*Development and Psychopathology* **24** (2012), 389–409
© Cambridge University Press 2012
doi:10.1017/S0954579412000065

*SPECIAL SECTION ARTICLE*

# Distinguishing differential susceptibility from diathesis–stress: Recommendations for evaluating interaction effects

GLENN I. ROISMAN, DANIEL A. NEWMAN, R. CHRIS FRALEY, JOHN D. HALTIGAN, ASHLEY M. GROH, AND KATHERINE C. HAYDON
*University of Illinois at Urbana–Champaign*

**Abstract**

This report describes the state of the art in distinguishing data generated by differential susceptibility from diathesis–stress models. We discuss several limitations of existing practices for probing interaction effects and offer solutions that are designed to better differentiate differential susceptibility from diathesis–stress models and quantify their corresponding implications. In addition, we demonstrate the utility of these methods by revisiting published evidence suggesting that temperamental difficulty serves as a marker of enhanced susceptibility to early maternal caregiving across a range of outcome domains in the NICHD Study of Early Child Care and Youth Development. We find that, with the exception of mother reports of psychopathology, there is consistent evidence in the Study of Early Child Care and Youth Development that the predictive significance of early sensitivity is moderated by difficult temperament over time. However, differential susceptibility effects emerged primarily for teacher reports of academic skills, social competence, and symptomatology. In contrast, effects more consistent with the diathesis–stress model were obtained for mother reports of social skills and objective tests of academic skills. We conclude by discussing the value of the application of this work to the next wave of Gene × Environment studies focused on early caregiving experiences.

Statistical moderation has been central to theoretical accounts of risk, resilience, and (mal)adaptation from the beginning of its formal study (Garmezy, Masten, & Tellegen, 1984). Indeed, in many respects the study of resilience is *by definition* the search for protective factors that moderate the impact of psychosocial stressors on development (Luthar, Cicchetti, & Becker, 2000; Masten, 2001). Similarly, one of the dominant models of risk and developmental adaptation—the *diathesis–stress* or *dual-risk* model—explicitly posits moderation (Heim & Nemeroff, 1999). Specifically, the diathesis–stress model suggests that poor developmental experiences (e.g., low quality parenting) are most likely to impact the development of individuals who carry vulnerability factors, which are latent diatheses that result in maladaptation when "triggered" by poor experience. Many potential diatheses have been the focus of empirical inquiry, including temperamental traits, such as difficulty (Belsky, Hsieh, & Crnic, 1998), and specific molecular genetic polymorphisms, such as the short allele of the serotonin transporter gene (Caspi et al., 2003).

One common theme in this research is that such moderators are "vulnerability factors" that inhibit successful adaptation in the face of adversity. Thus, one of the more important developments in the last decade is the evolutionary psychology-inspired theory of *differential susceptibility*, which is a theory that assumes that these factors are not diatheses but are actually plasticity agents (Belsky, 1997). More specifically, differential susceptibility theory posits that many putative sources of vulnerability (e.g., difficult temperament) are more accurately conceptualized as plasticity factors because they not only amplify risk for maladaptation given poor caregiving experiences (as in the diathesis–stress account), but also increase the probability of *positive* adaptation given high-quality caregiving experiences (Belsky & Pluess, 2009). In short, the basic logic of differential susceptibility theory is that individuals vary in their susceptibility to childhood socialization agents (i.e., parents and surrogate caregivers), both "for worse" *and* "for better" (Belsky, Bakermans-Kranenburg, & van IJzendoorn, 2007).

A growing body of research suggests that certain risk factors may function in this manner (e.g., Belsky & Pluess, 2011; Bradley & Corwyn, 2008; Dopkins-Stright, Cranley Gallagher, & Kelley, 2008; Pluess & Belsky, 2009, 2010). For example, Kochanska, Kim, Barry, and Philibert (2011) recently demonstrated that variants of the serotonin transporter linked promoter polymorphism (*5-HTTLPR*) moderate the association between maternal responsiveness and moral

This study was sponsored by Grant BCS-0720538 from the National Science Foundation (to G.I.R. and R.C.F.), whose support is gratefully acknowledged. We also thank Jay Belsky and Michael Pluess for their thoughtful comments on an earlier version of this paper.

Address correspondence and reprint requests to: Glenn I. Roisman, Department of Psychology, University of Illinois at Urbana–Champaign, 603 East Daniel Street, Champaign, IL 61820; E-mail: roisman@illinois.edu.



https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

internalization among children. Specifically, when maternal responsiveness was low, children with one or more short forms of this variable number tandem repeat exhibited lower moral internalization relative to children with two long forms of the promotor polymorphism. When maternal responsiveness was high, however, children with one or more short forms of this variable number tandem repeat exhibited *greater* moral internalization relative to children with two long forms.

As noted earlier, foundational to attempts to distinguish between alternative accounts of risk and development is the concept of statistical moderation. Formally, moderation refers to the case in which the magnitude of the association between a predictor *X* and outcome *Y* varies as a function of moderator *Z*, which is typically evaluated by testing the statistical significance of the interaction between *X* and *Z* on *Y* in the context of multiple regression (Aguinis & Gottfredson, 2010; Aiken & West, 1991; Baron & Kenny, 1986). However, both the diathesis–stress and differential susceptibility models posit statistical interactions between *X* and *Z*. Thus, the presence of an interaction per se does not allow these alternative models to be differentiated. Instead, the models must be compared by probing the interactions in depth (see Dearing & Hamilton, 2006).

Although various methods for distinguishing between diathesis–stress and differential susceptibility models have been recommended in the published literature (e.g., Belsky et al., 2007; Belsky & Pluess, 2009), these recommendations do not provide unambiguous guidelines on how evidence for and against differential susceptibility should be evaluated and presented. The purpose of this article is to discuss these issues in depth and to offer a unified set of guidelines for future research. We begin by reviewing the basic concepts of statistical moderation and their relevance for evaluating diathesis–stress and differential susceptibility models. We then discuss in depth four critiques of approaches that have been used in published studies to evaluate evidence for differential susceptibility over diathesis–stress (e.g., Bradley & Corwyn, 2008; Dopkins-Stright et al., 2008). Specifically, these four critiques of past differential susceptibility interpretations, which we discuss in detail below, are labeled (a) the *visual inspection critique*, (b) the *quantification critique*, (c) the *Type I error critique*, and (d) the *nonlinearity critique*.

After presenting these critiques, we offer some solutions to the problems we identify and provide guidelines that researchers should follow when attempting to differentiate diathesis–stress from differential susceptibility accounts of development. Although our focus here is on methods for establishing evidence for differential susceptibility, we note from the outset that all of our recommendations are potentially important in the evaluation of evidence for diathesis–stress interactions as well. Finally, in order to demonstrate the utility of these specific guidelines, we revisit published evidence from the NICHD Study of Early Child Care and Youth Development (SECCYD) that temperamental difficulty serves to make individuals more susceptible to early maternal caregiving across a range of outcome domains (Bradley & Corwyn, 2008; Dopkins-Stright et al., 2008; Pluess & Belsky, 2010). Based on these illustrative reanalyses, we conclude that although some of the current differential susceptibility evidence survives when implementing these newly formalized criteria, alternative conclusions may be reached in some cases, depending on the specific methods used to examine these models.

## Differential Susceptibility or Diathesis–Stress?

The distinction between the predictions of the diathesis–stress and the differential susceptibility models are illustrated in Figure 1.[1] The leftmost panel illustrates a prototypical case of diathesis–stress. According to this figure, individuals low on a moderator variable *Z* (e.g., difficult temperament) do not exhibit an association between a focal predictor *X* (e.g., maternal sensitivity) and an outcome variable *Y* (e.g., social competence). People who are high on the moderator variable (*Z*), however, exhibit a positive relation between maternal sensitivity (*X*) and social competence (*Y*). The rightmost panel of Figure 1 illustrates prototypical differential susceptibility predictions. In this example, there is again no association between *X* and *Y* (maternal sensitivity and social competence) among people who have low scores on the moderator variable *Z*. However, there is a positive association between *X* and *Y* among people with high scores on the moderator variable *Z*. In this particular situation, the two regression lines intersect near the middle of the distribution of maternal sensitivity. Thus, individuals with a difficult temperament are *more* well adjusted than others at high levels of parental sensitivity, but they are *less* well adjusted compared to others at low levels of parental sensitivity.

In a series of recent publications (e.g., Belsky et al., 2007; Belsky & Pluess, 2009), Belsky and colleagues have developed criteria for establishing differential susceptibility, and in particular, distinguishing evidence for it from support for other models of development involving statistical moderation, most notably diathesis–stress accounts. These scholars note that making this determination includes, first, conducting a formal statistical test for a crossover (or, more formally, *disordinal*) interaction, and second, comparing the resultant plot with the prototypical ones displayed in Figure 1 (and distinguishing these from other kinds of interactions, including contrastive effects).[2]

1. In all of the discussion below and figures that accompany it, we use the prototypic case where high values on *X* and *Y* indicate high-quality developmental inputs and outcomes, respectively, and high levels of the moderator (e.g., difficult temperament) are hypothesized to reflect high levels of the susceptibility/risk factor.
2. Although differential susceptibility theory requires a significant crossover interaction, it does allow for a significant association between *X* and *Y* at *both* low and high levels of *Z* that varies in magnitude. That is, although Figure 1a (i.e., the strong form of the differential susceptibility hypothesis) suggests that hyposusceptible individuals are not at all affected by a given experience, at least in terms of its impact on a given dependent

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

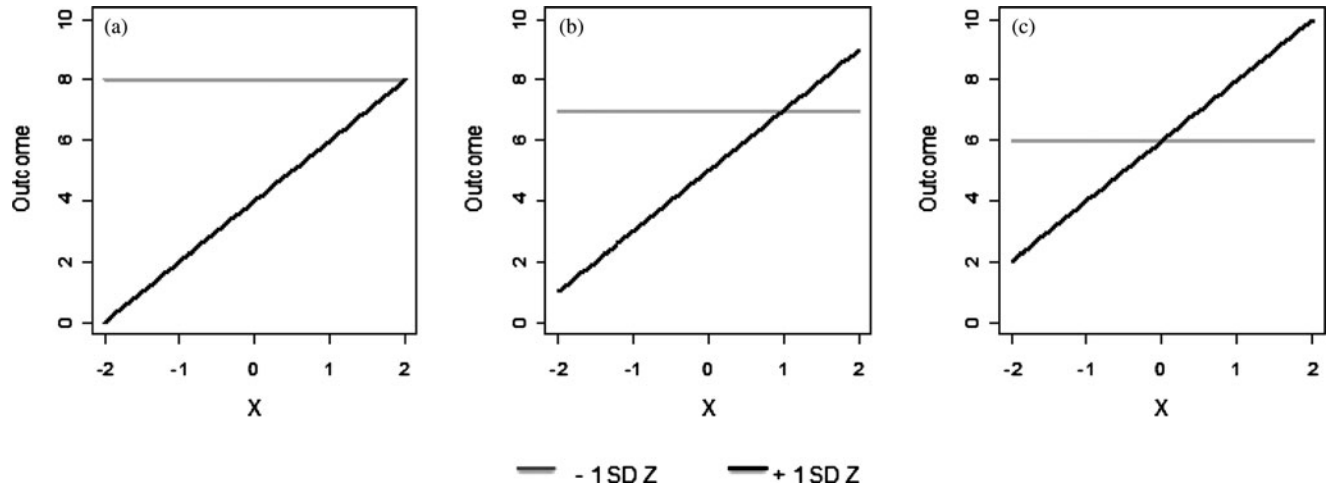



**Figure 1.** (a) Prototypical diathesis–stress and (c) differential susceptibility interaction plots based on *X*–*Y* simple slopes at ±1 *SD* from the mean of Z (moderator). (b) A visually ambiguous plot. Higher values on *X* (standardized) and Y (unstandardized) indicate positively valenced developmental inputs and outcomes, respectively. Higher values on *Z* represent higher levels of susceptibility/risk factor.

Evaluating one's data in light of these criteria can be accomplished in part by following the well-known recommendations of Aiken and West (1991), which involve regressing the dependent measure of interest simultaneously on (centered) independent and moderator variables and their interaction (i.e., the product of the centered predictor and moderator variable). The regression model is

$$Y = b_0 + b_1X + b_2Z + b_3XZ, \quad (1)$$

where *X* represents developmental experiences (e.g., parenting quality), *Z* represents the moderator variable (e.g., temperament), and *Y* represents the outcome variable (e.g., social competence). A significant interaction term ($b_3$) indicates that the association between *X* and *Y* varies as a function of *Z*, but it does not specify the *form* of the interaction. To better understand the form of the interaction it is valuable to plot the regression of *Y* on *X* at various levels of *Z* (e.g., 1 *SD* above and below the mean of *Z*). Traditionally, researchers have supplemented such plots by conducting a s*imple slopes analysis* (e.g., Aiken & West, 1991, p. 78). Simple slopes analysis determines whether the association between the predictor and outcome is statistically significant (and estimates the magnitude and valence of that association) at relatively high or relatively low levels of the moderator (or both; again, by convention, at ±1 *SD* from the mean value of the moderator when *Z* is continuously distributed).[3]

Developmental methodologists (e.g., Dearing & Hamilton, 2006) have recently begun advocating that, instead of conducting simple slopes analysis, in which significant interactions are probed at (for continuous variables) two arbitrarily specified levels of the moderator (i.e., ±1 *SD*), researchers should instead examine *regions of significance* (RoS). In the standard case, RoS, which can be computed using the Johnson–Neyman technique and online calculators that implement it (http://www.quantpsy.org; Preacher, Curran, & Bauer, 2006), refer to the range of values of the moderator for which the independent and dependent variable are significantly associated (what we refer to here as the RoS on *Z*). The potential value of focusing on RoS on *Z* is that these regions highlight *all* possible values of *Z* for which there are significant regressions of *Y* on *X*, instead of probing only two possibilities (i.e., values of *Z* that are 1 *SD* above and below the mean).

Although we agree that the criteria described above for identifying evidence for differential susceptibility are essential,[4] implementing them is less than straightforward

measure (i.e., the correlation between *X* and *Y* at −1 *SD Z* is depicted as zero), it would be in keeping with a differential susceptibility account if the association was also positive for these hyposusceptible individuals but, crucially, smaller in magnitude than for relatively more susceptible individuals. Researchers in this area, however, do not consider *contrastive* interactions—evidence that the valence of the associations between independent and dependent variables runs in the opposite direction at relatively high versus low levels of the moderator variable (i.e., they are the opposite valence and depart significantly from zero)—as consistent with differential susceptibility effects (Belsky et al, 2007).

3. Probing interaction effects involving categorical moderator variables is somewhat more straightforward as predicted linear functions can be generated for each level of the moderator (e.g., for boys and girls if sex is the moderator variable). In this case, there is no need to conduct an RoS on *Z* test (described immediately below in text).
4. Belsky et al. (2007) suggest that three additional criteria be satisfied: (a) the moderator and independent variable should be uncorrelated to rule out, for example, gene–environment correlation; (b) the moderator and the outcome variable should be uncorrelated; and (c) the specificity of the susceptibility (plasticity) factor should be tested by replacing susceptibility factors and outcomes. (a) and (b) strike us as unnecessarily conservative, as a significant interaction term in multiple regression is unique (net) of (a) any association between the independent variable and the moderator and (b) any main effect of either the independent variable or moderator on the dependent measure. When *X* and *Z* are both mean-centered

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

in ways that may not be fully appreciated by developmental psychologists interested in evaluating accounts of differential susceptibility. In light of this, we present four critiques of methods used to identify evidence for differential susceptibility in turn. Our goal in presenting these critiques is to highlight a set of standards by which to distinguish a differential susceptibility effect from (a) a diathesis–stress effect and (b) random data fluctuations. After presenting these critiques, we illustrate how to evaluate the merits of each critique using data from the NICHD SECCYD.

## Critique 1: The Visual Inspection Critique

Much of the research conducted with a focus on examining differential susceptibility has involved simple visual inspection of the interaction plot for evidence of crossover (e.g., Bradley & Corwyn, 2008; Dopkins-Stright et al., 2008). There are two problems with this approach. First, *all* interactions would eventually cross over if extrapolated to a wide enough range of *X*. As such, it is essential that there is strong a priori agreement among researchers about what constitutes the *research range of interest* on the independent (*X*) variable across studies of differential susceptibility (Pedhazur, 1982, p. 461).[5] In other words, there must be consensus on what constitutes the normative range of the independent variable that should be probed. To clarify why this is so critical, consider Figure 2a. A researcher who bounded the range of interest at −1 and +1 *SD* from the mean (dotted vertical lines) on the independent variable would conclude based on visual inspection alone that these (simulated) data were consistent with a prototypical diathesis–stress model. In contrast, a researcher who bounded the range of interest at −2 and +2 *SD* from the mean (solid vertical lines) would interpret the *very same data* as evidence for differential susceptibility. Figure 2b shows that, given a broad enough sampling of the range of the independent variable (however rare some of those values are likely to be actually represented in the population), interactions probably best understood as evidence for diathesis–stress (i.e., using the ±2 *SD* boundaries, solid vertical lines) would resemble a function suggestive of differential susceptibility (i.e., using bounds of ±3 *SD*, dashed vertical lines).

The principal recommendation that follows from the analysis above is that the *same* range of interest must be interrogated across studies of differential susceptibility. It is often difficult to discern what range of interest was used by investigators who plot raw (even if centered) variables, rather than standardized ones, especially if researchers label the *x* axis of interaction figures with terms like "low" and "high" without further explanation, as is commonplace. As such, we endorse the recommended practice (Aiken & West, 1991) of probing interactions from −2 to +2 *SD* from the mean for the independent variable because, given a normally distributed variable, 2 *SD* below and above the mean covers over 95% of the sample.

Assuming that all investigators adopt the same range of interest on *X*, a second problem remains regarding how to objectively probe the interactions to determine whether they are more consistent with diathesis–stress or differential susceptibility models. The problem becomes clear in examining Figure 1b. It is almost certainly the case that, by simply eyeballing this plot, some observers would see this figure as consistent with evidence for differential susceptibility, whereas others would see it as suggestive of diathesis–stress (the simulated plot was designed to be ambiguous in this respect, but is fairly representative of actual interaction effects reported in the literature; see Dick et al., 2011).

In order to counter this level of unacceptable subjectivity, Kochaska et al. (2011) recently suggested that researchers follow the recommendations for probing interactions described in Dearing and Hamilton (2006). Specifically, they endorsed the view that investigators identify what we refer to here as *regions of significance on* X (RoS on *X*), an approach that inverts the RoS on *Z* test to examine the values of *X* for which the moderator and outcome variable are significantly related. The benefit of conducting this analysis is that it provides a test of the statistical significance of the association between the moderator and the outcome at all values of *X* that fall within the range of interest.[6] If the association between the moderator and the outcome is significant at the low but not high end of *X*, there is evidence that the data support the diathesis–stress model (see Figure 3a in which, as is standard practice, regions of significance on *X* in gray are overlaid on a plot of an interaction depicted using simple slopes at −1 and +1 *SD*

(Aiken & West, 1991), *the crossover point will be near the midpoint of* X *whenever* $b_2$ *is near zero*. That is, the plot in Figure 1c (differential susceptibility) is more plausible to the extent that the moderator (*Z*, e.g., temperament) has no main effect on the outcome variable (*Y*). The stronger the main effect of *Z* (e.g., temperament) on *Y*, the less likely the researcher will be able to conclude differential susceptibility (and the more likely an *XZ* interaction will be interpreted as diathesis–stress only). Finally, (c) is not in our view a criterion for ascertaining whether a given interaction effect constitutes evidence of differential susceptibility per se, but rather an examination of its boundary conditions.

5. It is interesting that this is partly what inspired differential susceptibility theory in the first place. Belsky and Pluess (2011) note that, because many psychological variables fail to capture the full range of developmental inputs and outputs, evidence that would seem to support the diathesis–stress model may well mask evidence for differential susceptibility because the range of values of the independent and/or dependent variables might be censored on the positive end of the distribution of these variables. For example, in a given study, the real crossover point might have been realized well within the range of interest, but because the high end of, for example, the *X* distribution was poorly (or not at all) assessed in a given study, the observed (rather than actual) point of crossover made the resultant interaction appear consistent with the diathesis–stress model.

6. In the dichotomous categorical moderator variable case, the RoS on *X* test examines the range of values on *X* within the range of interest for which there is significant spread in the predicted values associated with the two levels of the moderator.

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

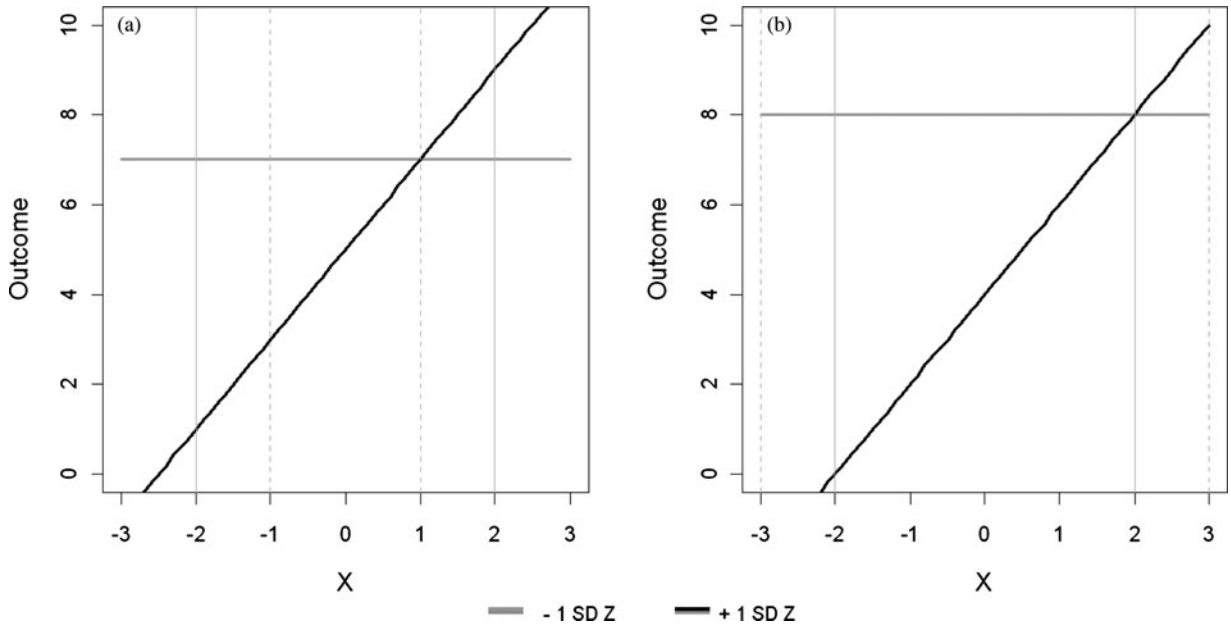



**Figure 2.** An illustration that conclusions regarding the form of an interaction vary as a function of the range of interest on the *X* variable. Dashed and solid vertical lines denote investigator-specified choices about ranges of interest on *X* (see text).

on a continuously distributed moderator variable). If the association between the moderator and the outcome is significant at both the low *and* high ends of the distribution of the independent variable within the normative range (i.e., ±2 *SD*), there is evidence of differential susceptibility (see Figure 3b; see Kochanska et al., 2011, for an early example of the implementation of the RoS on *X* test applied to the differential susceptibility hypothesis).

Although there is emerging awareness of RoS on *X* analyses among differential susceptibility researchers (e.g., Belsky & Pluess, in press; Kochanska et al., 2011), visual inspection of interaction plots to identify evidence for differential susceptibility remains widespread in recent publications (e.g., Pluess & Belsky, 2010). We recommend that researchers discontinue this practice. Moreover, when RoS on *X* analyses are used, it is crucial that the same *range of interest* is

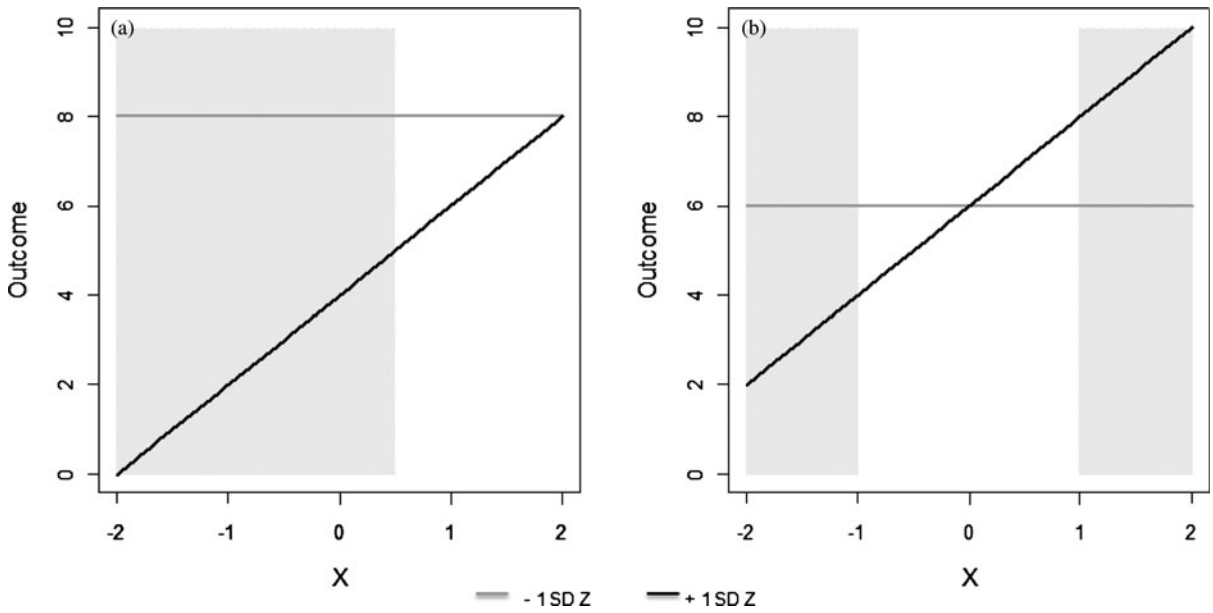



**Figure 3.** Regions of significance (RoS) on *X* for (a) diathesis–stress and (b) differential susceptibility models. Gray shaded area denotes regions where the two lines significantly differ.

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

examined across all studies of differential susceptibility, which we recommend should be bracketed by ±2 *SD* from the mean of *X*.

*Critique 1:* Visually apparent differential susceptibility evidence in an interaction plot may not be statistically distinguishable from data generated by a diathesis–stress model, a problem that is compounded if different researchers choose to adopt different ranges of interest on the X variable under investigation.

*Proposed Critical Test 1*: Researchers should apply a Regions of Significance (RoS) on *X* test (Dearing & Hamilton, 2006; Kochanska et al., 2011) to demonstrate that *Y* (e.g., social competence) and *Z* (e.g., difficult temperament) are correlated both at the high and low ends of the distribution of *X* (e.g., maternal sensitivity) bounded by a conventional range of interest that is ±2 *SD* from the mean of *X*.

## Critique 2: The Quantification Critique

Although the RoS on *X* analysis would appear to make differentiating between differential susceptibility and diathesis–stress more objective when faced with the ambiguity of, say, the plot depicted in Figure 1b, an underappreciated problem with this approach that flows from null hypothesis significance testing generally is that the test in not invariant to sample size (Cohen, 1988; Fraley & Marks, 2007). For example, in the relatively small sample size case it will be quite common to observe plots that would seem highly consistent with a differential susceptibility account (i.e., that look much like Figure 1c) but, due to low statistical power, might be interpreted (wrongly) as evidence for diathesis–stress.

What is potentially more concerning, however, is the problem that can emerge in a large *N* study. The problem in this empirical context is that even trivial evidence for a crossover interaction at the highest values of the predictor variable within the range of interest will eventually be detected as statistically significant as sample size increases. What this means is that, with enough *N*, *any* interaction that features some crossover at the high end of the predictor distribution—however trivial the magnitude of this difference—will be viewed through the prism of RoS on *X* testing as consistent with the differential susceptibility account. To be clear, large *N* designs are valuable because they allow theoretical parameters to be estimated with high precision. In the context of null hypothesis significance testing, however, they can be abused to highlight trivial differences that may actually have no practical or theoretical significance.

### *Two novel metrics for quantifying evidence for differential susceptibility*

To avoid sample size driving the interpretation of interaction effects, we offer here two novel solutions to distinguish between differential susceptibility and diathesis–stress that can be used to supplement the RoS on *X* approach (particularly in the large *N* empirical context). Both proposed metrics are unitless, and they are also independent of sample size. The first of these involves creating an index[7] of the degree to which an interaction conforms to the prototypic one for differential susceptibility. One straightforward way to accomplish this is to measure the proportion of the total area represented in the relevant figures (see Figure 4a and b) uniquely attributable to differential susceptibility. We refer to this proposed metric, $b/(b + w)$, as the proportion of interaction (PoI) index.

Both the differential susceptibility and diathesis–stress accounts anticipate the area labeled *w* (i.e., for *worse*) in Figure 4a and b. That is, both the differential susceptibility and diathesis–stress accounts require that individuals who experience poor developmental inputs (i.e., low levels of maternal sensitivity) and who are otherwise at risk (i.e., due to difficult temperament) should fare less well than individuals with only one of those risk factors. What the differential susceptibility hypothesis uniquely predicts, however, is that it will also be the case that the interaction effect will be driven by the area of the plotted interaction that we label *b* (for *better*) in Figure 4b, which is evidence that (for reasons previously explained) individuals who have the same putative risk factor (difficult temperament in this example) actually should be expected to be *more* successful than their temperamentally easy counterparts given positive developmental inputs. Furthermore, prototypic differential susceptibility and diathesis–stress plots allow us to place a point prediction on the expected values of the areas of *w* and *b*. Specifically, *w* should be equal to 100% in the prototypic case of diathesis–stress (i.e., $b/[b + w]$ should equal 0.00), whereas *w* and *b* should both equal 50% in the prototypic case of differential susceptibility (i.e., $b/[b + w]$ should equal 0.50).[8] (This assumes that, once *X* is standardized, a value of zero anchors the true mean for the construct of interest tapped by the mea-

7. The metrics we propose below should not be confused with the true effect size associated with the interaction term: the change in $R^2$ associated with the additional prediction of the dependent variable over and above the main effects of *X* and *Z*, which we report in Table 1 (see also Aguinis, Beaty, Boik, & Pierce, 2005, p. 96; Aiken & West, 1991, p. 157). Instead, PoI and PA should be understood as quantifying the degree to which a given interaction effect resembles evidence for differential susceptibility over diathesis–stress, regardless of how large the overall interaction effect is (in the present case, the significant interaction effects we identified were fairly modest in magnitude, explaining in general less than 1% of the variance in the dependent measures). Our metrics also require that the researcher first rule out contrastive interaction effects.
8. As a rough marker, values on the PoI metric between about 0.40 and 0.60 represent an interaction effect highly consistent with differential susceptibility. As PoI approaches 0.00, this suggests strong evidence for diathesis–stress. As PoI approaches 1.00 it suggests strong evidence for what might be considered a counterintuitive diathesis–stress model—one in which high levels of the putative risk factor, combined with positive developmental inputs, are especially promotive of development, but that such a putative risk factor provides no role in amplifying risk given poor developmental inputs.

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

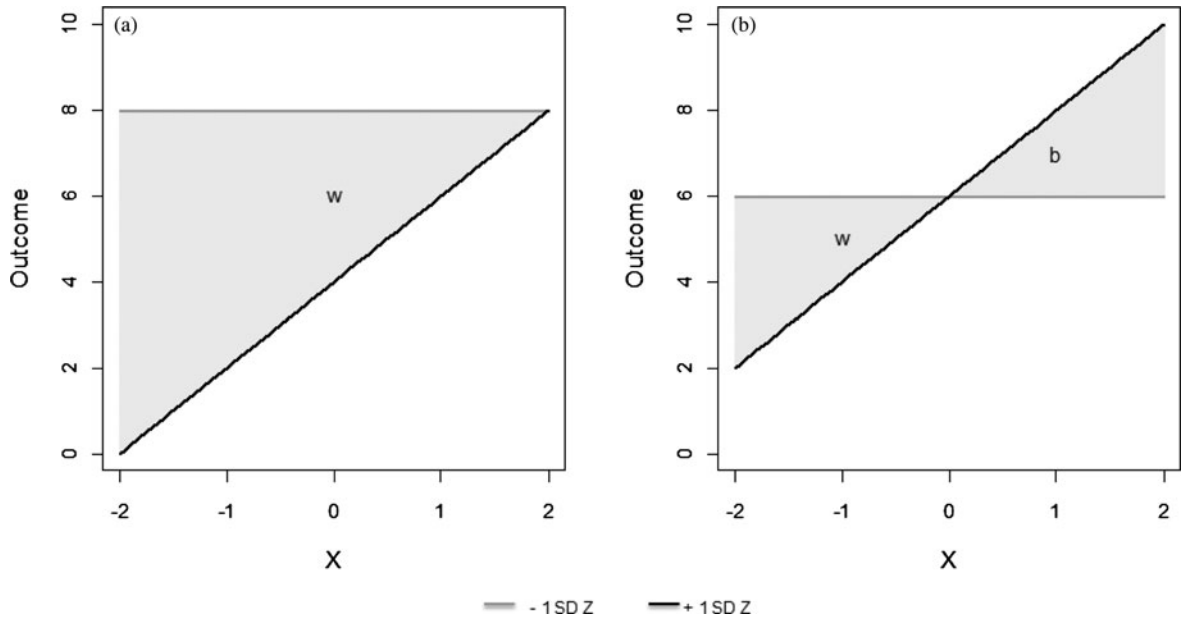



**Figure 4.** Proportion of the interaction (PoI) for (a) diathesis–stress and (b) differential susceptibility models. w, worse; b, better.

sure of *X*; see footnote 5 and the Discussion Section for elaboration on this important issue.[9]

One of the major strengths of the PoI metric for quantifying evidence for differential susceptibility (and diathesis–stress) is that, unlike the *p* values used in RoS tests, PoI is not an explicit function of sample size. However, like regions of significance, it will produce different results if different researchers define their ranges of interest using different criteria (e.g., the results for $\pm 1$ *SD* differ from those produced using $\pm 2$ *SD*). Thus, we recommend that researchers consistently probe $\pm 2$ *SD* on the *X* variable so the PoI metric requires no additional assumptions beyond those already in use when plotting interactions generally.

*Critique 2:* There is no standard index for the quantification of differential susceptibility (or diathesis–stress) effects. Thus, inferences of diathesis–stress versus differential susceptibility are driven largely by statistical significance testing and sample size using the RoS on *X* test or, more commonly, visual inspection of the interaction plot.

*Proposed Critical Test 2a:* In addition to RoS on *X* testing, researchers should apply a diathesis–stress/differential susceptibility index, PoI, to help distinguish differential susceptibility evidence from diathesis–stress evidence. PoI values close to 0.50 suggest strong evidence for differential susceptibility. Values closer to 0.00 suggest strong evidence for diathesis–stress.

An alternative to the PoI index, which is one that avoids the need to specify a range of interest altogether, is to estimate the proportion of the population that is differentially affected by the moderator: the proportion affected (PA) index. The logic of this approach is relatively straightforward. The goal is to identify first the value of *X* at which the regression lines cross over. The crossover value is located at

$$\text{crossover point} = -b_2/b_3, \tag{4}$$

where $b_2$ is the semipartial main effect of *Z* on *Y* from Equation 1, and $b_3$ is the interaction effect from Equation 1.

The crossover point is the threshold at which the association between *Z* and *Y* changes meaning as a function of *X*.

9. At this point, it is worth mentioning that the PoI index, $b/(b + w)$, is a ratio of *improved* outcomes for difficult temperament ($b$ = better) over the sum of improved outcomes ($b$ = better) and harmful outcomes ($w$ = worse; see Figure 4). This index will depend upon the range of *X* (e.g., maternal sensitivity) that the researcher arbitrarily chooses (i.e., how $X_{low}$ and $X_{high}$ are defined). It is of interest that we can demonstrate mathematically that the new effect size index works out to be related to the squared ratio of the *Z*–*Y* simple slope at $X_{high}$ (e.g., conditional effect of temperament at high sensitivity) to the *Z*–*Y* simple slope at $X_{low}$ (e.g., conditional effect of temperament at low sensitivity)

$$\text{PoI} = \frac{b}{(b+w)} = \left(1 + \frac{(b_3 X_{low} + b_2)^2}{(b_3 X_{high} + b_2)^2}\right)^{-1}, \tag{2}$$

where $(b_3 X_{high} + b_2)$ and $(b_3 X_{low} + b_2)$ are the simple slopes of *Y* on *Z* at high and low *X*, respectively.

Another way to index the beneficial component of differential susceptibility is to express the region *b* (see Figure 4) in a straightforward fashion, without reference to region *w*. This option would amount to a squared standardized simple slope of *Y* on *Z*, at high *X*.

$$\text{betterness index} = (b_3 X_{high} + b_2)^2 (s_z^2/s_y^2), \tag{3}$$

where $(b_3 X_{high} + b_2)$ is still the simple slope of *Y* on *Z* at high *X* (e.g., 2 *SD* above the mean of *X*), and $(s_z^2/s_y^2)$ is the ratio of the variance of *Z* over the variance of *Y*.

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

Thus, the number of people who are differentially affected (i.e., the proportion affected or the PA index) can be computed by sorting the dataset with respect to *X* and calculating the proportion of cases that fall above the crossover point. If the number/proportion of people above the crossover exceeds a prespecified threshold, then one concludes in favor of differential susceptibility. If the proportion of people above the crossover is less than that threshold, one concludes in favor of diathesis–stress.

The potential value of the PA index is that it offers a pragmatic way of evaluating evidence for differential susceptibility. Namely, it could be argued that, even if there is evidence for differential susceptibility according to other criteria (e.g., RoS on *X* tests), the model is of limited use if only a small number of individuals experience the theorized benefits. What threshold should be used? Typically, the PA index will range from around 0.50 (indicating strong evidence for differential susceptibility) to 0.00 (in the clear case of diathesis–stress). However, in the interest of establishing default conventions, we suggest that if only 16% of cases fall above the point at which the regression lines intersect, investigators should begin to question whether their data are consistent with differential susceptibility theory; when less than 2% of cases fall above the point at which the regression lines intersect, such evidence should definitely not be interpreted as evidence for differential susceptibility.[10] Although the selection of any threshold is necessarily arbitrary, our justification for these particular thresholds is that approximately 16% and 2% of cases will fall 1 and 2 *SD*, respectively, above the mean on a normal distribution.

*Proposed Critical Test 2b:* Researchers should estimate the percentage of the population above the crossover point. Larger percentages suggest stronger evidence for differential susceptibility relative to diathesis–stress.

## Critique 3: The Type I Error Critique

### *Capitalizing on chance?*

When testing a differential susceptibility hypothesis, it is not uncommon for researchers to have access to data on many different variables that might alternatively be used to operationalize *X*, *Z*, and *Y*. In such instances, it can be possible to conduct many, separate tests for differential susceptibility. For instance, if there are 2 *X* variables (e.g., two different measures of parental sensitivity), 2 *Z* variables (e.g., two different measures of temperament), and 15 *Y* variables (e.g., three different outcome measures, *Y*, each taken at five different points in time), then a researcher might ostensibly conduct $2 \times 2 \times 15 = 60$ separate tests of the differential susceptibility effect. With so many tests, the probability of committing a Type I error (i.e., concluding the effect exists when, in fact, it does not) for any single test of differential susceptibility (or diathesis–stress) is quite high. For example, at a nominal a level of 0.05 (5% Type I error rate) for each test, the familywise Type I error rate for falsely rejecting the null hypothesis for *any one of* 12 independent significance tests is $1 - (1 - 0.05)^{12} = 46\%$ (Sidak, 1967), which much larger than 5%. Because the significance tests for 60 different operationalizations of the interaction effect (described above) involves many nonindependent tests, it can be complicated to estimate the exact overall Type I error rate across multiple comparisons (see Benjamini & Yekutieli, 2001). Nevertheless, we note that Type I error rate is extremely high in traditional tests of differential susceptibility and diathesis–stress. This issue is potentially exacerbated in *developmental* tests of differential susceptibility and diathesis–stress, because the same outcome variable is often measured longitudinally—creating a strong possibility that differential susceptibility may appear at one point in time, but not at others.

### *True developmental sensitive periods?*

One issue that applies to the evidence for differential susceptibility and diathesis–stress is that of the relative inattention in this literature to date on development itself (but see Belsky & Pluess, in press). In a typical study in this literature, caregiving experiences (e.g., maternal sensitivity) and a putative plasticity agent (e.g., difficult temperament) are measured at some point early in the life course and the correlates of those experiences are assessed at another point later in time. The caveats we have thus far offered aside, if the predictive significance of those early experiences are significantly moderated by the plasticity factor in a manner resembling Figure 1c, researchers conclude that they have documented evidence for differential susceptibility (conclusions that as noted earlier have often not involved the application of the RoS on *X* test; see, e.g., Pluess & Belsky, 2010).

10. We feel the need to elaborate on the rationale behind the 16% cutoff for the PA index. Undeniably, if one can rule out sampling error (e.g., with null hypothesis significance testing, using RoS on *X*), any evidence for (noncontrastive) crossover in the range of interest on *X* is consistent with individuals +1 *SD* on Z being accurately described as "differentially susceptible," even if the crossover occurs at the very high end of the distribution of *X* (e.g., if Figure 1b was based on the entire population of interest, rather than the intended sample drawn from that population, it provides clear evidence of differential susceptibility). However, two other judgments must be brought to bear even when sampling error is no longer an issue. The first is whether the magnitude of the crossover in the range of interest on *X* is trivial, which our PoI and PA metrics are meant to quantify. The second is the question of what differential susceptibility *theory*, as articulated by Belsky and colleagues, predicts. It is the interpretation of the first author of this report that differential susceptibility theory, at least in its strong form, makes the prediction that PA should be equal to 50%: the crossover should occur at the population mean of *X* (see footnote 5). By this logic, a crossover at 1 *SD* (and certainly 2 *SD*) above the mean of *X* departs notably from this prediction and suggests that differentially susceptible individuals only stand to benefit when the quality of their caregiving experiences is extraordinarily positive in nature. Nonetheless, we acknowledge there are other interpretations that could be offered for the range of PA values that would be compatible with differential susceptibility theory.)

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

However, what may not be obvious in such analyses is that such effects may either fade or endure over time (for a broader discussion of this problem, see Fraley, Roisman, & Haltigan, in press). What seems especially problematic is that even when multiple assessments of a given dependent variable across time and informant are available (e.g., in longitudinal studies such as the NICHD SECCYD), most studies of differential susceptibility tend to be focused on the way the predictive significance of early caregiving experiences is moderated by difficult temperament in relation to a set of outcomes all of which are assessed at a single assessment, sometimes by the same informant (but see Belsky & Pluess, in press; Pluess & Belsky, 2009, 2010). Dopkins-Stright et al. (2008), for example, found evidence that individuals with difficult temperament in the NICHD SECCYD were more susceptible to the effects of early maternal sensitivity in the sense that they benefitted more from high quality caregiving and were disproportionately negatively affected by insensitive caregiving than those low on temperamental difficulty. However, all of the outcomes examined in that report were assessed at a single assessment (Grade 1) by the same informant (i.e., the participants' teachers).

What might be discovered by examining the same outcome variables available on the SECCYD in multiple assessments across time and informant through midadolescence? In advance of examining this question empirically, it is impossible to know. However, it is possible that the evidence for differential susceptibility might have been transient. One way to examine this empirically is to examine how the interaction effect varies as a function of when the outcome variable in question was assessed. Second, assuming evidence for moderation is consistent over time, the potentially dynamic shape of significant interaction effects can be evaluated by plotting the PoI or PA metrics described above as a function of the age at which the outcome was assessed. For example, if the PoI metric tends to be close to a value of 0.50 and relatively stable over time, this would provide compelling evidence that early temperament is a marker of differential susceptibility to early caregiving experiences, in a way that is reflected in an *enduring* manner in developmental adaptation. The presentation of all available dependent data in longitudinal studies also helps rule out the possibility that evidence for differential susceptibility or diathesis–stress at any given assessment point is merely Type I error.

Moreover, two methodological protections against developmental Type I error (i.e., finding statistically significant effects for an outcome measured at one point in time, but not for outcomes measured at others) are also available. First, when the researcher is interested in a true developmental sensitive period (i.e., a true effect that obtains at some time points but not others) then it may be valuable to apply a more conservative a level (e.g., Dunn, 1961; Sidak, 1967) to rule out the possibility that this particular time period appears significant by chance alone. We recommend a sequential Bonferroni-type procedure (Benjamini & Hochberg, 1995; Simes, 1986; see Appendix A), which controls the false discovery rate but can be considerably more powerful than the traditional Bonferroni procedure. Second, when the researcher is interested in differential susceptibility in general (at the potential cost of ignoring any developmental patterns over time), then we recommend aggregating outcome variables across time, via a multilevel model (intercepts as outcomes model; see Pluess & Belsky, 2010; Raudenbush & Bryk, 2002).

*Critique 3:* Differential susceptibility effects that appear in only a few out of many statistical tests attempted are likely the result of Type I error (i.e., a high probability of getting at least one false positive result when many tests are conducted).
*Proposed Critical Test 3a:* Researchers should apply multilevel models to determine overall differential susceptibility for outcomes, averaged across the entire developmental course (i.e., using intercepts as outcomes models).
*Proposed Critical Test 3b*: If an apparent differential susceptibility effect emerges at some points in time but not at others, then researchers should apply a more conservative procedure for Type I error control (Benjamini & Hochberg, 1995).

At this point, an important caveat is in order concerning Critical Test 3b. That is, one possible reason for an interaction effect to emerge at one point in time but not at others could be the existence of interindividual differences in intraindividual change (i.e., linear and quadratic slope effects over time; Baltes, Reese, & Nesselroade, 1977; Raudenbush & Bryk, 2002). As an example of this, if the Sensitivity × Temperament interaction effect gets systematically stronger at later time points, this might reflect a positive association between the interaction term and the linear slope of the outcome variable. As such, in addition to Critical Test 3b, researchers should also inspect whether the interaction terms, PoI, and PA indices follow a systematic pattern (e.g., monotonically rising or declining).

## Critique 4: The Nonlinearity Critique

Finally, we note that apparent differential susceptibility effects might logically be an artifact of imposing a linear model on a plausibly nonlinear phenomenon. This is illustrated in Figure 5. In this figure we observe a diathesis–stress effect in which there is a *nonlinear X–Y* relationship. Suppose for the sake of the argument that there are diminishing returns to parental sensitivity ($X$), such that children benefit *more* from a unit increase in caregiver sensitivity when sensitivity is scarce than when sensitivity is more available (i.e., the $X$–$Y$ relationship is stronger at lower levels of $X$). In addition, we might reasonably suppose, as is consistent with the diathesis–stress perspective, that any nonlinear $X$–$Y$ effect (where low parental sensitivity becomes increasingly harmful when sensitivity is already scarce) is especially likely to emerge among children with difficult temperaments (i.e., high $Z$).

The point of Figure 5 is that, if such a nonlinear phenomenon existed, imposing a linear predictor model on the data (as is standard; see the dashed lines in Figure 5) might lead

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

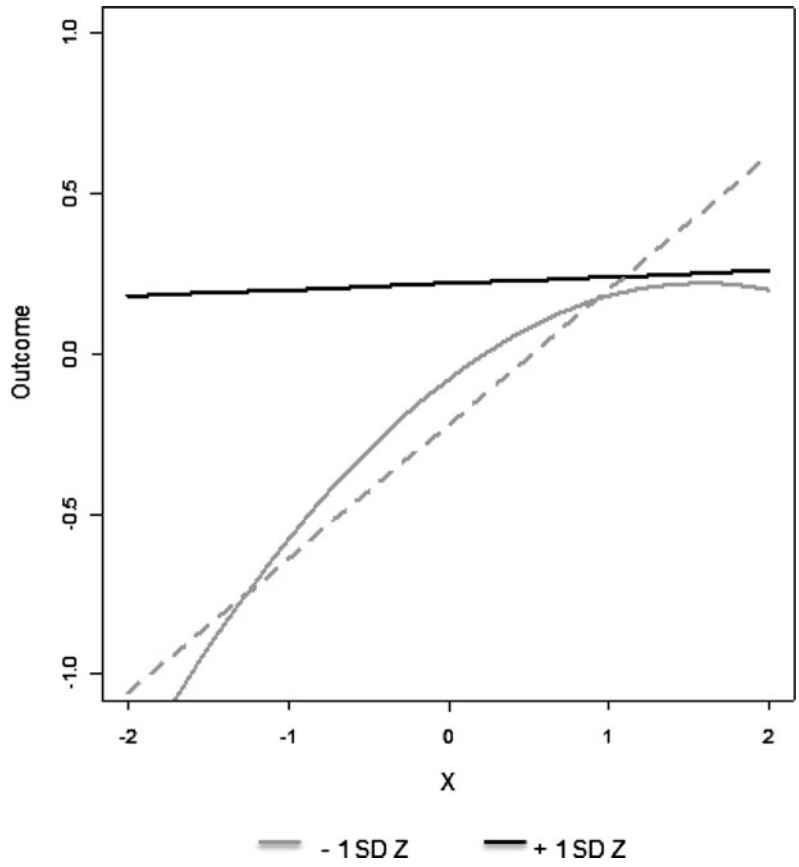



**Figure 5.** Nonlinear relations masquerading as differential susceptibility. The solid lines represent predictions from a nonlinear model. The dashed lines represent predictions from a linear model. (Note the linear and nonlinear lines for +1 *SD* Z overlap.)

researchers to observe a differential susceptibility effect that does not really exist. To test this possibility, it would be necessary to estimate differential susceptibility while demonstrating the absence of nonlinear terms $X^2$ and $ZX^2$ (see Cortina, 1993).

*Critique 4:* Apparent differential susceptibility effects can be an artifact of imposing a linear predictor model on a nonlinear diathesis–stress phenomenon.
*Proposed Critical Test 4:* When testing differential susceptibility, researchers should estimate an additional model that includes $X^2$ and $ZX^2$. The results of this model must show that neither $X^2$, $ZX^2$, nor the combination of both $X^2$ and $ZX^2$ together is statistically significant. If one of these nonlinear $X^2$ terms is significant, then the researcher must demonstrate that the *XZ* interaction term remains statistically significant even after controlling for the nonlinear terms.

## An Empirical Example in the NICHD SECCYD

To provide a demonstration of how the substantive conclusions in this area of research are potentially subject to change as a function of the critiques provided above, in the remainder of this report we revisit the literature on difficult temperament as a susceptibility factor in relation to early maternal sensitivity using data drawn from the NICHD SECCYD. To our knowledge, three published studies have highlighted the role of early difficult temperament (assessed using a maternal report questionnaire at 1 and/or 6 months in all three studies) in amplifying the predictive significance of early observations of maternal sensitivity, for better and for worse, in the SECCYD dataset. Dopkins-Stright et al. (2008) focused on academic competence, social skills, and relationships with teachers and peers measured during first grade. Bradley and Corwyn (2008) examined externalizing problems, also assessed by teachers in first grade. Finally, and most recently, Pluess and Belsky (2010) focused on assessments of cognitive–academic and social skills. Data presented in each of these studies were interpreted as supporting the differential susceptibility model over the diathesis–stress account, although none used the RoS on *X* test as a basis for that conclusion (but see Belsky & Pluess, in press).

The present study revisits and extends these recent reports, as well as our own analyses of the NICHD SECCYD (Fraley et al., in press; Haltigan, Roisman, & Fraley, in press) in which we have found that early observed maternal sensitivity assessed through 36 months is associated in an enduring manner with academic skills (teacher reported through Grade 6 and based on objective assessments through age 15), social competence (mother reported through age 15 and teacher

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

reported through Grade 6), and teacher-reported symptoms of psychopathology (through Grade 6). (The predictive significance of early maternal sensitivity for *mother* reports of symptomatology was demonstrated in this work to fade over time in the SECCYD through age 15; that is, early sensitivity was an increasingly weaker predictor of mother reports of symptomatology as children in the cohort aged; Haltigan et al., in press).

More specifically, we apply the methods described above to examine whether early difficult temperament reliably moderates the association between early observed sensitivity and developmental adaptation over time within and across the key domains of (mal)adaptation described above in the SECCYD. Along the way, we seek to assess whether the differential susceptibility hypothesis can survive the tests we articulated above. We present the results of RoS on *Z* and RoS on *X* tests, novel differential susceptibility/diathesis–stress indices PoI ($b/[b + w]$) and PA, significance tests with conservative critical $\alpha$ levels and multilevel intercepts as outcomes models, and nonlinear predictor models, to quantify the evidence for differential susceptibility versus diathesis–stress. We will consider if and how such metrics change as a function of when the dependent measures were assessed in order to explore the developmental course of these moderated associations. In addition to providing new data on the SECCYD cohort (i.e., through age 15, where applicable), this analysis provides a representative test of the differential susceptibility and diathesis–stress hypotheses in the SECCYD in relation to early maternal sensitivity, early difficult temperament, and a fairly comprehensive range of markers of developmental (mal)adaptation.

## Method

### *Participants*

Families were recruited for the NICHD SECCYD in 1991 from hospitals located in or near Little Rock, Arkansas; Orange County, California; Lawrence, Kansas; Boston, Massachusetts; Pittsburgh, Pennsylvania; Philadelphia, Pennsylvania; Charlottesville, Virginia; Seattle, Washington; Morganton, North Carolina; and Madison, Wisconsin. During selected 24-hr sampling periods, 8,986 women who gave birth were screened, 5,416 of whom met the eligibility criteria for the study. Families were excluded if (a) the mother was younger than 18 years of age, (b) the family planned to move, (c) there was a multiple birth, (d) the infant had a known disability or remained in the hospital more than 7 days, (e) the mother acknowledged substance abuse, (f) the mother did not speak English, and/or (g) the mother lived more than 1 hr from the laboratory site or in an extremely unsafe neighborhood, as determined by local police. From that group, 1,364 families became study participants upon completing an interview when their infants were 1 month old. Details about recruitment and selection procedures are available in prior publications from the study (see NICHD ECCRN, 2005) and the Web (http://secc.rti.org). Note that, although large, demographically diverse, and methodologically rich, the SECCYD was not designed to be a nationally representative study.

*Analysis sample.* The analysis sample for this study included the 1,306 children (52% males) for whom any observational data were collected at ages 6, 15, 24, and/or 36 months that focused on the quality of the maternal caregiving they experienced in the first 3 years of life. In terms of race/ethnicity, 77% of the children from the analysis sample for this study were White/non-Hispanic. Note that for all modeling analyses described below, multiple imputation ($m = 20$ imputations) was used with raw case-level analytic data as input, to produce less biased and more efficient and consistent parameter estimates than techniques such as pairwise or listwise deletion for longitudinal missing data (Newman, 2003; Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002). These analyses mirror the multiple imputation approach used by Pluess and Belsky (2010). Further, we followed Collins, Schafer, and Kam's (2001) inclusive strategy of using all study variables in the imputation model, rather than just a subset of variables (see Graham, 2003), as well as Enders' (2010) and Graham's (2009) recommendation to incorporate the interaction term into the imputation model when testing moderated effects, to avoid a known problem where the linear imputation model biases interaction effects toward zero. Predictor variables were standardized using the expectation–maximization algorithm (maximum likelihood) means and standard deviations, as advised by Enders (2010). Moderation analyses were performed in SAS 9.1 Proc MI, Proc Reg, and Proc Mixed; syntax files are available from the second author. RoS on *Z*, RoS on *X*, PoI, the crossover point for the interaction, and the PA indexes were generated using a Web-based program developed by the third author. The application is freely available at http://www.yourpersonality.net/interaction/ and can also be used to generate and explore interaction plots and to conduct simple slope tests.

### *Measures*

Measures are presented in four sets corresponding to their function in the analytic plan, as follows: (a) variables used to create a composite measure of the observed quality of participants' experience with caregivers in the first 3 years of life; (b) measures of social competence, as assessed by primary caregivers and teachers through the latest assessment for which these data were available (mother: age 15; teacher: Grade 6); (c) measures of academic skills, as rated by teachers and assessed using standardized tests through the latest assessment for which relevant data were acquired (teacher reports: Grade 6; standardized assessments: age 15); and (d) reports of total problem behavior, as assessed by primary caregivers and teachers from the first through the latest assessment for which these data were available (mother: age 15; teacher: Grade 6). In all cases we selected variables that

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

were measured multiple times by multiple reporters using standard assessment tools.

*Early maternal sensitivity.* Mother–child interactions were videotaped during 15-min semistructured tasks at 6, 15, 24, and 36 months. A number of scales were used to rate the mothers' behavior from these videotapes. More specifically, at 6 months, mothers and children were instructed to play together, first with toys available in the home (or none at all) and then with a standard set of toys. At 15, 24, and 36 months, mothers were asked to show their children age-appropriate toys in three containers in a set order. As in prior studies of this sample (e.g., NICHD ECCRN, 2001), observations of maternal sensitivity from the first 3 years of life (6, 15, 24, and 36 months) were standardized and averaged to create a composite of the *observed early sensitivity*. Note that at 6, 15, and 24 months, the a priori maternal sensitivity composites were constructed by summing ratings for sensitivity to nondistress, positive regard, and intrusiveness (reversed). At 36 months the supportive presence, respect for autonomy, and hostility (reversed) scales were composited (as reported in NICHD ECCRN, 2001, internal consistencies of composites were 0.75, 0.70, 0.79, and 0.78 for the 6, 15, 24, and 36 month composites, respectively, and intercoder reliabilities on scales >0.80). Within age composites showed stability over time (*r*s range = .30–48; standardized α for composite measure = 0.73).

*Early temperament.* At the 6-month visits, mothers completed the Carey and McDevitt (1978) Infant Temperament Questionnaire, which contained 38 items rated on a 6-point scale. The composite total score, with higher scores reflecting a more irritable and difficult child, had adequate reliability (α = 0.81). We selected the 6-month Infant Temperament Questionnaire data in the current analyses rather than a composite of the 1- and 6-month data as has been done in at least one prior investigation of differential susceptibility (i.e., Bradley & Corwyn, 2008) based on both psychometric considerations: the 1-month temperament composite possessed only modest reliability relative to the 6-month temperament composite (α = 0.67 vs. α = 0.81, respectively; Roisman et al., 2010) and the 1- and 6-month temperament composites are only modestly correlated (*r* = .32; Roisman, Susman, et al., 2009). We also selected it because Pluess and Belsky (2010) explicitly noted in a prior examination of the 1- and 6-month temperament data with respect to differential susceptibility that "temperament scores at 6 months were not averaged with a similar measure at 1 month because preliminary hierarchical linear model analyses showed that the 6-month measure yielded stronger interaction effects than the average of the two measurement occasions" (p. 381).

*Total problem symptomatology.* Symptoms of psychopathology from childhood to adolescence were assessed with the total problem scale of the Child Behavior Checklist (CBCL) obtained using the parent and Teacher Report Form (TRF) versions (Achenbach, 1991; Achenbach & Edelbrock, 1986; Achenbach, Edelbrock, & Howell, 1987). We used *T* scores. In the current study, maternal reports on the CBCL were used from the following assessment points: a mean composite of the 24- and 36-month CBCL ($r = .73$; $p < .01$); 54 months; kindergarten; Grades 1, 3, 4, 5, and 6; and age 15. Teacher reports were used from the following assessment points: kindergarten and Grades 1, 2, 3, 4, 5, and 6. The total problem behavior scale showed adequate reliability across time and had a coefficient α averaging 0.93 for maternal reports and 0.96 for teacher reports across all assessments.

*Social competence.* Mothers completed the 38-item Social Skills Questionnaire from the Social Skills Rating System (SSRS; Gresham & Elliott, 1990) when children were age 54 months and in kindergarten; Grades 1, 3, 4, 5, and 6; and at age 15, indexing general social competence with adults and other children. To obtain a standardized measure of *total social skills*, an a priori scale was created by summing all items that index social competence with other children, with higher scores indicating more socially skilled children (αs range = 0.87–0.91; $M = 0.89$). In a parallel fashion, teachers completed the 30-item school version of the Social Skills Questionnaire from the SSRS when children were in kindergarten as well as Grades 1, 2, 3, 4, 5, and 6. As with the mother reports, a standardized *total social skills* scale was created by summing items indexing social competence (αs range = 0.93–0.94; $M = 0.94$).

*Academic skills.* Teachers also used the SSRS to rate participants' academic skills when children were in kindergarten and Grades 1, 2, 3, 4, and 5 (αs range = 0.94–0.95; $M = 0.95$). For use in this study, an a priori, standardized academic competence scale was created by summing items indexing academic success. In addition, in order to examine a relatively objective assessment of academic skills we used Woodcock–Johnson Psycho-Educational Battery—Revised (WJ-R; Woodcock, 1990; Woodcock & Johnson, 1989) scores at 54 months; Grades 1, 3, and 5; and age 15. Note that for the WJ-R, a slightly different subset of scales was used at each assessment point. For purposes of this analysis, we averaged the *W* (standard) scores for all available subscales at each time point (within time αs range = 0.81–0.91; $M = 0.87$).

*Control variables.* Although there are a large number of potential control variables to consider, we selected four that we have examined in previous research (e.g., Fraley et al., in press; Roisman, Booth-LaForce, et al., 2009; Roisman, Susman, et al., 2009) and which are known to correlate with maternal sensitivity and the outcomes examined in this report: child gender, child ethnicity, maternal education, and family income to needs. Child gender was coded in a binary fashion (1 = *female*, 0 = *male*). Because the majority of the children in the sample were White/non-Hispanic, we created a binary variable to represent ethnicity (1 = *White/non-Hispanic*,

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

0 = *otherwise*). Maternal education was coded on an ordered metric representing the number of years of education/highest degree. Family financial resources were operationalized as an income to needs ratio. Income to needs was computed separately within each of four assessment waves (6, 15, 24, and 36 months) and averaged to create a mean income to needs index for early childhood. Each of these variables was entered into all analyses reported below.

## Results

Results of the tests of differential susceptibility and diathesis stress are presented in Table 1. For each outcome variable (e.g., mother-reported CBCL, teacher-rated social competence), we first report the results for a regression model at each point in time (54 months, kindergarten, Grade 1, etc.), followed by the overall (mean) effect from an intercepts as outcomes multilevel model (i.e., a random intercepts model in which the intercept [average over time] of the outcome is treated as the dependent variable). These results in Table 1 show (a) whether the Parental Sensitivity × Difficult Temperament interaction effect is statistically significant; (b) whether the RoS on parental sensitivity (RoS on $X$) is within the bounds of −2 to +2 *SD* (Critical Test 1); (c) how large the differential susceptibility/diathesis–stress indices are (PoI and PA; Critical Tests 2a and 2b); (d) whether the effects emerge across the entire developmental course, on average (Critical Test 3a, intercepts as outcomes); (e) whether the effect observed at a single point in time is likely due to familywise Type I error (Critical Test 3b; Benjamini & Hochberg test); and (f) whether $X^2$ and $ZX^2$ nonlinear terms are statistically significant, and, if so, whether the $XZ$ interaction term remains statistically significant in the model with significant nonlinear predictor terms (Critical Test 4). The results of these diathesis–stress/differential susceptibility tests are reported below, organized with respect to the outcome variables.

### *Maternal reports of total problems*

Regarding the possibility of differential susceptibility and diathesis stress effects for mother-reported CBCL, we see in Table 1 that the only statistically significant Parental Sensitivity × Difficult Temperament interaction effect appears when the outcome variable was measured at 54 months of age. The upper-bound RoS (RoS on $X$) for this effect appears above 2 *SD*, suggesting that this effect fails Critical Test 1 for differential susceptibility. Both the PoI and the PA are equal to 0.00, suggesting the observed interaction effect also fails Critical Tests 2a and 2b for differential susceptibility. Regarding Critical Test 3a, we see no statistically significant interaction effect for overall mother-reported CBCL using the intercepts as outcomes model ($b_3 = -0.417$; $p = .074$). Nevertheless, there was still a possibility that the interaction might reflect a true period of developmental sensitivity that emerges uniquely at 54 months of age. To test this, we used Critical Test 3b to calculate a conservative alpha level (critical $p$ value) to control for the overall false discovery rate (Type I error) across multiple tests (see Appendix A for description of sequential Bonferroni procedure; Benjamini & Hochberg, 1995). Noting that nine separate tests of the interaction effect were conducted for mother-rated CBCL, across nine waves of assessment (24 and 36 months; 54 months; kindergarten; Grades 1, 3, 4, 5, and 6; and 15 years), then because only one of these nine $p$ values was below .05, the alpha level for the Benjamini and Hochberg (1995) sequential Bonferroni test would be .05 × (1/9 tests) = 0.0056 (see Appendix A). Because the observed $p$ value at 54 months was $p = .022 > .0056$, we conclude that the observed interaction effect at 54 months fails Critical Test 3b. Last, in terms of the nonlinearity test (Critical Test 4), we find that neither $X^2$, $ZX^2$, nor the change in $R^2$ for the set of $X^2$ and $ZX^2$ together was statistically significant. In sum, the Parental Sensitivity × Difficult Temperament interaction effect on Mother CBCL failed Critical Tests 1, 2, and 3 for differential susceptibility and passed Critical Test 4 (i.e., lack of nonlinearity). The form of the overall interaction effect is depicted in Figure 6, first panel (labeled CBCL Mother), where we see that there is no suggestion of differential susceptibility for this outcome variable within 2 *SD* of the mean of $X$. Instead, Mother CBCL results seem more consistent with either diathesis–stress (note in Figure 6 that the RoS on $X$ is shaded for low $X$ only [i.e., high and low difficult temperament only differ at low levels of $X$, not at high levels of $X$]) or, perhaps more likely, any apparent interaction effect is due to Type I error.

### *Teacher reports of total problems*

Next, for the outcome variable teacher-reported total problems (TRF teacher), there is a statistically significant early Sensitivity × Temperament interaction at Grades 1, 4, and 5, as well as a significant overall effect (averaged over time). The RoS on $X$ test suggests significant differences between high and low $Z$ (temperament) within the range from −2 to +2 *SD* on $X$ (sensitivity) at Grade 1, Grade 5, and overall (passing Critical Test 1). In contrast at Grade 4, the upper bound for RoS on $X$ was above +2 *SD*, meaning that temperament did not become a significant predictor until $X$ (sensitivity) exceeded 2.519 *SD* above the mean (failing Critical Test 1). Regarding the magnitude of the differential susceptibility effect, PoI indices were uniformly above 18%, and PA was uniformly above 16% (passing Critical Tests 2a and 2b). Critical Test 3a was also passed for TRF Teacher, given that the intercepts as outcomes overall effect was significant and consistent with differential susceptibility. For Critical Test 3b (familywise Type I error), there were seven significance tests conducted and three $p$ values less than .05 ($p = .002$, .004, and .012). Because the third largest $p$ value .012 < .0214 (see Appendix A), then Grade 1, Grade 4, and Grade 5 effects remain statistically significant beyond chance alone. Finally, $X^2$, $ZX^2$, and the change in $R^2$ for $X^2$ and $ZX^2$ together were not statistically significant (passing Critical Test 4).

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press



**Table 1.** *Regression estimates and differential susceptibility/diathesis–stress indices by outcome domain over time in the NICHD Study of Early Child Care and Youth Development*

| | | | | | | | Differential Susceptibility/Diathesis–Stress Indices | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Regression Estimates | | | | | | RoS *Z* | | RoS *X* | | | | | |
| Outcome | $b_0$ | $b_1$ | $b_2$ | $b_3$ | $XZ$ $\Delta R^2$ | $p$ | Lower Bound | Upper Bound | Lower Bound | Upper Bound | PoI | PA | Crossover | $X^2$ or $ZX^2$? |
| CBCL mother | | | | | | | | | | | | | | |
| 24/36 months | 51.68 | –1.26* | 2.18* | –0.395 | .002 | .094 | — | — | — | — | — | — | — | — |
| 54 months | 50.83 | –1.19* | 1.41* | –0.660 | .004 | .022* | <–3.00 | –0.595 | 0.974 | >–3.000 | .00 | .00 | 2.136 | No |
| K | 48.93 | –0.881* | 1.11* | –0.161 | .000 | .595 | — | — | — | — | — | — | — | — |
| G1 | 48.29 | –1.18* | 1.26* | –0.303 | .000 | .331 | — | — | — | — | — | — | — | — |
| G3 | 47.46 | –1.56* | 1.16* | –0.551 | .002 | .113 | — | — | — | — | — | — | — | — |
| G4 | 46.48 | –1.21* | 1.10* | –0.403 | .001 | .225 | — | — | — | — | — | — | — | — |
| G5 | 46.22 | –1.10* | 1.26* | –0.325 | .000 | .367 | — | — | — | — | — | — | — | — |
| G6 | 45.90 | –0.826* | 1.11* | –0.083 | .000 | .812 | — | — | — | — | — | — | — | — |
| 15 years | 44.90 | –0.500 | 1.40* | –0.051 | .000 | .892 | — | — | — | — | — | — | — | — |
| Overall (time intercept) | 46.59 | –1.22* | 1.47* | –0.417 | .000 | .074 | — | — | — | — | — | — | — | — |
| TRF teacher | | | | | | | | | | | | | | |
| K | 47.64 | –1.94* | –0.034 | –0.414 | .001 | .195 | — | — | — | — | — | — | — | — |
| G1 | 50.04 | –1.47* | –0.023 | –0.752 | .006 | .012*† | <–3.00 | –0.746 | –1.328[a] | 1.091[a] | .51 | .57 | –0.031 | No |
| G2 | 49.73 | –1.90* | –0.251 | –0.402 | .001 | .222 | — | — | — | — | — | — | — | — |
| G3 | 51.60 | –2.03* | 0.060 | –0.575 | .003 | .054 | — | — | — | — | — | — | — | — |
| G4 | 50.71 | –1.71* | 0.597* | –0.864 | .008 | .004*† | <–3.00 | –0.894 | 0.012 | 2.519 | .19 | .17 | 0.691 | No |
| G5 | 50.56 | –1.32* | 0.326 | –0.941 | .009 | .002*† | <–3.00 | –0.522 | –0.364[a] | 1.481[a] | .33 | .36 | 0.346 | No |
| G6 | 50.23 | –0.790* | 0.052 | –0.498 | .003 | .145 | — | — | — | — | — | — | — | — |
| Overall (time intercept) | 50.78 | –1.67* | 0.117 | –0.579 | .009 | .005*† | <–3.00 | –1.479 | –0.635[a] | 1.312[a] | .40 | .44 | 0.202 | No |
| Mother-rated social competence | | | | | | | | | | | | | | |
| 54 months | 98.26 | 2.40* | –2.77* | 0.350 | .000 | .398 | — | — | — | — | — | — | — | — |
| K | 102.56 | 2.03* | –2.58* | 1.20 | .007 | .008*† | <–3.00 | –0.643 | 1.093 | >–3.000 | .00 | .00 | 2.150 | No |
| G1 | 105.29 | 2.28* | –2.23* | 1.07 | .005 | .013*† | <–3.00 | –0.845 | 0.989 | >–3.000 | .00 | .00 | 2.084 | No |
| G3 | 106.06 | 2.92* | –2.78* | 1.14 | .005 | .017*† | <–3.00 | –1.128 | 1.182 | >–3.000 | .00 | .00 | 2.439 | No |
| G4 | 106.92 | 3.60* | –3.00* | 1.02 | .004 | .039* | <–3.00 | –1.568 | 1.353 | >–3.000 | .00 | .00 | 2.941 | No |
| G5 | 107.35 | 2.22* | –2.45* | 0.796 | .002 | .091 | — | — | — | — | — | — | — | — |
| G6 | 106.88 | 2.39* | –2.17* | 0.755 | .002 | .103 | — | — | — | — | — | — | — | — |
| 15 years | 103.49 | 2.40* | –2.12* | 0.702 | .002 | .167 | — | — | — | — | — | — | — | — |
| Overall (time intercept) | 105.32 | 2.46* | –2.52* | 0.837 | .004 | .019*† | <–3.00 | –1.338 | 1.528 | >–3.000 | .00 | .00 | 3.011 | No |
| Teacher-rated social competence | | | | | | | | | | | | | | |
| K | 103.22 | 2.28* | –0.294 | 0.894 | .004 | .049* | <–3.00 | –0.949 | –1.449 | >–3.000 | .34 | .38 | 0.329 | No |
| G1 | 103.11 | 1.87* | 0.394 | 1.35 | .010 | .003*† | <–3.00 | –0.533 | –1.311[a] | 0.396[a] | .62 | .68 | –0.259 | No |
| G2 | 104.65 | 2.80* | 0.592 | 1.58 | .012 | .002*† | <–3.00 | –0.781 | –1.398[a] | 0.174[a] | .68 | .72 | –0.375 | $X^2$ |
| G3 | 102.08 | 1.74* | 0.014 | 0.937 | .004 | .041* | <–3.00 | –0.543 | <–3.000 | >–3.000 | .50 | .57 | –0.015 | No |

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G4 | 102.02 | 2.20* | –0.758 | 1.27 | .008 | .005*† | <–3.00 | –0.674 | –0.089 | 2.344 | .22 | .22 | 0.597 | No |
| G5 | 102.68 | 1.95* | –0.431 | 1.29 | .008 | .003*† | <–3.00 | –0.571 | –0.401[a] | 1.523[a] | .33 | .37 | 0.334 | No |
| G6 | 102.51 | 2.11* | –0.032 | 0.634 | .002 | .242 | — | — | — | — | — | — | — | — |
| Overall (time intercept) | 102.56 | 2.16* | –0.084 | 1.10 | .011 | .000*† | <–3.00 | –1.065 | –0.522[a] | 0.690[a] | .46 | .51 | 0.076 | No |
| Teacher-rated academic skills | | | | | | | | | | | | | | |
| K | 98.39 | 2.32* | –0.253 | 0.682 | .003 | .065 | — | — | — | — | — | — | — | — |
| G1 | 98.23 | 2.12* | –0.050 | 1.09 | .009 | .002*† | <–3.00 | –0.932 | –0.746[a] | 0.861[a] | .47 | .52 | 0.046 | No |
| G2 | 99.09 | 1.70* | 0.011 | 1.09 | .008 | .008*† | <–3.00 | –0.530 | –1.001[a] | 0.907[a] | .50 | .56 | –0.010 | No |
| G3 | 98.84 | 1.74* | 0.006 | 0.939 | .007 | .007*† | <–3.00 | –0.760 | –1.138[a] | 1.047[a] | .50 | .56 | –0.006 | No |
| G4 | 98.48 | 1.73* | –0.554 | 0.815 | .005 | .023*† | <–3.00 | –0.809 | –0.246 | >–3.000 | .19 | .18 | 0.680 | No |
| G5 | 99.28 | 1.85* | –0.470 | 0.982 | .008 | .008*† | <–3.00 | –0.790 | –0.237 | 2.216 | .27 | .28 | 0.479 | No |
| Overall (time intercept) | 99.16 | 2.03* | –0.189 | 0.876 | .007 | .002*† | <–3.00 | –1.200 | –0.475[a] | 1.115[a] | .39 | .43 | 0.216 | No |
| WJ academic skills | | | | | | | | | | | | | | |
| 54 months | 434.30 | 3.43* | –0.457 | 0.231 | .000 | .549 | — | — | — | — | — | — | — | — |
| G1 | 476.88 | 2.00* | –0.603 | 0.671 | .004 | .038*† | <–3.00 | –1.283 | –0.030 | >–3.000 | .12 | .09 | 0.899 | No |
| G3 | 495.75 | 2.03* | –0.767* | 0.891 | .008 | .002*† | <–3.00 | –1.128 | 0.175 | 2.674 | .14 | .10 | 0.861 | No |
| G5 | 508.42 | 2.23* | –1.05* | 1.13 | .012 | .000*† | <–3.00 | –1.032 | 0.320 | 2.311 | .12 | .08 | 0.929 | $X^2$ |
| 15 years | 521.94 | 2.57* | –0.674 | 1.02 | .009 | .004*† | <–3.00 | –1.279 | –0.016 | 2.333 | .20 | .18 | 0.661 | No |
| Overall (time intercept) | 497.98 | 2.68* | –0.836* | 0.755 | .044 | .008*† | <–3.00 | –1.837 | 0.356 | >–3.000 | .08 | .03 | 1.107 | No |

*Note:* RoS, regions of significance; PoI, proportion of interaction; PA, proportion affected; CBCL, Child Behavior Checklist; TRF, Teacher Report Form; K, kindergarten assessment; G1–G6, Grade 1–Grade 6 assessments; 15 years, 15-year-old assessment; WJ, Woodcock–Johnson. The regression equation $Y = b_0 + b_1X + b_2Z + b_3XZ$, where $X$ is parental sensitivity and $Z$ is difficult temperament (control variables not reported); RoS $Z$ refers to the RoS with respect to temperament (Z). Specifically, the upper (lower) bound represents the value of temperament (standardized) above (below) which the regression of outcome on early sensitivity is statistically significant. RoS $X$ refers to the RoS with respect to early sensitivity. Specifically, the upper (lower) bound represents the value of $X$ (early sensitivity, standardized) above (below) which the regression of outcome on temperament is statistically significant. PoI falls above the cross-over point for the regressions. Cross-over represents the value of $X$ (early sensitivity, standardized) at which the regression lines intersect. PA is the proportion of participants who had early sensitivity scores that fell above the crossover point. $X^2$ or $ZX^2$ denotes whether $X^2$, $ZX^2$, or the set of both nonlinear terms together was statistically significant in the equation $Y = b_0 + b_1X + b_2Z + b_3XZ + b_3XZ + b_4X^2 + b_5ZX^2$. Overall represents the average effect across all assessment waves, estimated from an intercepts as outcomes multilevel model centered at fifth grade (Pluess & Belsky, 2010). The $XZ$ $\Delta R^2$ for the intercepts as outcomes model was estimated as 1 – (intercept variance with predictor/intercept variance without predictor).

[a]Outcomes for which both the lower and upper regions of significance with respect to early sensitivity (RoS $X$) fell within $\pm 2$ *SD* of the mean of early sensitivity.

*$p < .05$.

†False discovery rate $p < .05$ (controlling Type I error across multiple tests; Benjamini & Hochberg, 1995).

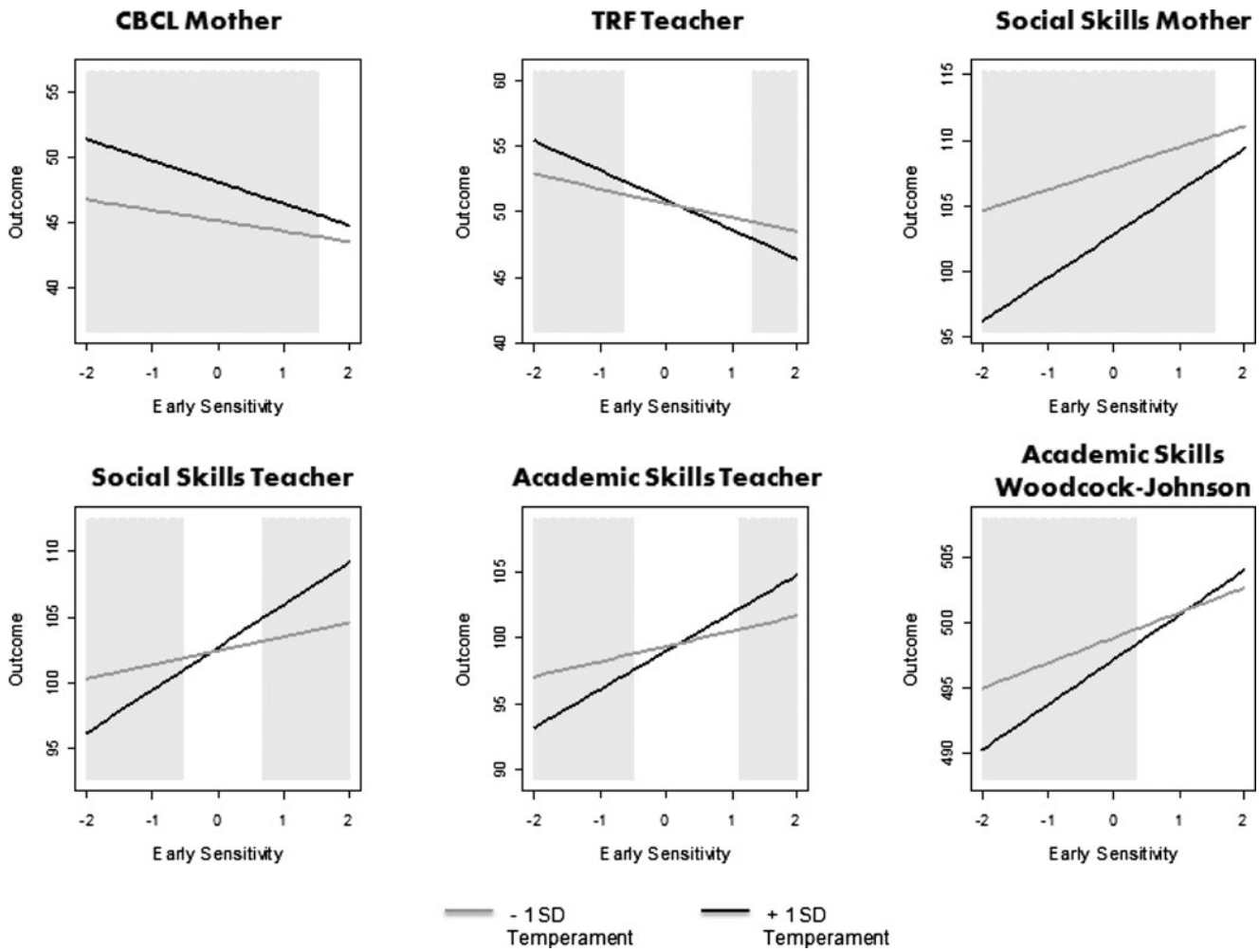



**Figure 6.** Empirical results for each of the six outcome domains (overall effects, i.e., time intercept), with regions of significance on *X* tests (gray) superimposed on simple slopes ($\pm 1$ *SD* on difficult temperament; $Z =$ *difficult* temperament). Note that all interactions are statistically significant, except mother-reported total problems on the CBCL, which is included here only for purposes of comparison. CBCL, Child Behavior Checklist; TRF, Teacher Report Form.

To summarize, differential susceptibility appears to be supported according to all critical tests at Grade 1, Grade 5, and in the overall (mean across time) effect. In the plot of the overall interaction in Figure 6, second panel (TRF teacher), we see shaded regions of significance (RoS on *X*) at both the high and low ends of X (Critical Test 1 for differential susceptibility). In other words, difficult temperament both (a) increases teacher-reported problems at low levels of sensitivity and (b) decreases teacher-reported problems at high levels of sensitivity. The data on teacher-reported problems clears all four hurdles for demonstrating differential susceptibility.

*Maternal reports of social competence*

For mother-rated social competence, the RoS on *X* test (Critical Test 1) fails under all conditions, because high and low temperament (*Z*) do not differ at high levels of *X* (+2 *SD* above the mean of parental sensitivity; see Figure 6, panel 3, Social Skills Mother). Consistent with this lack of evidence for differential susceptibility at high *X*, both PoI and PA are 0.00 for all conditions (Critical Tests 2a and 2b). The overall time intercept shows a significant interaction (Critical Test 3a), and this overall effect survives the conservative familywise Type I error test (Critical Test 3b); note that for the overall (time intercept) effects, we are testing six overall time intercept effects in Table 1, mother-rated social competence exhibits the fifth largest *p* value ($p = .019$), and this *p* value $< .0417$ (see Appendix A). There is no evidence of nonlinearity effects (passing Critical Test 4). To recap, mother-rated social competence seems consistent with a diathesis–stress account, rather than a differential susceptibility account. Regarding the chance that these observed interaction effects are due to familywise Type I error, note that three of these significant interaction effects pass the Benjamini–Hochberg test (see Appendix A, eight tests, third lowest *p* value $= .017 <$

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

.0188), supporting the interaction at kindergarten, Grade 1, and Grade 3.

### *Teacher reports of social competence*

For teacher-rated social competence, six out of seven assessment waves exhibited statistically significant interactions, and three of these (Grades 1, 2, and 5) passed the RoS on *X* test (Critical Test 1; temperament has significant effects in two different directions, within the range of +2 to −2 *SD* on early sensitivity). PoI and PA effect size indices were also large (uniformly above 21%, and above 50% in Grades 1, 2, and 3; passing Critical Test 2). The overall (time intercept) effect was significant (passing Critical Test 3a), and four of the seven waves were statistically significant beyond the familywise false discovery (Type I error) rate (passing Critical Test 3b). Only one assessment wave showed evidence of nonlinearity ($X^2$ effect was statistically significant at Grade 2), but the *XZ* interaction effect remained statistically significant after controlling both $X^2$ and $ZX^2$ (passing Critical Test 4). In total, teacher-rated social competence passes all four critical tests and shows strong evidence of differential susceptibility. Figure 6 corroborates this conclusion.

### *Teacher reports of academic skills*

Teacher-rated academic skills (a) passed Critical Test 1 (RoS on *X*) at Grades 1, 2, 3, and overall, (b) passed Critical Tests 2a and 2b at all time points (PoI always >18% and PA always >17%), (c) passed Critical Test 3a (overall time intercept), (d) passed Critical Test 3b (familywise Type I error) at five out of six assessment waves, and (e) exhibited no nonlinearity. Teacher-rated academic skills reveal strong evidence of differential susceptibility (see Figure 6).

### *Objective reports of academic skills*

Scores on the WJ-J test showed a statistically significant interaction effect at four out of five assessment waves, but failed Critical Test 1 (the RoS fell below +2 *SD* on *X*; see Figure 6). In other words, the crossover point was simply too high on *X* to support a strong inference of differential susceptibility. PoI was approximately 15%, and the proportion of the sample affected (PA) was around 11% (i.e., differential susceptibility indices were much lower here than in the cases of teacher ratings of total problems, social competence, and academic skills reported above). The overall effect was statistically significant (passing Critical Test 3a), and *p* values were lower than conservative sequential Bonferroni levels at four out of five assessment waves (passing Critical Test 3b). Finally, at Grade 5 there was evidence of nonlinearity ($X^2$ and the change in $R^2$ for $X^2$ and $ZX^2$ together were both statistically significant), but the *XZ* interaction remained statistically significant after controlling these effects (passing Critical Test 4). Altogether, it appears that the Woodcock–Johnson results are more consistent with evidence of diathesis–stress than differential susceptibility (see Figure 6). There is a robust Sensitivity × Temperament interaction effect, but high and low temperament do not differ statistically at the high end of the normal range of *X* (sensitivity).

## Discussion

This paper was designed to highlight four problems endemic to the methods used in most published studies examining evidence for *differential susceptibility*, which is the hypothesis that factors (e.g., difficult temperament) long thought only to amplify risk given poor developmental inputs (e.g., low quality parenting) actually enhance adaptation, given supportive caregiving experiences. Specifically, we revisited evidence that temperamental difficulty serves as a marker of enhanced susceptibility to early maternal caregiving across a range of outcome domains in the NICHD Study of Early Child Care and Youth Development (SECCYD; Bradley & Corwyn, 2008; Dopkins-Stright et al., 2008; Pluess & Belsky, 2010), examining whether published findings held up to the scrutiny of the use of a standard range of interest on *X*, RoS on *X* testing, our new metrics for quantifying the degree of evidence for differential susceptibility over diathesis–stress, the possibility that some findings in the literature could have been generated by Type I errors, and the potential for nonlinear associations consistent with diathesis–stress effects to masquerade as evidence of differential susceptibility.

The key finding in our analysis is that many of the results published in this emerging literature, at least those based on the NICHD SECCYD, held up quite well to our recommended standards. In particular, we found no evidence that nonlinear effects were masquerading as evidence for differential susceptibility, and relatively little evidence for Type I error due to multiple testing, although we should note that it was not always the case that a given pattern of findings replicated for each and every assessment of a given outcome domain. The most crucial point is that we found consistent evidence in the NICHD SECCYD, using RoS on *X* testing augmented by our new differential susceptibility metrics, that temperamentally difficult infants were differentially susceptible to the effects of early maternal sensitivity, both for good and for ill, with respect to *teacher-reported* symptoms of psychopathology (Bradley & Corwyn, 2008), social competence (Dopkins-Stright et al., 2008; Pluess & Belsky, 2010), and academic skills (Pluess & Belsky, 2010). These empirical findings are impressive to us, particularly given that different teachers rated participants over time. As such, reliable differential susceptibility effects across developmental periods cannot be easily dismissed to rater biases.

First, our more representative analysis of assessments of adaptation in the SECCYD also demonstrated little evidence that difficult temperament moderated the predictive significance of early sensitivity for outcomes of mother-reported symptoms of psychopathology. Second, a *diathesis–stress* model clearly best characterized the way in which difficult

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

temperament moderated the association between early sensitivity and mother reports of social competence over time (diathesis–stress interaction effects were consistently significant in the overall test, and from kindergarten through Grade 3 for this outcome domain).

Third, and importantly, Pluess and Belsky (2010) recently reported that a differential susceptibility model best characterized the way in which early temperament moderates the predictive significance of early sensitivity for objective reports of broad reading in the SECCYD based on the WJ-R assessment (they focused on WJ-R data through Grade 5). Although it is indeed the case based on visual inspection that there is some crossover in such interactions, RoS testing and our metrics for quantifying differential susceptibility effects suggest to us that a diathesis–stress model in fact better captures the way in which early temperament moderates the predictive effects of early sensitivity on objective tests of academic skills over time. Said another way, we found no consistent evidence in the SECCYD (particularly when we focused on RoS on *X* testing) that individuals with early difficult temperaments show enhanced academic functioning as assessed using standardized testing when they experience high-quality early parenting, as reported in Pluess and Belsky (2010). We did, however, find evidence that difficult infants, compared to their relatively easy counterparts, may be disproportionately negatively affected by low-quality early caregiving in terms of their standardized test scores, a pattern of findings consistent with the diathesis–stress account.

It is not entirely clear why the patterns of moderation we identified (i.e., differential susceptibility vs. diathesis–stress) differed by outcome measure informant in the SECCYD, with differential susceptibility effects predominating for the teacher-report data and diathesis–stress effects being more conspicuous for mother-reported social skills and objective assessments of academic skills. However, one speculative possibility is that these findings were in part driven by context-specific child behaviors, where difficult temperament is only (purely) a risk factor in contexts that include the individual who rated the temperamental variation of the target child. It is also noteworthy that corroborative findings emerged in terms of the main effects results, which we have underemphasized in this report. As reported elsewhere (see Fraley et al., in press; Haltigan et al., in press), early maternal sensitivity was consistently associated with developmental adaptation across outcome domains and informants. However, the main effects of temperament differed notably by informant. Mother-reported outcomes were, perhaps not surprisingly, significantly predicted by variation in temperament, which was also mother reported. In contrast, teacher-report outcomes were not consistently associated with difficult temperament, although objective assessments of academic skills were. As we have noted earlier (see footnote 4), main effects of the moderator on the outcome make the differential susceptibility account less tenable.

Another issue that should be emphasized is that, as discussed more completely in footnote 2, the strong form of the differential susceptibility account could be understood to suggest that individuals low on the plasticity factor (e.g., easy temperament) would be hyposusceptible to caregiving experiences in the absolute sense (as per Figure 1c). However, the results of our analysis consistently demonstrated that significant associations between maternal sensitivity and aspects of developmental mal(adaptation) obtained *both* for individuals relatively high *and* low on difficult temperament, even though those effects were significantly stronger for relatively difficult children. This pattern of results can be discerned through inspection of the upper bounds of the RoS on *Z* tests in Table 1. Of interest is whether those values range from relatively low to relatively high values on *Z* (e.g., −1 to +1 *SD*). As can be seen in Table 1, particularly for the overall effects across outcomes, this is the case as the upper bounds of the RoS on *Z* included values that ranged from below −1 *SD* from the mean of difficult temperament (*Z*) to all possible positive values high on *Z*. In other words, although early parenting seems to matter more for temperamentally difficult children, parenting still had predictive significance for temperamentally easy children.

One final issue concerns development. As is clear from the discussion above, we identified scant evidence in the NICHD SECCYD that age of assessment of a given aspect of developmental (mal)adaptation was systematically associated with the form of the interaction effect obtained. Nonetheless, we suspect that attention to the possibly dynamic nature of interactions between caregiving experiences and plasticity/vulnerability factors within a domain of adaptation is well placed. In particular, as differential susceptibility models continue to move into the domain of Gene × Environment interaction (G × E) research (e.g., Kochanska et al., 2011), it will be important to document whether G × E effects endure over development, and, if so, whether a given variable number tandem repeat or single nucleotide polymorphism changes from a vulnerability to plasticity factor (or vice versa) over time. The new metrics proposed in this paper for quantifying evidence for differential susceptibility and diathesis–stress (i.e., PoI and PA) will make plotting such age-related functions relatively straightforward. (Also potentially valuable is plotting the *b*s associated with the interaction term *over time*, although to be clear such a function would provide no information about the specific form of the interaction, just whether the G × E interaction effect endures, wanes, or increases in magnitude over development.)

Indeed, we suspect that the analysis applied in this paper will be more generally valuable as G × E research proceeds. We agree with other scholars that thinking about genetic variants as potential plasticity enhancers rather than merely sources of vulnerability is an attractive and fundamentally developmental idea (Belsky & Pluess, 2011). However, without reference to each of the critiques of the methods used to conclude in favor of differential susceptibility described in this paper, there is significant risk that many of the conclusions reached in such work will prove unreplicable, a concern that has received some empirical support in recent meta-analytic

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

work on G×E effects (Munafò, Durrant, Lewis, & Flint, 2009; Risch et al., 2009).

Before concluding, we offer several caveats regarding this work. First, we do not wish to be seen as implicitly advocating that the diathesis–stress model should be assumed as a default when significant interaction effects obtain. As noted earlier, attention to the issues discussed in this article are just as important in providing evidence for diathesis–stress as they are for differential susceptibility effects. For example, investigators who desire to provide clearer evidence that their data are consistent with a diathesis–stress account would benefit from quantifying the evidence for this model using our PoI and/or PA metrics. Likewise, as noted earlier, alpha inflation via multiple testing has the potential to undermine inferences regarding evidence for diathesis–stress, much as it does conclusions in favor of differential susceptibility.

A related point is that, although we have focused our analysis on differential susceptibility theory and its predictions, the methods described in this paper are likely to be equally valuable in other kinds of studies of risk and resilience that focus on interaction effects. For example, resilience researchers have described a variety of interaction effects pertinent to risk, resilience, and (mal)adaptation (see Luthar er al., 2000; Masten, in press). However, the evidence in favor of such models has been rarely quantified using the kinds of diagnostics described in this report. Similarly, we would be remiss in failing to note that biological sensitivity to context theory (Boyce & Ellis, 2005), while based on a different set of evolutionary psychology-inspired principles, actually makes precisely the same prediction as the differential susceptibility account with respect to the expectation that some children are likely to be disproportionately affected by both negative and positive life circumstances. Thus, all the methods described in this report should be valuable in evaluating evidence for biological sensitivity to context theory.

Second, one focus of this analysis was the developmental question of whether interactions between early sensitivity and temperament leave an enduring, or merely transient, mark on developmental adaptation in the SECCYD across outcome domains. However, it is important to emphasize that we did not examine sensitive caregiving in a dynamic fashion in this report (i.e., by incorporating later measures of sensitive caregiving into this analysis; but see Fraley et al., in press; Haltigan et al., in press). It is possible, for example, that one might find that the predictive significance of sensitive caregiving measured in close temporal proximity to assessments of developmental outcomes is moderated by temperament (whether measured early or later in development), even if interactions between *early* sensitivity and *early* temperamental difficulty have merely short-term predictive significance. Both kinds of analyses are relevant in addressing questions related to sensitive periods in development, an issue that goes beyond the scope of this report.

Third and finally, perhaps the most important issue that we have sidelined in this paper is how the sampling of *X*, *Y*, and *Z* all constrain what one might be able to conclude about evidence for diathesis–stress versus differential susceptibility in any given dataset. Here we would note that the differential susceptibility account emphasizes that susceptible individuals who, *in the absolute sense*, receive below- and above-average caregiving would be expected to achieve, respectively, both less *and* more favorable outcomes compared to less susceptible individuals. As such, it is debatable whether it is even wise to attempt to distinguish differential susceptibility from diathesis–stress accounts if (for example) the full range of the *X* variable (parent sensitivity) is not well represented within a given sample or if the midpoint on *X* (i.e., 0 after standardization) is not reasonably reflective of the average score for that construct in the population (see footnote 5 for elaboration). None of the methods we describe in this report allows one to sidestep the limitations of one's sampling frame (e.g., failing to capture the positive end of the distribution of *X*), and indeed some of the approaches we advocate (e.g., standardization of *X* and *Z*) might actually exacerbate them.

In summary, we have provided four constructive critiques of differential susceptibility evidence and proposed a set of critical tests that can be used to better establish the phenomenon. First, because all interactions would eventually cross over if extrapolated to a wide enough range of *X*, it is imperative that researchers choose the same range of interest on *X*. Second, RoS on *X* testing combined with quantifying the size of any differential susceptibility effects allows the investigator to sidestep the misleading conclusions that she or he might reach about differential susceptibility through visual inspection alone. Third, there are oftentimes a large number of independent variables, dependent variables, and moderators available in studies of risk and resilience, particularly if constructs have been assessed multiple times across development. Careful consideration of the possibility that observed moderated effects (whether interpreted in line with differential susceptibility or diathesis–stress accounts) might be attributable to chance alone leads to using the approaches described in this report to preserve familywise Type I error rates (or false discovery rates) at 0.05. Fourth, the possibility that nonlinear effects consistent with the diathesis–stress model are masquerading as differential susceptibility should be routinely considered. Attention to these issues will help build a robust and replicable set of results related to the differential susceptibility hypothesis, an idea that has great potential to shape the next decades of developmental science.

## References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4–18 and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & Edelbrock, C. (1986). *Manual for the Teacher's Report Form and Teacher Version of the Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

Achenbach, T. M., Edelbrock, C., & Howell, C. (1987). Empirically-based assessment of the behavioral/emotional problems of 2–3 year old children. *Journal of Abnormal Child Psychology*, *15*, 629–650.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*, 94–107.

Aguinis, H., & Gottfredson, R. K. (2010). Best-practice recommendations for estimating interaction effects using moderated multiple regression. *Journal of Organizational Behavior*, *31*, 776–786.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury, CA: Sage.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Lifespan developmental psychology: Introduction to research methods*. Monterey, CA: Brooks/Cole.

Baron, R. M., & Kenny, D. A. (1986). The mediator–moderator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Belsky, J. (1997). Variation in susceptibility to environmental influences: An evolutionary argument. *Psychological Inquiry*, *8*, 182–186.

Belsky, J., Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). For better *and* for worse: Differential susceptibility to environmental influence. *Current Directions in Psychological Science*, *16*, 300–304.

Belsky, J., Hsieh, K., & Crnic, K. (1998). Mothering, fathering, and infant negativity as antecedents of boys' externalizing problems and inhibition at age 3: Differential susceptibility to rearing influence? *Development and Psychopathology*, *10*, 301–319.

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: Differential susceptibility to environmental influences. *Psychological Bulletin*, *135*, 885–908.

Belsky, J., & Pluess, M. (2011). Beyond adversity, vulnerability, and resilience: Individual differences in developmental plasticity. In D. Cicchetti & G. I. Roisman (Eds.), *The origins and organization of adaptation and maladaptation: Minnesota Symposia on Child Psychology* (Vol. 36, pp. 379–422). Hoboken, NJ: Wiley.

Belsky, J., & Pluess, M. (in press). Differential susceptibility to long-term effects of quality of child care on externalizing behavior in adolescence. *International Journal of Behavioral Development*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*, 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1165–1188.

Boyce, W. T., & Ellis, B. J. (2005). Biological sensitivity to context: A. An evolutionary–developmental theory of the origins and functions of stress reactivity. *Development and Psychopathology*, *17*, 271–301.

Bradley, R. H., & Corwyn, R. F. (2008). Infant temperament, parenting, and externalizing behavior in first grade: A test of the differential susceptibility hypothesis. *Journal of Child Psychology and Psychiatry*, *49*, 124–131.

Carey, W. B., & McDevitt, S. C. (1978). Revision of the Infant Temperament Questionnaire. *Pediatrics*, *61*, 735–739.

Caspi, A., Sugden, K., Moffitt, T. E., Mill, J., Taylor, A., Craig, I. W., et al. (2003). Influence of life stress on depression: Moderation by a polymorphism in the *5-HTT* gene. *Science*, *301*, 386–389.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Cortina, J. M. (1993). Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. *Journal of Management*, *19*, 915–922.

Dearing, E., & Hamilton, L. C. (2006). Contemporary approaches and classic advice for analyzing mediating and moderating variables. *Monographs of the Society for Research in Child Development*, *71*, 88–104.

Dick, D. M., Meyers, J. L., Latendresse, S. J., Creemers, H. E., Lansford, J. E., Pettit, G. S., et al. (2011). CHRM2, parental monitoring, and adolescent externalizing behavior: Evidence for gene–environment interaction. *Psychological Science*, *22*, 481–489.

Dopkins-Stright, A., Cranley Gallagher, K., & Kelley, K. (2008). Infant temperament moderates relations between maternal parenting in early childhood and children's adjustment in first grade. *Child Development*, *79*, 186–200.

Dunn, O. J. (1961). Multiple comparison among means. *Journal of the American Statistical Association*, *56*, 52–64.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Fraley, R. C., & Marks, J. (2007). The null hypothesis significance testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 149–169). New York: Guilford Press.

Fraley, R. C., Roisman, G. I., & Haltigan, J. D. (in press). The legacy of early experiences in development: Formalizing alternative models of how early experiences are carried forward over time. *Developmental Psychology*.

Garmezy, N., Masten, A. S., & Tellegen, A. (1984). The study of stress and competence in children: A building block for developmental psychopathology. *Child Development*, *55*, 97–111.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80–100.

Graham, J. W. (2009). Missing data analysis: Making it in the real world. *Annual Review of Psychology*, *60*, 549–576.

Gresham, F. M., & Elliott, S. N. (1990). *The social skills rating system*. Circle Pines, MN: American Guidance Service.

Haltigan, J. D., Roisman, G. I., & Fraley, R. C. (in press). The predictive significance of early caregiving experiences for symptoms of psychopathology through midadolescence: Enduring or transient effects? *Development and Psychopathology*.

Heim, C., & Nemeroff, C. B. (1999). The impact of early adverse experiences on brain systems involved in the pathophysiology of anxiety and affective disorders. *Biological Psychiatry*, *46*, 1509–1522.

Kochanska, G., Kim, S., Barry, R. A., & Philibert, R. A. (2011). Children's genotypes interact with maternal responsive care in predicting children's competence: Diathesis–stress or differential susceptibility? *Development and Psychopathology*, *23*, 605–616.

Luthar, S. S., Cicchetti, D., & Becker, B. (2000). The construct of resilience: A critical evaluation and guidelines for future work. *Child Development*, *71*, 543–562.

Masten, A. S. (2001). Ordinary magic: Resilience processes in development. *American Psychologist*, *56*, 227–238.

Masten, A. S. (in press). Risk and resilience in development. In P. D. Zelazo (Ed.), *Oxford handbook of developmental psychology*. Oxford: Oxford University Press.

Munafò, M. R., Durrant, C., Lewis, G., & Flint, J. (2009). Gene × Environment interactions at the serotonin transporter locus. *Biological Psychiatry*, *65*, 211–219.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*, 328–362.

NICHD Early Child Care Research Network. (2001). Child-care and family predictors of preschool attachment and stability from infancy. *Developmental Psychology*, *37*, 847–862.

NICHD Early Child Care Research Network. (2005). *Child care and child development*. New York: Guilford Press.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and predication* (2nd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Pluess, M., & Belsky, J. (2009). Differential susceptibility to rearing experience: The case of child-care. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *50*, 396–404.

Pluess, M., & Belsky, J. (2010). Differential susceptibility to parenting and quality child care. *Developmental Psychology*, *45*, 379–390.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., et al. (2009). Interaction between the serotonin transporter gene (*5-HTTLPR*), stressful life events, and risk of depression: A meta-analysis. *Journal of the American Medical Association*, *301*, 2462–2471.

Roisman, G. I., Booth-LaForce, C., Cauffman, E., Spieker, S., & the NICHD Early Child Care Research Network. (2009). The developmental signifi-

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press

cance of adolescent romantic relationships: Parent and peer predictors of quality and engagement at age 15. *Journal of Youth and Adolescence*, *38*, 1294–1303.
Roisman, G. I., Monahan, K. C., Campbell, S. B., Steinberg, L., Cauffman, E., & the NICHD Early Child Care Research Network. (2010). Is adolescence-onset antisocial behavior developmentally normative? *Development and Psychopathology*, *22*, 295–311.
Roisman, G. I., Susman, E., Barnett-Walker, K., Booth-LaForce, C., Owen, M. T., Belsky, J., et al. (2009). Early family and child-care antecedents of awakening cortisol levels in adolescence. *Child Development*, *80*, 907–920.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.
Sidak, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626–633.
Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*, 751–754.
Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment*, *8*, 231–258.
Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Psycho-Educational Battery—Revised.* Allen, TX: DLM.

## Appendix A

*Critical α values for simple sequential Bonferroni-type procedure to control Type I error across multiple tests*

| | Sequential Position Lowest → Highest $p$ ($j$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| No. Tests ($c$) | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th |
| 10 | 0.005 | 0.01 | 0.015 | 0.02 | 0.025 | 0.03 | 0.035 | 0.04 | 0.045 | 0.05 |
| 9 | 0.0056 | 0.0111 | 0.0167 | 0.0222 | 0.0278 | 0.0333 | 0.0389 | 0.0444 | 0.05 | |
| 8 | 0.0063 | 0.0125 | 0.0188 | 0.025 | 0.0313 | 0.0375 | 0.0438 | 0.05 | | |
| 7 | 0.0071 | 0.0143 | 0.0214 | 0.0286 | 0.0357 | 0.0429 | 0.05 | | | |
| 6 | 0.0083 | 0.0167 | 0.025 | 0.0333 | 0.0417 | 0.05 | | | | |
| 5 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | | | | | |
| 4 | 0.0125 | 0.025 | 0.0375 | 0.05 | | | | | | |
| 3 | 0.0167 | 0.0333 | 0.05 | | | | | | | |
| 2 | 0.025 | 0.05 | | | | | | | | |
| 1 | 0.05 | | | | | | | | | |

*Note:* Table entries are $\alpha(j/c)$, or $.05(j/c)$. The Benjamini–Hochberg (1995) procedure involves three steps. First, order all the observed $p$ values from lowest $p$ value (1st, or $j$=1 in top row of the table) to highest $p$ value. Second, find $k$, which is the largest value of $j$ for which $p_j < \alpha(j/c)$. Third, declare the first $k$ (i.e., 1st, 2nd, . . . , $k$th) $p$ values statistically significant. For example, if four significance tests were conducted ($c = 4$), and the $p$ values from lowest to highest were $p =$ .02, .02, .03, and .06, then $k = 3$, because the third largest $p$ value (i.e., .03) is $<.0375$. Thus, we reject the null hypothesis for the three lowest $p$ values.

https://doi.org/10.1017/S0954579412000065 Published online by Cambridge University Press