origins versus a model that assumes a simple multiplicative function with incorrectly specified origins. Note that the focus of this analysis has shifted from empirically testing the theoretical viability of the true generating process to one of assuming that the multiplicative process is true and then testing the measurement assumption of correctly specified origins. The test of the coefficients associated with the product-term components neither has anything to do with testing for possible additivity in a theoretical sense nor evaluates the viability of the multiplicative generating process. Rather, it tests measurement properties under the assumption of a known generating function.

The above reflects one possible rationale for the Greenwald et al. (2002) approach. Some might argue that the lack of significant improvement in the fit of Equation 10 relative to Equation 9 provides support for correctly specified origins with respect to the measurement assumption of rational zeros. This is true only if one has complete confidence that the multiplying model correctly characterizes the generating process. However, it is the generating process that one is uncertain about and that one desires to evaluate.<sup>4</sup>

If one wanted, one could conduct formal tests that focus on measurement assumptions. Psychologists use a variety of strate-

<sup>2</sup> The rationale for focusing on standard deviation units from a presumed predictor origin of 0 will become apparent in the next section. In terms of standardized effect size indices for explaining  $Y$ , the effect for the two component parts map onto a medium effect size as characterized by Cohen (1978).

<sup>3</sup> These power estimates were obtained through computer simulation.

<sup>4</sup> If one asserts the presence of the multiplying rule, then there should exist a different linear transformation for each of the observed variables that will concentrate the regression sums of squares for each of the three regression analyses in the bilinear interaction component. Minimizing the sums of squares for the these three components could be a criterion for an iterative search for the three slopes and three intercept values that provide the transformations from the observed scale to the latent scale for each variable with the goal of removing the appropriate linear effects as well as the nonbilinear interaction effects. The scale transformation for a given variable would need to remain the same in each of the three analyses that the variable is involved in.

gies to evaluate the nature of metrics. For example, if one wanted to identify the measured value on a scale that corresponds to the “true zero” on an underlying dimension, one might use a strong, accepted, and well-developed theory that permits one to make predictions about how data for variables should pattern themselves as one moves across the dimension of interest and through the true zero-point. To isolate where the true zero-point occurs on the observed metric, the theory must predict a distinct data pattern for the zero-point. When one consistently observes the data pattern that should result at the true zero-point at a particular scale value (and not at others), then one has a basis for interpreting that number as being reflective of the true zero-point. However, note that the focus of this strategy is not on theory testing, because it requires the use of an already well accepted theory. Instead, the focus is on evaluating metric properties. The Greenwald et al. (2002) scenario is different in that it is precisely the theory one wishes to test.

As with the previous subtest, support at this step for the multiplicative model is inferred from a nonsignificant hierarchical test result, which suggests support for the theory based on the logic of accepting the null hypothesis. At the least, the statistical power of such tests should be reported. For a sample size of 95 (which is near the largest sample size used by Greenwald et al., 2002) and for the case where the squared  $R$  as a result of adding the component terms to the equation increases from 0.10 to 0.15 (representing 5% additional explained variance), the hierarchical test has statistical power of approximately 0.53. For a sample size of 55 under comparable conditions, the approximate statistical power is 0.32.

Parenthetically, the hierarchical test of Greenwald et al. (2002) as applied to our simulation example also was nonsignificant, providing yet additional “support” for the multiplicative model (even though an additive model was operative).

In sum, the hierarchical analysis of Greenwald et al. (2002) has some justification if one wants to test measurement properties (valid origins) under the assumption that the true generating process is known and is multiplicative. One could, of course, argue that a nonsignificant increment in fit for the two equations simultaneously affirms the multiplying generating process as well as the origins of the scales, but this is a large inferential leap. The nature of the violations of the measurement assumptions may, in principle, offset violations of the multiplying generation process, producing a net result that appears consistent with the theory. The analysis also leaves ambiguity as to what to conclude when the hierarchical difference in equation fit is large. Is it because the theory is wrong, because the scaling assumptions are wrong, or both?

### An Analytic Strategy With Fewer Auxiliary Assumptions

Fortunately, simple multiplicative models like Greenwald et al.’s (2002) can be evaluated without recourse to assumptions about scale origins and in ways that are not susceptible to the type of specification error noted above. Consider the case of a criterion variable,  $Y$ , and two predictor variables,  $X$  and  $Z$ , that are thought to influence  $Y$  in a multiplicative fashion. Anderson (1981) has shown that multiplicative models of this nature predict interactions of a specific form among the predictors, namely bilinear interactions. For bilinear interactions, the slope of  $Y$  on  $X$  exhibits a linear

fan pattern when the slope of  $Y$  on  $X$  is plotted as a function of different values of  $Z$  (see the examples in Anderson, 1981, and see Figure 6 in Greenwald et al.). Anderson (1981) refers to this as the linear fan theorem. In algebraic terms, the slope of  $Y$  on  $X$  is said to be a linear function of  $Z$ , such that

$$\beta_{Y \cdot X} = \alpha + \beta Z, \quad (11)$$

where  $\beta_{Y \cdot X}$  is the slope of  $Y$  on  $X$ . In this model, for every one unit that  $Z$  changes, the slope of  $Y$  on  $X$  is predicted to change by  $\beta$  units. A test of the linear fan theorem implied by a multiplicative model can be evaluated using a statistical analysis that isolates and yields significance tests for  $\beta$  in Equation 11. It turns out that the value of  $\beta$  in Equation 11 is equivalent to the value of  $\beta_3$  in the equation

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon. \quad (12)$$

A test of the multiplicative model should therefore focus on the significance of the regression coefficient for the product term in a model that includes both the component parts of the product term as well as the product term. Importantly,  $\beta_3$  in Equation 12 reflects  $\beta$  in Equation 11 irrespective of the origins of  $X$  and  $Z$ , so no assumptions about “rational zero-points” are required. Rather, the assumption is simply that the observed measure is a linear function of the underlying latent construct (i.e., the data are interval level). The origins can be arbitrary.

Applying this method of analysis to the simulation example noted earlier, the crucial test of the multiplicative model resides in the test of significance for the product term in the three-term regression equation. This regression coefficient was not statistically significant,  $B = 1.37$ ,  $t(91) = 1.06$ ,  $p < .29$ . The multiplicative model in this case is questioned because of a statistically nonsignificant bilinear interaction. This is directly opposite to what the Greenwald et al. (2002) test strategy concluded. When a main effect only model was fit to the data (by regressing self-esteem onto group attitude and group identification), the analysis yielded a multiple correlation of 0.50, a statistically significant coefficient for group attitude,  $B = 0.73$ ,  $t(92) = 4.57$ ,  $p < .001$ , and a statistically significant coefficient for group identification,  $B = 0.42$ ,  $t(92) = 2.65$ ,  $p < .010$ . These analyses reasonably reflect the true dynamics that are operating, given sampling error. This approach to the data yields results that are more in line with the true generating processes than the Greenwald et al. test strategy.

As noted, the test of the linear fan theorem using the product term in the full three-term model is invariant to scale origin. For example, if one defines a new set of origins by mean centering  $X$  and  $Z$ , recalculates the product term, and recalculates the regression equation, the multiple correlation is the same and the regression coefficient and significance test for the product term is the same. This invariance of results does not hold for the Greenwald et al. (2002) analysis of the product term alone.

To evaluate a multiplicative model, we suggest that researchers conduct traditional product term regression analyses of the form of Equation 12 and evaluate the statistical significance of the interaction predicted by the theory. Simple multiplicative models predict that the data should pattern themselves in accord with a bilinear interaction, so the above test provides perspectives on whether the data conform to this pattern.

In addition to a test of the product term coefficient in the three-term equation, researchers also should routinely perform

diagnostics to confirm that the patterning of data are in accord with the linear fan theorem. Pursuing such diagnostics embraces the notion of strong versus weak inference as developed by Anderson (1981, 2001), who eschews sole reliance on significance tests of correlations. Anderson (1981) suggests conducting formal statistical tests to ensure that the residual explained variance of the interaction that does not include the bilinear component is trivial in magnitude. For example, a model that includes terms representing quadratic or cubic interaction forms should exhibit nonsignificant effects for those terms. An informal diagnostic of departures from the bilinear form uses bandwidth regression (Hamilton, 1992). In this approach, one of the predictor variables ( $Z$ ) is grouped into 5 to 10 roughly equal sized, ordered categories. The mean  $Z$  is calculated for each group. A regression analysis is then performed regressing  $Y$  onto  $X$  for each of the  $Z$  groups separately. Examination of the coefficients for  $Y$  on  $X$  across the 5–10 groups defined by  $Z$  should reveal a trend whereby the coefficient increases or decreases roughly as a linear function of the mean  $Z$  for each group across the groups. If one plots from such an analysis the  $Y$  on  $X$  coefficients against the mean (or median)  $Z$  values, a linear trend should be evident. If this is not the case, then an interaction form other than bilinear may be operating and should be modeled accordingly. This technique requires that the sample size in each group be sufficiently large that the estimated coefficients are not too unstable. We believe that the assumption of a bilinear functional form is somewhat arbitrary and that other multiplicative functions may be operating in the theoretical scenarios considered by Greenwald et al. (2002).<sup>5</sup> The bandwidth regression approach is illustrated in the Appendix, where we present a reanalysis of the Greenwald et al. data.

When summarizing the results of their analyses across experiments, Greenwald et al. (2002) did not focus on the statistical significance of the product term coefficient in the three-term equation, only with whether it is positive in sign. Even though these researchers stressed the importance of examining this coefficient, they only stressed the need to examine its sign. Ironically, the one coefficient that we find to be most critical in the analysis of multiplicative models, they choose to downplay by ignoring its significance test in their statistical summary. In the case of a purely additive process in which there is no interaction effect operating at all, one would expect to observe a positively signed product term in a three-term equation about 50% of the time. Given this, a simple examination of the sign of the coefficient is questionable.<sup>6</sup>

One might think that Greenwald et al.'s (2002) procedure somehow represents a more stringent test than the traditional test of a bilinear interaction because it imposes multiple statistical criteria (or subtests) that the data must satisfy, whereas we emphasize only one. Such a conclusion is undermined by the fact that each of the Greenwald et al. criteria has nontrivial shortcomings. A collection of four weak tests is not necessarily better than a single strong(er) test. The cost of performing our recommended analysis relative to testing a purely multiplicative model without product terms (the first step of the Greenwald et al., 2002, analytic strategy) is a tiny loss of statistical power if the multiplicative model is true and if the measures do indeed have rational zeros. By contrast, if either the multiplicative model or the measurement assumption of non-arbitrary origins is not perfectly true, then the product-term only approach for step one of the Greenwald et al. strategy can yield bias in effect estimates such that both spurious effects may be

obtained and true effects not detected. In general, we believe that the traditional analysis of bilinear interaction is preferable to the Greenwald et al. strategy for testing a multiplicative model unless one has confidence in the metric assumptions and unless additive models are not viable.

### Conclusions and Comments

Strong theories that make unequivocal statements about an underlying generating process often are easier to test than weak theories that are vague about such processes. However, a strong theory alone cannot overcome bias in theory tests due to weak measurement nor accommodate tests that lack sensitivity to misspecified models (Birnbaum, 1973; Shanteau, 1977). Because psychologists deal with unobservable constructs that often are scaled on arbitrary metrics (e.g., attitudes on 1 to 7 scales and associations measured in milliseconds), we believe that the statistical strategy suggested by Greenwald et al. (2002) has real shortcomings when evaluating multiplicative models. Not only does it obscure theory tests with the validity of measurement assumptions about scale origins, but it also can lead too often to the acceptance of incorrect models because it is insensitive to specification error for simple additivity (see Busemeyer & Jones, 1983, and Aiken & West, 1991, for the underlying mathematics).<sup>7</sup>

A straightforward test of theory predictions that is not susceptible to these problems is the traditional test of bilinear interaction in regression analysis. Although this test assumes that the measures are interval level (a nontrivial assumption in its own right), it neither makes assumptions about scale origins nor is subject to the type of specification error inherent in the Greenwald et al. (2002) analysis.

<sup>5</sup> Both our approach and the Greenwald et al. (2002) approach assume, of course, that the basic assumptions of ordinary least squares linear regression hold, and if the assumptions are nontrivially violated, then alternative methods for estimating coefficients are necessary (see Cohen et al., 2003, for a discussion of strategies for evaluating assumptions and Wilcox, 2003, for a description of robust estimation procedures). Bias in coefficients also can result given the presence of random measurement error and analytic methods that can accommodate such error in product term models are desirable (see Wall & Amemiya, 2000, for a promising approach). The approaches also must assume that the multiplicative process occurs for all study participants (Anderson, 1981), otherwise special interaction terms must be introduced to accommodate such individual differences. Finally, both approaches require that there is sufficient statistical power to detect meaningful effects. Also, Greenwald et al. conducted many of their analyses on reaction time data, which have special analytic needs (see Ratcliff, 2002).

<sup>6</sup> Greenwald et al. (2002) reported that across 16 tests of the product term in Equation 10, the coefficient was positive in 15 of them, a result that is not easily attributable to chance and that is consistent with the Greenwald et al. theory. Most researchers will apply the Greenwald et al. strategy in experiments that have only a few such tests and where formal sign tests of coefficient direction are problematic, especially given the dependencies among the estimating equations.

<sup>7</sup> Although our simulation represented an example where the Greenwald et al. (2002) tests affirmed a multiplicative model when an additive model was the true generating process, it also is possible that their strategy will reject a true multiplying model given the presence of nonarbitrary origins in the observed measures. Their test can thus lead to incorrect conclusions in either scenario.

Supposed counterexamples may come to mind from the physical sciences where purely multiplicative models are tested without concern for including component parts in a prediction equation. These include Einstein's famous  $E = mc^2$ , Newton's formula that force = mass  $\times$  acceleration, and the area of a rectangle being equal to its length  $\times$  its width. All of these formulations include just product terms, and one might argue that it would be nonsensical to include the component parts of the terms in a prediction equation. Such is not the case. Consider the latter example of rectangles. If one predicts the area of rectangles from a simple product term model that multiplies length times width, then fitting such a model is legitimate and diagnostic *given that* one has measures whose origins map onto true rational zeros and *given that* one is not concerned with potential specification error in the form of an additive model. If one does not have such measures or if such specification error is a concern, then the analysis of a simple product term model using regression strategies could be misleading. Fortunately, even without measures whose origins map onto rational zeros, one can still gain perspectives on whether, for example, area is a simple multiplicative function of length times width. If it is, one should observe a linear fan when one plots the linear relationship between area and length at different values of width. Stated another way, if one regresses area onto width, length, and the width  $\times$  length product term, then the coefficient for the product term should be nonzero (and statistically significant, if random error is present). Such a result maps onto a bilinear interaction and is consistent with the generating multiplicative process. So, even with measures without rational zeros, one can gain perspectives on the generating process using the strategy we propose. Had Einstein or Newton been stuck with the types of measures that psychologists typically work with, and if their experimental designs were such that they had to work with correlational data and regression models, then they would indeed have needed to include component parts of the product term in their empirical tests of their models.

## References

- Aiken, L. N., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Ajzen, I., & Fishbein, M. (1981). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.
- Anderson, N. H. (1981). *Methods of information integration*. New York: Academic Press.
- Anderson, N. H. (2001). *Empirical direction in design and analysis*. Mahwah, NJ: Erlbaum.
- Anderson, N. H., & Butzin, C. A. (1974). Performance = Motivation  $\times$  Ability: An integration-theoretical analysis. *Journal of Personality and Social Psychology*, 30, 598–604.
- Asendorpf, J. B., Banse, R., & Muecke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, 83, 380–393.
- Banaji, M. R., Greenwald, A. G., & Rosier, M. (1997, October). *Implicit esteem: When collectives shape individuals*. Paper presented at the Preconference on Self, Toronto, Ontario, Canada.
- Biernat, M., Manis, M., & Nelson, T. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, 60, 485–499.
- Birnbaum, M. H. (1973). The devil rides again: Correlation as an index of fit. *Psychological Bulletin*, 79, 239–242.
- Busemeyer, J., & Jones, L. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, 93, 549–562.
- Cohen, J. (1978). Partialled products are interactions; partialled powers are curve components. *Psychological Bulletin*, 85, 858–866.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.
- Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analysis recently proposed. *Psychological Bulletin*, 102, 414–417.
- de Jong, P. J., van den Hout, M. A., Rietbroek, H., & Huijding, J. (2003). Dissociations between implicit and explicit attitudes toward phobic stimuli. *Cognition and Emotion*, 17, 521–545.
- Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, 82, 835–848.
- Edwards, W., & Fasolo, B. (2000). Decision technology. *Annual Review of Psychology*, 52, 581–606.
- Farnham, S. D., & Greenwald, A. G. (1999, June). *In-group favoritism = Implicit Self-Esteem  $\times$  In-Group Identification*. Paper presented at the 11th annual meeting of the American Psychological Society, Denver, CO.
- Gray, N. S., MacCulloch, M. J., Smith, J., Morris, M., & Snowden, R. J. (2003, May 29). Forensic Psychology: Violence viewed by psychopathic murderers. *Nature*, 423, 497–498.
- Greenwald, A. G., Banaji, M., Rudnam, L., Farnham, S., Nosek, B. A., & Mellott, D. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3–25.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Gupta, M., & Singh, R. (1981). An integration theoretical analysis of cultural and developmental differences in attribution of performance. *Developmental Psychology*, 17, 816–825.
- Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. Belmont, CA: Brooks/Cole.
- Irwin, J., & McClelland, G. H. (2002). Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research*, 38, 100–109.
- Jaccard, J. (1998). *Interaction effects in factorial analysis of variance*. Newbury Park: Sage.
- Jaccard, J., & Turrissi, R. (2002). *Interaction effects in multiple regression*. Newbury Park: Sage.
- Karpinski, A. (2004). Measuring self-esteem using the Implicit Association Test: The role of the other. *Personality and Social Psychology Bulletin*, 30, 22–34.
- Karpinski, A., & Hilton, J. L. (2001). Attitudes and the implicit association test. *Journal of Personality and Social Psychology*, 81, 774–788.
- Kenny, D. A., & Judd, C. M. (1984). The nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96, 201–210.
- McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the Implicit Association Test. *Social Cognition*, 20, 483–510.
- Mellott, D. S., & Greenwald, A. G. (2000). *But I don't feel old! Implicit self-esteem, age identity, and ageism in the elderly*. Unpublished manuscript, University of Washington, Seattle.
- Nosek, B., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math  $\mu$  me. *Journal of Personality and Social Psychology*, 83, 44–59.
- Oden, G. C. (1978). Semantic constraints and judged preferences for interpretations of ambiguous sentences. *Memory & Cognition*, 3, 336–352.

- Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measure of prejudice: What are we measuring? *Psychological Science, 14*, 636–639.
- Phelps, E. A., Cannistraci, C. J., & Cunningham, W. A. (2003). Intact performance on an indirect measure of face bias following amygdala damage. *Neuropsychologia, 41*, 203–208.
- Prentice, D., & Miller, D. T. (1992). When small effect sizes are impressive. *Psychological Bulletin, 112*, 160–164.
- Ratcliff, R. (2002). Estimating parameters of the diffusion model: Approaching to dealing with contaminant reaction and parameter variability. *Psychonomic Bulletin and Review, 9*, 438–481.
- Rogers, J. L., Howard, K., & Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553–565.
- Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test (IAT): Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*, 139–165.
- Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin, 27*, 1164–1178.
- Shanteau, J. (1977). Correlation as a deceiving measure of fit. *Bulletin of the Psychonomic Society, 10*, 134–136.
- Teachman, B. A., Gapinski, K. D., Brownell, K. D., Rawlins, M., & Jeyaram, S. (2003). Demonstrations of implicit anti-fat bias: The impact of providing causal information and evoking empathy. *Health Psychology, 22*, 68–78.
- Tukey, J. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83–91.
- Wall, M. M., & Amemiya, Y. (2000). Estimation for polynomial structural equation models. *Journal of the American Statistical Association, 95*, 929–940.
- Wiers, R. W., VanWoerden, N., Smulders, F. T., & de Jong, P. J. (2002). Implicit and explicit alcohol-related cognitions in heavy and light drinkers. *Journal of Abnormal Psychology, 111*, 648–658.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. New York: Academic Press.

## Appendix

### Reanalysis of Greenwald et al.'s (2002) Data

This appendix presents a reanalysis of data summarized in Greenwald et al. (2002). The data were graciously provided to us by Anthony G. Greenwald.

The theory in Greenwald et al. (2002) predicts a bilinear interaction between two variables when predicting a third variable. The focus of our analysis is thus on the three-term regression equation that predicts an outcome variable from a product term and the component parts of that product term. Of primary interest is the statistical significance of the coefficient associated with the product term in the three-term equation, as this reflects a bilinear interaction without assuming rational zeros in the scales. In their summary of analyses across experiments, Greenwald et al. do not describe the results of significance tests of this product term coefficient. Rather, they describe the number of times that it was positive in sign without regard to its statistical significance. Descriptions of the results of the significance test for the product term occur sporadically for some of the experiments but not others. Our contention is that the significance test of the coefficient for the product term in the three-term equation is central to theory evaluation.

#### The Product-Term Coefficient in the Three-Term Equation

Table A1 presents the means and standard deviations of the major variables used in each of the studies reported by Greenwald et al. (2002) and Table A2 presents the results of a regression analysis for the three-term equation in each experiment. Within a given experiment and for a given type of measure (implicit or explicit), support for a bilinear interaction is present if the coefficient for the product term for three-term equations is statistically significant (see  $b_3$  in Column 5 of Table A2). Table A2 follows tradition for interaction analysis by presenting unstandardized coefficients, whereas Greenwald et al. presented the partial correlations between  $Y$  and  $XZ$  holding constant  $X$  and  $Z$  when characterizing the interaction. The unstandardized coefficient for the product term directly describes the nature of the moderating influences operating (as shown below), whereas this is not necessarily true of standardized indices (see Aiken & West, 1991). We adopt the more traditional strategy of reporting unstandardized coefficients.

Across the 25 regression equations, the product-term coefficient was statistically significant 9 times or 36% of the time and statistically non-

significant 16 times or 64% of the time. Greenwald et al. (2002) argue that theoretical predictions are more likely to be upheld for the implicit measures as opposed to the explicit measures. Of the 16 regression equations that used implicit measures, the product term coefficient was statistically significant 8 times or 50% of the time.

Across the five studies in Table A2, the theory of Greenwald et al. (2002) predicts that the bilinear interaction pattern should replicate when each of three variables in a set of three is treated as the outcome variable.<sup>A1</sup> Across the 8 sets of three regression equations, the product term coefficient was statistically significant in all three equations once, or 12% of the time, with one or more of the coefficients being nonsignificant 88% of the time. Focusing on just the implicit measures, the product term coefficient was statistically significant in all three equations one of six times, or 17% of the time, with one or more of the coefficients being nonsignificant 83% of the time.

In the above summary, we use only descriptive “vote counting” for characterizing the number of significant results. We do not apply meta-analytic procedures as they introduce nontrivial complexities for these data. Traditional meta-analytic statistical models cannot be applied (either in the form of inferential vote counting or effect size analysis) because some of the 27 equations are based on the same respondents, whereas others are not (i.e., there is a mixture of independent and dependent equations). Some of the equations involve the exact same variables (but in different roles as predictors and criteria), while others do not. Hence, the dependency structure among the equations is complex and inconsistent with the statistical theory underlying traditional meta-analytic methods. In addition, the topic areas studied were diverse (gender identification, racial identification, age identification) and there may exist topic differences in the magnitude of the interactions that a meta-analysis could obscure. For example, assume in eight populations that the null hypothesis is true in four of the populations but not the other four. A researcher might apply traditional fixed effects meta-analysis across the eight studies to evaluate the null hypothesis of no effect across them. The test result could well be statistically significant (because half of the populations have a false null hypothesis) but it would

<sup>A1</sup> The summary of the study by Nosek, Banaji, & Greenwald (2002) reported four equations because of the presence of an additional variable.

Table A1  
Means, Standard Deviations, and Sample Sizes for Key Variables

Study	Measure type	Construct	<i>M</i>	<i>SD</i>	<i>N</i>
FG	Implicit	Self-positive	0.36	0.18	65
FG	Implicit	Female-positive	0.34	0.14	65
FG	Implicit	Female-self	0.25	0.18	65
BGR	Implicit	Self-positive	0.20	0.13	61
BGR	Implicit	White-positive	0.10	0.11	61
BGR	Implicit	White-self	0.05	0.20	61
MG	Implicit	Self-positive	-0.16	0.20	98
MG	Implicit	Old-positive	-0.23	0.20	98
MG	Implicit	Old-self	0.25	0.20	98
RGM	Implicit	Self-warm	0.06	0.14	95
RGM	Implicit	Female-warm	0.20	0.15	95
RGM	Implicit	Female-self	0.01	0.28	95
NBG	Implicit	Math-positive	-0.01	0.21	91
NBG	Implicit	Math-self	0.26	0.19	91
NBG	Implicit	Male-self	-0.25	0.21	91
NBG	Implicit	Male-math	-0.01	0.16	91
FG	Explicit	Self-positive	1.41	0.80	57
FG	Explicit	Female-positive	0.22	0.70	57
FG	Explicit	Female-self	4.39	1.00	57
MG	Explicit	Self-positive	1.75	0.97	91
MG	Explicit	Old-positive	0.02	0.71	91
MG	Explicit	Old-self	-0.35	0.97	91
RGM	Explicit	Self-warm	0.92	0.96	95
RGM	Explicit	Female-warm	1.11	0.87	95
RGM	Explicit	Female-self	0.36	1.58	95

*Note.* Initials in the “Study” column refer to the last names of the authors to whom the study was attributed in Greenwald et al. (2002). FG = Farnham and Greenwald (1999); BGR = Banaji, Greenwald, and Rosier (1997); MG = Mellott and Greenwald (2000); RGM = Rudman, Greenwald, and McGhee (2001); NBG = Nosek, Banaji, and Greenwald (2002).

be an error to assume from this that the effect is robust across the eight studies. Such an analysis can obscure important effect differences as a function of topic domain or study characteristics and this possibility needs to be explored carefully.

#### The Form of the Interaction

A potential limitation of the three-term equation is that it only models interactions that are bilinear in form. It is possible for two continuous variables to interact with one another in the prediction of a third but for the interaction to take a form that departs from the bilinear pattern. Consistent with the philosophy of Tukey (1969), we recommend that any test of a bilinear interaction be complemented by exploratory analyses using regression diagnostics that permit the data to speak to potential violations of the predicted functional form. To the extent that observed violations make theoretical sense, they may lessen the confidence in the initially posited model.

To illustrate this point, consider an equation from the Rudman, Greenwald, and McGhee (2001) study reported in Greenwald et al. (2002) where the outcome variable was the strength of the association between self and the attribute warmth (*Y*) and the predictors were the strength of the association between female and warmth (*X*) and the strength of the association between self and female (*Z*). For the three-term equation, the regression coefficient for the product term *XZ* was 1.12 ( $p < 0.01$ ). Some researchers believe that this coefficient reflects an interaction in its most general sense, when, in fact, it reflects a very specific interaction form. Specifically, a bilinear interaction implies symmetry in the way that a predictor variable moderates the impact of the other predictor variable on the outcome. To be concrete, the product term coefficient estimates how the slope of *Y* on *X* is assumed to vary as a function of changes in *Z*. For the simple linear equation

$$Y = a_x + b_x X,$$

the value of the coefficient  $b_x$  is assumed to increase by 1.12 units every time *Z* increases by one unit. The product term coefficient also estimates how the slope of *Y* on *Z* is assumed to vary as a function of changes in *X*. For the simple linear equation

$$Y = a_z + b_z Z,$$

the value of the coefficient  $b_z$  also is assumed to increase by 1.12 units every time *X* increases by one unit. A bilinear interaction (and a multiplicative model) assumes symmetry in the two moderated influences in that they are assumed to be linear, identical in sign and identical in magnitude. It is possible, however, that the moderating influences are not symmetrical but rather asymmetrical, with the moderating influence of *X* on the *Y-Z* relationship being different than the moderating influence of *Z* on the *Y-X* relationship.

Another way of conceptualizing this dynamic is in terms of the function relating different values of the slope of  $b_x$  to values of *Z*. A bilinear interaction (and by implication, the simple multiplicative model) assumes that  $b_x$  is a linear function of *Z*, such that

$$b_x = a_{xz} + b_{xz} Z.$$

But, perhaps the relationship between  $b_x$  and *Z* is not linear. For example, the function might be quadratic such that

$$b_x = a_{xz} + b_{xz1} Z + b_{xz2} Z^2.$$

If the function is indeed nonlinear, then the model with a bilinear interaction is misspecified. It is possible to obtain a statistically significant product-term coefficient in a model that assumes that  $b_x$  varies as a linear function of *Z*, even when the true function is nonlinear in form. It thus is

Table A2  
Results of Three-Term Analyses

Study	Measure	$b_1$	$b_2$	$b_3$	$R^2$	$N$
FG	Implicit	-0.268	-0.034	1.341	0.20	65
FG	Implicit	-0.144	-0.305	1.477*	0.24	65
FG	Implicit	0.108	0.046	0.416	0.21	65
BGR	Implicit	0.533	-0.354	2.358	0.46	61
BGR	Implicit	0.014	0.031	1.361**	0.49	61
BGR	Implicit	-0.280*	0.059	2.018**	0.16	61
MG	Implicit	0.158	-0.285	0.092	0.17	98
MG	Implicit	-0.391**	0.306	-0.443	0.18	98
MG	Implicit	-0.246	-0.290*	0.151	0.22	98
RGM	Implicit	-0.004	-0.004	1.075**	0.13	95
RGM	Implicit	-0.634*	-0.105	4.772**	0.18	95
RGM	Implicit	-0.028	-0.195*	1.116**	0.13	95
NBG	Implicit	-0.029	-0.966	1.332*	0.07	91
NBG	Implicit	0.111	-0.006	0.711	0.19	91
NBG	Implicit	-0.119	0.320	0.655	0.19	91
NBG	Implicit	-0.102	0.115	1.674**	0.21	91
FG	Explicit	0.216	-0.085	0.023	0.08	57
FG	Explicit	-0.581	0.279	0.133	0.17	57
FG	Explicit	-0.666	-0.417	0.158	0.12	57
MG	Explicit	-0.422	0.139	0.185	0.05	91
MG	Explicit	0.089	-0.389*	0.178	0.05	91
MG	Explicit	0.156	0.062	0.230	0.02	91
RGM	Explicit	-0.215**	0.081	0.048	0.12	95
RGM	Explicit	-0.121	-0.104**	0.377*	0.17	95
RGM	Explicit	0.089	-0.024	0.087	0.04	95

Note.  $b_1$  is the coefficient for  $X$ ,  $b_2$  is the coefficient for  $Z$ , and  $b_3$  is the coefficient for  $XZ$ . Initials in the “Study” column refer to the last names of the authors to whom the study was attributed in Greenwald et al. (2002). FG = Farnham and Greenwald (1999); BGR = Banaji, Greenwald, and Rosier (1997); MG = Mellott and Greenwald (2000); RGM = Rudman, Greenwald, and McGhee (2001); NBG = Nosek, Banaji, and Greenwald (2002). \*  $p < .05$ . \*\*  $p < .01$ .

important to examine regression diagnostics to ensure that the function relating  $b_X$  to  $Z$  is indeed linear.

One informal diagnostic uses bandwidth regression, which we illustrate using the Rudman et al. (2001) data. This approach is not a formal test of departure from bilinear interaction form, but rather it is a diagnostic that one can perform to gain a sense of whether the interaction is approximately bilinear. The first step was to create 5 to 10 strata that group together individuals with similar scores on  $Z$  and where the groups can be ordered on the underlying dimension that  $Z$  is assumed to reflect. Given the small sample size in Rudman et al. ( $N = 95$ ), we decided to create five groups of respondents with  $N = 19$  in each group. Group 1 was the 19 respondents with the lowest  $Z$  scores; group 2 was the 19 respondents with the next lowest  $Z$  scores; and so on, until we reached group 5 that was the 19 respondents with the highest  $Z$  scores. For each group separately, we regressed  $Y$  onto  $X$  and noted the value of the regression coefficient. The values for each group, as well as the mean  $Z$  score for that group were:

	Mean $Z$	Slope of $Y$ on $X$
Group 1	-0.39	-0.55
Group 2	-0.18	-0.09
Group 3	0.06	0.07
Group 4	0.17	0.43
Group 5	0.38	0.22

A bilinear interaction should exhibit roughly a linear relationship between the two columns of numbers. This is crudely in place (allowing for sampling error because of the small  $N$  per group), except for a somewhat glaring departure for the slope for Group 5 relative to Group 4. Rather than increasing, it tends to flatten off or decrease relative to Group 4.

This diagnostic suggests that the interaction may not be bilinear in form. A more formal test of this can be pursued. It appears from the above that

the slope changes are a quadratic function of  $Z$  rather than a linear function. Using the logic developed in Kenny and Judd (1984), we fit a quadratic interaction model using the equation:

$$Y = a + b_1X + b_2Z + b_3Z^2 + b_4XZ + b_5XZ^2.$$

If  $b_5$  is statistically significant, then this is consistent with the proposition that the interaction is not bilinear. When the above equation was tested for the Rudman et al. (2001) equation referenced above,  $b_5$  was statistically significant ( $p < 0.05$ ). It is beyond the scope of this article to describe the methods by which the above equation can be decomposed to isolate the predicted slope of  $Y$  on  $X$  at different values of  $Z$  (interested readers are referred to Jaccard and Turrisi, 2002) or the complex issues involved in pursuing non-linear interactions. The main point we make is that even though there was a statistically significant coefficient for the product term in the three-term equation, further diagnostics and evaluation suggest that the form of the interaction predicted by the multiplicative model may be violated.

The above analysis is counter to a simple multiplicative model, but it is not necessarily damaging to the Greenwald et al. (2002) theory. Although one might expect the slope of  $Y$  on  $X$  to increase as  $Z$  gets stronger, there probably is a ceiling effect where the values of the slopes reach an asymptote and don't increase any more as  $Z$  increases. This would produce a flattening of the curve so that the values of the slopes stay about the same with increases in  $Z$ . This asymptote may have been reached and exceeded in the Rudman et al. (2001) analysis.

For the experiments reported by Greenwald et al. (2002), it is difficult to explore and test for deviations from bilinear interactions because the sample sizes are small and the studies are underpowered for such purposes. This problem can be illustrated by considering the two hypothetical “curves” in Figure A1 that characterize how a slope changes as a function of a moderator variable  $Z$ . Let the solid line be the true nonlinear function that describes the relationship and the dotted line represent the line that

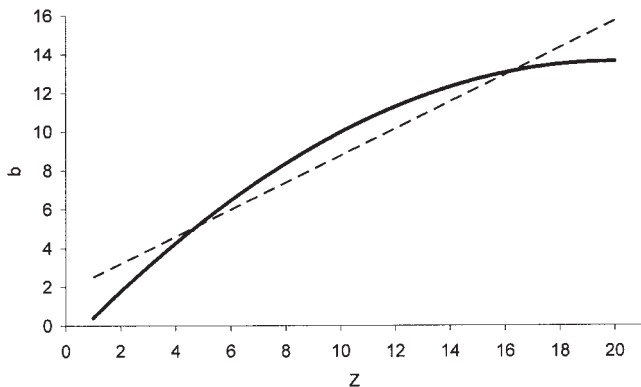


Figure A1. The change in slope of  $Y$  on  $X$ , as a function of a moderator variable  $Z$ . The solid line represents the true nonlinear function that describes the relationship. The dotted line represents the results when a linear model is fit to the values.

results when a linear model is fit to the values. The sample size must be sufficiently large to be sensitive to departures from linearity such as the one illustrated in the figure. In addition, replication of the above result depends, in part, on the range of  $X$  and  $Z$  values activated within an experiment. In Figure A1, for instance, if a study activates  $Z$  values in the 1 to 8 range only, then the true function within these values is essentially linear. If a study activates values from 12 to 20, then the true function is nonlinear. In the Greenwald et al. studies, the choice of stimuli and different features of

the task may affect which values of the association strength are activated. Because of low power and our primary intent to illustrate the use of regression diagnostics (and because of the theoretical complexities of interpreting the non-linear interactions in the substantive domains considered by Greenwald et al. relative to space constraints), we do not pursue these analyses further.

### Outliers and Nonnormality

In all analyses, we conducted additional diagnostics to explore the data for possible adverse effects of outliers and for bias due to non-normality. For outliers, we examined  $df$ Betas associated with each coefficient in the three-term regression model and found two cases where outliers raised issues. First, for the last row of entries in Table A2 (Rudman et al., 2001), the  $b_3$  coefficient became statistically significant when an individual with a nontrivial  $df$ Beta was deleted from the data. Second, in our analysis of the quadratic interaction term for the Rudman et al. data, we identified an individual who, if eliminated, resulted in the  $b_5$  coefficient becoming statistically nonsignificant. These results suggest caution when interpreting these particular analyses as the inferential results appear to be somewhat fragile. For non-normality, we reestimated each model using bootstrap procedures to estimate coefficient standard errors and bias corrected confidence intervals to test the null hypothesis (by examining if zero was contained within the interval). These analyses yielded results comparable to those portrayed in Table A2.

Received October 4, 2002

Revision received May 11, 2004

Accepted May 12, 2004 ■

DOI: 10.1037/0033-295X.113.1.166

### Postscript: Perspectives on the Reply by Greenwald, Rudman, Nosek, and Zayas (2006)

Hart Blanton  
Texas A&M University

James Jaccard  
Florida International University

In their reply to our article, Greenwald, Rudman, Nosek, and Zayas (2006) offered comments on the use of their procedure relative to one that focuses on standard multiple regression (SMR) methods. Here we address five issues where we think a balance of perspective is called for: (a) issues of measurement, (b) the structuring and reporting of the simulations, (c) testing “pure” multiplicative models, (d) the meta-analysis, and (e) tests for deviations from bilinearity.<sup>1</sup> On several occasions in their reply, Greenwald et al. stated our positions in the extreme (e.g., that we argued that no psychological measure can have ratio level properties). We will not react to these statements here, but we urge the reader to not be swayed by the setting up of straw men.

### Issues of Measurement

#### Theory Tests and Measurement Assumptions

Greenwald et al. (p. 170 and Footnote 1) stated that because the results reported in Greenwald et al. (2002) conformed to expectations of their theory, this therefore confirms both their theory as

well as the measurement assumption of a rational zero. They argue that “empirical tests of a theory will simultaneously appraise the theory and any auxiliary measurement assumptions made in testing it” (Greenwald et al., 2002, Footnote 1). We question this argument. For example, it is well known that ordinal level measures can produce spurious interaction effects at the measured level in the absence of a true interaction at the latent variable level, just as such measures can mask true interaction effects (Busemeyer & Jones, 1983). A predicted interaction effect can just as easily be the result of faulty measures as it can be the result of the truth of a theory. Greenwald et al. (2002) attributed lack of support for their theory as possibly being due to the violation of scaling assumptions. However, in their reply, Greenwald et al. took the position that as long as the result is consistent with theory, then one need not consider possibilities of measurement artifact because the result affirms both theory and measurement. Apparently, scaling assumptions are relevant when results are counter to theory but are not relevant when results are consistent with theory. We disagree with this logic. Ironically, Greenwald et al.’s own simulation that used mean centering suggests that a result that is consistent with theory does not simultaneously validate theory and underlying measurement assumptions. In this simulation, Greenwald et al. used measures that did not have rational zeros in the presence of a true multiplicative model. When they applied their tests, they

<sup>1</sup> Throughout this postscript, reference to Greenwald et al. means Greenwald et al. (2006), unless otherwise noted.