The construct of psychophysiological reactivity: Statistical and psychometric issues

Keith B. Burt a,*, Jelena Obradovic b

aUniversity of Vermont, Burlington, VT, United States
bStanford University, Stanford, CA, United States

ARTICLE INFO

Article history:
Received 1 February 2012
Revised 18 October 2012
Available online 16 November 2012

Keywords:
Psychophysiology
Reactivity
Measurement of change
Difference score
Residual score
Latent variable modeling

ABSTRACT

The purpose of this paper is to review major statistical and psychometric issues impacting the study of psychophysiological reactivity and discuss their implications for applied developmental researchers. We first cover traditional approaches such as the observed difference score (DS) and the observed residual score (RS), including a review of classic and recent research on their reliability and validity from two related bodies of work: the measurement of change and the Law of Initial Values. Second, we review several types of latent variable modeling in this context: latent difference score (LDS) models, latent residual score (LRS) models, latent state-trait (LST) models, and latent growth curve (LGC) models. Finally, we provide broad guidelines for applied researchers broken down by key stages of a psychophysiological project: study planning, data analysis, and reporting of results. Our recommendations highlight the need for (1) increased attention to the ubiquitous nature of measurement error in observed variables and the importance of employing latent variable models when possible, and (2) increased specification of theories relating to the construct of reactivity, especially in regards to the distinction between baseline arousal and change over time in broader systems of variables.

© 2012 Elsevier Inc. All rights reserved.

Introduction

Recent years have seen a proliferation of research studies that aim to link processes of children’s adaptation across multiple levels of analysis, with a focus on integrating psychophysiological assess-
ments into the study of developmental psychology (e.g., Beauchaine, 2001; Curtis & Cicchetti, 2003; Obradović, 2012). Analytically, there is substantial interest in connecting youth’s physiological responses to stressors/challenges to both behavioral outcomes as well as environmental influences. Although this research draws from decades of psychophysiological research in adults, it lacks consistency in conceptualizing and measuring indices of youth physiological responsivity. The goal of the present paper is to review the strengths and weaknesses of different approaches to the reduction and analysis of physiological data, in the hope of motivating better tests and refinement of existing developmental theories as well as more precise conceptualization of developmental changes in stress responsivity. In particular, we discuss statistical and psychometric issues that have relevance for researchers seeking connections between psychophysiological variables and other constructs of interest.

Following a brief introduction to the collection of psychophysiological data, the first half of this paper focuses on psychometric issues related to traditional measures of reactivity: the observed difference score and the observed residual score. This section draws on sources from the broad measurement of change as well as the psychophysiological Law of Initial Values, literatures with both a rich history (e.g., Heath & Oken, 1965; Wilder, 1931) and considerable recent revival (e.g., May & Hittner, 2010; Zimmerman, 2009). Since many of the publications discussing these topics are not in journals widely read by applied developmental psychologists, we hope that our review will provide developmental researchers with not only increased appreciation for the complex issues involved in such data reduction, but also practical guidance for research situations favoring a particular score. A recurring theme in this discussion is the ubiquitous yet unmeasured presence of error in observed scores.

The second half of this paper reviews newer, promising approaches to the analysis of reactivity data: latent difference score modeling, latent residual score modeling, latent growth curve modeling, and latent state-trait modeling. These methods offer great promise for the testing and refinement of developmental theory. Yet, with few exceptions, applications of these approaches remain rare in youth psychophysiology research; thus, we hope that our review will encourage the increased adoption and testing of these powerful analytic techniques. To that end, we provide sample programming syntax in Mplus (Muthén & Muthén, 1998–2010) for each of the discussed models in the Appendix (syntax partly adapted from Brown, Croudace, & Heron, 2011). We end our review with a set of general research recommendations.

Psychophysiological reactivity

We restrict our review primarily to statistical issues in studies that derive a single physiological reactivity score. To help ground our discussion, consider that such study paradigms often employ measures of the autonomic nervous system (ANS), a fast-acting physiological stress response system (Berntson, Quigley, & Lozano, 2007; Obradović, 2012). The ANS has traditionally been divided into two primary branches, the sympathetic and parasympathetic; the former initiates physiological arousal (termed the “fight-or-flight” response) and the latter restores homeostasis (termed the “rest-and-digest” response). Common markers of sympathetic ANS activity include cardiac pre-ejection period (PEP) and skin conductance level (SCL), whereas common markers of parasympathetic ANS activity include respiratory sinus arrhythmia (RSA), derived from polyvagal theory (Porges, 2001, 2007). However, the two ANS branches do not necessarily operate in a simple reciprocal manner, with co-activation and co-inhibition both possible. Although in recent years researchers have started to examine how the two systems jointly relate to developmental processes (Del Giudice, Hinnant, Ellis, & El-Sheikh, 2012), the majority of developmental research still focuses on a single measure of ANS activity at a time.

Although quite common, single-score operationalization is admittedly simplistic. In contrast, Berntson, Cacioppo, and Quigley (1991) outline a comprehensive theory of autonomic control that details various modes of SNS and PNS response, taken together. Among other points, Berntson et al. (1991) stress the importance of working towards independent assessment and manipulation of the SNS and PNS. The complexity of these and other psychophysiological systems has also been
highlighted by experts such as Porges (2007) and Cacioppo and Tassinary (1990), who note that there is not a simple one-to-one mapping between physical organ and physiological measurement nor from physiological measurement to psychological construct. Although a comprehensive review of these systems is beyond the scope of this paper, readers are encouraged to consult other resources for more details (e.g., Berntson et al., 1991, 2007; Obradović & Boyce, 2012).

In a typical developmental psychophysiological study, researchers bring participants to a laboratory setting to measure ANS physiological processes (e.g., heart rate, skin conductance level, or blood pressure). Baseline ANS activity levels are generally measured while children are relaxed and resting. Because young children often find it difficult to be physically still, researchers sometimes utilize the reading of a calming story as a means of inducing a state of quiet restfulness. Recently, however, some researchers have emphasized the importance of isolating psychological responses to the experimental stressor and controlling for peripheral triggers of cardiovascular activation, such as muscle movement, by calculating task-specific baseline values (Bush, Alkon, Obradović, Stamperdahl, & Boyce, 2011; Kamarck & Lovallo, 2003). Broader discussions of many practical considerations in obtaining psychophysiological data with children and adolescents are provided by Gavin and Davies (2008) as well as Miller and Long (2008). Throughout this paper, baseline scores are denoted by the variable $X$.

During one or more laboratory stressor or challenge tasks, the physiological process or processes of interest are re-assessed, leaving researchers with two scores for calculating reactivity, or short-term change between baseline and challenge response. In some instances, researchers then aggregate reactivity scores across multiple stressors. Throughout this paper, challenge scores are denoted by the variable $Y$.

**Traditional approaches: observed difference scores and residual scores**

In this section, we review prior areas of investigation in psychometrics and psychophysiology focused on the reduction and analysis of $X$ and $Y$ scores. A crucial starting point is the extended literature on the measurement of change (e.g., Cronbach & Furby, 1970; Rogosa, Brandt, & Zimowski, 1982). In addition, we consider a specific hypothesis from the physiological literature—the Law of Initial Values—which predicts that in certain contexts, baseline status and level of reactivity will be correlated (Jamieson & Howk, 1992; Wilder, 1931). Using both literatures, we compare and contrast the observed difference score and observed residual score, which remain two of the most common operationalizations of reactivity. Summarizing key points from this section, Table 1 presents various factors that influence the reliability and validity of these two scores. Although these factors are not uniquely relevant to youth samples, discussion of the reliability and validity of physiological scores is rarely included in applied developmental publications.

**Definitions**

Given the study design described in the preceding section, our discussion focuses on the difference score (hereafter termed DS), defined as

$$DS = Y - X$$

as well as the residual score (hereafter termed RS), defined as sample-specific residuals of a linear regression of $Y$ on $X$. Algebraically, the equation for a residual score follows from re-working the standard multiple regression equation as follows (e.g., Stemmler, 2003):

$$RS = Y - B_{XY}(X - \bar{X}) - \bar{Y}$$

where $\bar{X}$ is the sample mean of the baseline $X$, $\bar{Y}$ is the sample mean of the challenge $Y$, and $B_{XY}$ is the sample ordinary least squares (OLS) regression coefficient of $Y$ on $X$. In the physiological research context, the RS was originally introduced as the “autonomic lability score” by Lacey (1956), whereas the DS has been termed “delta”, “gain score”, or “change score” by some scholars. In the developmental physiological literature, researchers generally use the DS and/or RS to represent children’s physiological reactivity. However, as discussed below, other operationalizations of reactivity are also present,
such as separate entry of baseline and challenge variables to a regression equation predicting a desired outcome variable, or the addition of the baseline variable to a regression equation along with the DS or the RS.

Similarities and differences

Because a DS is the simple arithmetic difference between two variables typically measured on the same scale, it has a relatively straightforward numeric interpretation: a positive score means an increase in activity from baseline to challenge and a negative score means a decrease in activity from baseline to challenge. This transparency of interpretation can be viewed as an advantage (Rogosa et al., 1982), especially in the physiological literature where, for example, a net decrease in parasympathetic input to heart from baseline levels is interpreted as “vagal withdrawal” with specific theoretical significance (Beauchaine, Gatzke-Kopp, & Mead, 2007; Calkins, Graziano, & Keane, 2007). That said, the DS will not differentiate between participants with high and low baseline activity if they show equal arousal.

Conversely, an RS represents whether an individual has changed more or less “than expected” based on one’s baseline score (Cronbach & Furby, 1970). Calculating the residual score is thus a two-step process: first, one calculates an expected challenge score from the sample regression line of challenge (Y) on baseline (X); second, one calculates the difference between each participant’s observed and expected challenge scores. A positive RS then represents an observed challenge score that is larger than expected given the sample regression line, while a negative RS represents a challenge score that is smaller than expected.

However, it is not always conceptually clear what changing more or less “than expected” means, and this meaning may differ depending on one’s specific research context. In other words, because a given individual’s reactivity is judged against the sample regression line, the overall association between X and Y will influence interpretation of the RS. Substantively, this means that the RS cannot be interpreted simply as, for example, vagal augmentation or suppression, because the absolute meaning depends on predicted values set by the sample’s responses.

Table 1

<table>
<thead>
<tr>
<th>Factor</th>
<th>Definition</th>
<th>Influence on reliability/validity</th>
<th>Research recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{XY}$</td>
<td>Correlation of baseline and challenge</td>
<td>Higher $r_{XY}$ raises DS validity but lowers DS reliability$^a$</td>
<td>Report $r_{XY}$, $\lambda$, and their comparison; use this information as part of DS/RS judgments; latent variable models preferable</td>
</tr>
<tr>
<td>$\lambda$ (Lambda)</td>
<td>Ratio of baseline variance to challenge variance ($\text{var}_X / \text{var}_Y$)$^b$</td>
<td>Other things equal; if $r_{XY} &lt; \lambda$, RS more reliable; if $r_{XY} &gt; \lambda$, DS more reliable</td>
<td></td>
</tr>
<tr>
<td>$r_{XZ}$</td>
<td>Correlation of baseline with outcome</td>
<td>Higher $r_{XZ}$ values increase Type I (positive $r_{XZ}$) or II (negative $r_{XZ}$) error rates in the association between RS and Z$^c$</td>
<td>Consider using DS rather than RS if there is a strong association between baseline assessments and the outcome; latent variable models preferable</td>
</tr>
<tr>
<td>$r_{YZ}$</td>
<td>Correlation of challenge with outcome</td>
<td>Other things equal, differing $r_{XZ}$ and $r_{YZ}$ values increase validity of DS$^d$</td>
<td></td>
</tr>
<tr>
<td>$r_{XX}$</td>
<td>Reliability of baseline</td>
<td>Higher reliability of each component increases reliability of both DS and RS</td>
<td>Calculate and report reliability of X and Y; employ designs maximizing reliability (e.g. multiple tasks, standardizing conditions)$^e$</td>
</tr>
<tr>
<td>$r_{YY}$</td>
<td>Reliability of challenge</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: $X =$ baseline; $Y =$ challenge; $Z =$ outcome (variable to be associated with reactivity).

$^a$ May and Hittner (2010).
$^c$ Jamieson (1994).
$^e$ Kamarck, Jennings, and Manuck (1993).
More generally, because RSs are calculated based on the regression line of a particular sample, the same RS may have a different interpretation across samples. Fig. 1 depicts a hypothetical situation demonstrating this point. The top panel shows a target individual (depicted by an X symbol) in a community sample that shows a relatively high association between X and Y. The bottom panel shows the same individual in a clinical sample with attenuated association between X and Y but keeping his or her pattern of individual reactivity scores. Comparison of the two graphs show different RSs based solely on the changed sample. This could be seen as a disadvantage since in contrast to the DS, the RS does not allow for direct comparison of “raw” reactivity across different studies. On the other hand, given the lack of established absolute thresholds for what constitutes “high” (or “low”) reactivity, an RS may be judged to provide more meaningful information than other measures precisely because of its reference to the particular sample under investigation.

Although DSs and RSs are both mathematical combinations of the same two variables, they differ in how they “weight” X and Y scores for the prediction of a third variable. Assuming equal variances of X and Y, the DS weights the two scores equally, since one can rewrite Eq. (1) algebraically as a regression equation with weights of +1 for Y and −1 for X: \( DS = 1Y + (-1)X \). This visualization of the DS as an outcome predicted by X and Y with an added restriction of equal regression weights (with opposite sign) leads to the interpretation of a DS as a simple gain score.

**Fig. 1.** Illustration of changing regression slopes on the residual score.
Conversely, the RS "weights" the challenge score $Y$ more heavily, especially when the correlation between $X$ and $Y$ is low. As the correlation between $X$ and $Y$ approaches zero, the RS approaches the simple difference between $Y$ and its sample mean (Heath & Oken, 1965). Mathematically, because the regression weight of $Y$ on $X$ ($b_{YX}$) approaches zero as their correlation approaches zero, substituting a zero for the $b_{YX}$ term in Eq. (2) results in: $RS = Y - \bar{Y}$. This demonstrates another way to think about the meaning of Eq. (2): namely, the RS always removes observed baseline score variability from the challenge score. Thus, the RS and variants of it have been described as "baseline-corrected" change scores. However, since baseline scores always contain some amount of measurement error, removing observed baseline variance from challenge scores creates a more complicated situation than what is often described as simply a baseline correction. For example, depending on the nature of the error, the sample correlation between the RS and an outcome of interest could be either inflated or suppressed relative to the true correlation in the population.

Despite their differences, DSs and RSs are closely linked, which is not surprising given their joint derivation from $X$ and $Y$. A DS can be thought of as a special case of an RS, in which the regression line has an intercept of zero and slope of one, describing a diagonal line through the origin of a plot. In turn, the RS is akin to a DS where individuals' baseline scores are first equated statistically, in a similar fashion to statistically equating groups on a covariate in an ANCOVA. In other words, an RS can be thought of as a weighted DS, where scores are weighted by the regression line. Although often highly intercorrelated in applied research, DSs and RSs can still have varying patterns of correlations with other variables.

Generally speaking, use of the observed DS to represent reactivity is common in the developmental psychophysiology literature and often not accompanied by specific justification (although cf. Roisman, 2007, who discusses some of the statistical issues involved). Use of the observed RS is more commonly justified, but unfortunately the justification can include questionable or unstated assumptions, or incorrect statistical information. Commonly, the RS is justified based simply on an observed correlation between baseline and DS, which can result from a variety of factors that are not necessarily problematic for analysis. At other times the RS is described as inherently more reliable than the DS. As detailed below, this is not the case.

**Outliers**

It is somewhat surprising that the extensive psychometric literature on DSs and RSs has given little explicit attention to the topic of outliers. It is clear that psychophysiological researchers, like other data analysts, are aware of the potential importance of outliers on linear regression models. Informal inspection of research reports shows that researchers do sometimes report discarding or trimming extreme values on psychophysiological assessments. However, this type of reporting is not uniform. More formally, Rogosa et al. (1982) noted that because RSs are based on the least-squares regression of challenge on baseline, they can “make atypical points seem typical” (p. 739) via small residuals, while potentially inflating residuals of nonoutliers. Therefore, Rogosa et al. recommended that robust regression methods be applied in this context, which typically entails estimating a regression equation using methods other than ordinary least squares, and calculating residuals from that equation. To our knowledge, this has rarely, if ever, been done in applied psychophysiological research. Of greater importance, a systematic investigation of the influence of outliers on DSs and RSs has not been conducted.

**Reliability**

Comparing and contrasting DSs and RSs becomes more complicated when considering psychometric reliability. Reliability requires making a distinction between true scores and observed scores. From classical test theory, observed scores ($O$) equal true scores ($T$) plus measurement error ($E$):

$$O = T + E$$

Reliability, commonly denoted as $r_{OY}$ for the reliability of an observed score $O$, is defined as the proportion of variance in observed scores that is attributable to true scores:
Use of observed scores requires researchers to infer reliability using indirect methods. One common method, coefficient alpha, requires multiple items or scores per variable (Cronbach, 1951). Alpha for physiological data can be estimated using multiple baseline periods of assessment, or by dividing continuous measurements into multiple epochs (Stemmler, 2003).

Intuitively, one might suspect that reliability of a DS or RS depends on the reliability of X and Y, and this is true (e.g., Zimmerman & Williams, 1982; Table 1, above). However, the situation is made more complex by the often-present correlation between X and Y. Mathematical derivations and simulation studies show that the reliability of both the DS and the RS decreases somewhat as the correlation between X and Y increases. Therefore, accurate estimation of the reliability of the DS or the RS using formulas derived from classical test theory requires as input X and Y variances, reliabilities, and intercorrelation. In addition, some mathematical properties of DS and RS reliability are counterintuitive. For example, because statistical power is a function of observed variance rather than true variance, it is possible for a DS to have low reliability but high statistical power for rejecting a null hypothesis of no change (Overall & Woodward, 1975, 1976; Zimmerman, Williams, & Zumbo, 1993).

Applied studies on reliability of reactivity to stress are fairly rare, with some notable exceptions. For example, Kelsey, Ornduff, and Alpert (2007) examined internal consistency of DS over epochs both within and across tasks, generally finding high levels of within-task reliability but lower levels of across-task reliability. This finding is particularly important for developmental researchers, given that stressors used in psychophysiological research necessarily change based on the age and developmental stage of participants. Thus, it would be advisable to keep using the same or similar tasks as long as they remain developmentally appropriate.

In addition, Kamarck, Jennings, and Manuck (1993) provide a rich discussion of methods to increase reliability in studies of physiological reactivity. For example, they stress the importance of considering baseline measurements as a control condition rather than a “zero point”. Thus, they advocate making baseline and challenge assessment periods as similar as possible in all ways other than the target stressor or manipulation. This might include task-specific baseline assessments designed to control for extraneous physical movements (Bush et al., 2011). For example, a challenge task that requires standing and talking, such as a child version of the Trier Social Stress Test, should be matched with a control task that requires non-stressful standing and talking (e.g., naming colors). Such control of body movement may be of particular relevance in developmental work, given the higher overall activity level of youth relative to most adults.

Initial mathematical studies of DS reliability highlighted potentially low reliability as compared with the reliabilities of X and Y (Linn & Slinde, 1977; Lord, 1956, 1963). However, these early criticisms almost always focused on situations where X and Y themselves have both equal variances and equal reliabilities, conditions which place an upper limit on DS reliability (Rogosa, 1988). Subsequent examinations showed acceptable reliability values of DS in cases where these restrictions are relaxed (Rogosa & Willet, 1983; Zimmerman & Williams, 1982). Still, the misunderstanding that DS reliability is low under most conditions has persisted in some sources (e.g., Hauser-Cram & Krauss, 1991).

A systematic program of research by Williams, Zimmerman and colleagues has studied conditions under which the reliability of the DS and the RS may differ. Early work demonstrated that both scores can show reasonable reliability (>0.80) using real-life data for which multiple test forms were available (Williams, Zimmerman, Rich, & Steed, 1984; Zimmerman & Williams, 1982). Subsequent work focused on conditions under which the DS is more or less reliable than the RS (Williams, Zimmerman, & Mazzagatti, 1987), which hinges on comparing the baseline-challenge correlation ($r_{XY}$) to the ratio of observed standard deviations between baseline and challenge, termed lambda ($\lambda$):

$$\lambda = \frac{SD(X)}{SD(Y)}$$

Williams et al. demonstrated that when $r_{XY}$ is less than lambda, the RS has greater reliability than the DS; when $r_{XY}$ is greater than lambda, the DS has greater reliability than the RS; when $r_{XY}$ and lambda are equal, the two scores have equal reliability (see also Llabre, Spitzer, Saab, Ironson, & Schneiderman, 2013).
Thus, as noted in Table 1, given that $r_{XY}$ is typically expected to be at least moderate in magnitude, the DS will have greater reliability than the RS when $Y$ has substantially greater variability than $X$. For example, given an $r_{XY}$ of .5, the variance of $Y$ must be greater than twice the variance of $X$ for the DS to have a larger reliability coefficient than the RS.

Validity

The validity of a DS and RS against an arbitrary covariate or outside criterion variable, here denoted by $Z$, has also been a focus of psychometric research, and is of clear applied relevance. In fact, many developmental psychophysiological researchers are more interested in finding associations between reactivity and outside variables than in examining the reactivity scores per se. To date, formal mathematical derivation of high and low validity conditions has been conducted for the DS only, while simulation studies have examined the validity of both the DS and the RS.

Gupta, Srivastava, and Sharma (1988) noted that the validity correlation between a DS and $Z$, which we term $r_{DZ}$, is uniquely determined under classical test theory by four values: $r_{XY}$, $r_{XZ}$, $r_{YZ}$, and lambda (i.e., the ratio of observed baseline and challenge standard deviations). They considered a variety of combinations of these values (e.g., both positive and negative values for the three correlations noted above) and explored their consequences. Although it is difficult to reduce their conclusions to a brief statement, generally $r_{DZ}$ validity will be higher when the correlation between the baseline score and the outcome ($r_{XZ}$) differs from the correlation between the challenge score and the outcome ($r_{YZ}$) and when $X$ and $Y$ have unequal variances (i.e., lambda ≠ 0). Subsequent research on this topic has drawn attention to the counterintuitive finding that some conditions leading to higher reliability of the DS, such as lower $r_{XY}$, are the same conditions which lead to lower $r_{DZ}$ validity, and vice versa (May & Hittner, 2010). One clear implication of this finding is that researchers should do everything possible to maximize the reliability of $X$ and $Y$, since higher baseline and challenge reliability will increase reliability and validity of both DSs and RSs (see Table 1).

Although to our knowledge a full mathematical treatment of RS validity has not been laid out, Jamieson (1994) has provided a simulation study comparing the DS and the RS against a criterion $Z$. In this study, simulated $X$ and $Y$ scores were created by combining true score and random error components, as well as a constant representing change. Then, a third variable ($Z$) was created and correlated with both baseline and true change at varying levels by adding a function of both $X$ and $Y$ scores as well as random error to $Z$. Results suggested that when $X$ and $Z$ were correlated, RS methods either overestimated (increased Type I error) or underestimated (increased Type II error) the relation between true change and $Z$. This was further dependant on whether or not the $X – Z$ association was in the same or different direction from the true change-$Z$ association. This effect was subsequently referred to as the “regression bias” by Cribbie and Jamieson (2000). Jamieson (1994) notes that the regression bias occurs due to the difficulties in removing effects of a score such as $X$ from a third variable when part of that score is error variance. Latent variable methods, discussed further below, help mitigate this potential problem.

In contrast, in the Jamieson (1994) simulation study the DS was not affected by the regression bias. Based on these results, one may be tempted to draw the conclusion that the DS is preferable to the RS in situations where baseline and challenge scores are correlated. However, subsequent simulation work demonstrated that the RS can fare better, at least in terms of statistical power, when there is a dependency between true scores of baseline values and true change as defined in a simulation (Jamieson, 1995). These RS-preferable situations can be identified by a test of differing variances between $X$ and $Y$ (Geenen & van de Vijver, 1993). More specifically, situations where variance decreases from $X$ to $Y$ suggest greater potential problems for the observed DS relative to the RS, because they imply that the DS confounding of baseline and true change may be due to something besides measurement error.

The Law of Initial Values

A substantial body of psychophysiological research has been conducted on the Law of Initial Values (LIV)—the proposition that due to biological constraints on specific physiological systems, one's
baseline status on a physiological measurement will be associated with the amount of change occurring in response to a laboratory challenge (Wilder, 1931). Typically the association is negative, although evidence for positive LIV effects has occasionally been found (Furedy & Sher, 1989). Given a negative association, children who show very high baseline values tend to display smaller increases in response to a challenge, presumably because they are already physiologically aroused. Although LIV effects may seem straightforward to investigate—calculate change scores and then correlate baseline scores with change—in practice LIV effects are confounded with the statistical phenomenon of regression to the mean. Regression to mean, which occurs whenever repeated measurements are made on scores with random measurement error (Lord, 1967), also produces negative correlations between baseline scores and change scores. More specifically, given imperfect measurement, children who score very low at baseline will tend to score somewhat higher following a laboratory challenge simply due to fluctuations in random error. These fluctuations are conceptually separate, but statistically indistinguishable from biological constraints of the LIV. Regression to the mean will be present in all psychophysiological assessments to the extent that measurement error is also present. Therefore, recent LIV research has explored statistical means of differentiating these effects, while acknowledging that they are not incompatible (Geenen & van de Vijver, 1993).

The LIV phenomenon has sparked continuing debate and discussion in medical research, where treatment response over time for two or more groups is often a goal of analysis. Although the two-group scenario is not the focus of this review, recommended means of addressing this issue include correcting baseline values for measurement error before making group comparisons if reliability is known (Blomqvist, 1977) and comparing groups on the correlation between change (DS) and the mean of baseline and challenge scores (Oldham, 1962), which is essentially equivalent to the test of differing X and Y variances noted above (Geenen & van de Vijver, 1993). Tu and Gilthorpe (2007) provide more detailed discussion of the two-group treatment-response case.

Returning to single-group designs, Stemmler (2003) provides a cogent explanation of how the LIV relates to DSs, RSs, and regression to the mean, assuming one has reliability information available on X and Y. This is accomplished by distinguishing between two types of initial value dependency (IVD): “true IVD”, representing a real effect consistent with the LIV, and “statistical IVD”, representing regression to the mean (see also Foerster, 1995). Once this distinction is made, Stemmler notes that the DS can be mathematically split into three independent elements: (1) statistical IVD (which always equals zero if there is perfect measurement), (2) true IVD (true change that is related to baseline), and (3) the residual score (change not related to either IVD). Stemmler notes that which elements one might include in one’s statistical analysis depend on substantive considerations of the applied researcher and there are not easy rules to recommend one over the other. Here as well, latent variable methods provide a potential way forward, because they require the applied researcher to specify their model for change in more detail: as discussed further below, the distinction between baseline-associated and baseline-independent true change is sharply drawn by contrasting latent difference score and latent residual score models.

Further complicating the study of the LIV are findings suggesting operation of the “law” may vary depending on several aspects of the physiological system under study. For example, Hord, Johnson, and Lubin (1964) showed evidence for LIV-consistent results for heart rate and respiration rate in adults, but not for galvanic skin response. Berntson, Uchino, and Cacioppo (1994) demonstrated that LIV-consistent findings can vary depending on the source of baseline variability, especially when autonomic and non-autonomic influences on heart period are controlled experimentally, which was accomplished in their study through a posture manipulation where participants moved from sitting to standing positions (or vice versa). Thus, LIV findings may vary by system, experimental design, or within- versus between-subject analysis (Scher, Furedy, & Heslegrave, 1985). Moreover, Jin’s (1992) detailed discussion of the LIV notes that the LIV can encompass several different patterns of change from baseline to challenge (e.g., manipulations that reduce response rather than increase it). While our focus here is on the statistical issues related to the LIV, many of which are also reviewed by Jin (1992), we echo the call by Berntson et al. (1994) for researchers to consider in more detail what influences on physiological response they are attempting to manipulate in experimental designs.

Psychometrically, two broad points seem evident from the above discussion. First, researchers should strive to estimate true change rather than observed change, accounting for measurement error
when reliability estimates are available. Denoting the reliability of baseline $X$ as $r_{XX}$, an equation for the true difference score is as follows (Stemmler, 2003):

$$TDS = Y - r_{XX} (X - \bar{X}) - \bar{X}$$

(6)

Second, for physiological systems where the LIV may be expected to operate, researchers should consider more explicitly how they wish to represent individual differences in baseline scores. It is possible that situations in which one desires to incorporate individual differences in baseline would appear to favor use of the DS, whereas situations in which one desires to statistically equate baseline levels as part of the analysis of reactivity would appear to favor use of the RS. At the same time, it is important to remember that the LIV is neither a universal law nor relevant to certain psychophysiological research questions, such as determining overall group levels of reactivity (Geenen & van de Vijver, 1993), and that further exploration of the source of baseline differences is important (Berntson et al., 1994).

A program of Monte Carlo research by John Jamieson has investigated the statistical conditions that affect the magnitude of the LIV. Initial work demonstrated that effects on the correlation between baseline and change can be grouped into two major categories. The first category is measurement error, which affects regression to the mean. The second category is any factor that causes a change in variance from $X$ (baseline score) to $Y$ (challenge score), such as a skewed distribution of baseline values as well as floor or ceiling effects (Jamieson, 1993; Jamieson & Howk, 1992). As variance increases from $X$ to $Y$, the association between baseline and change tends to be positive, as one might expect from a “fanning out” distribution of scores where those individuals with higher initial status are also the ones who show the largest increase. Because this effect operates opposite to regression to the mean—which serves to “narrow in” the distribution of scores—the overall association between baseline and change could be positive, negative, or zero depending on the relative strength of each effect. The effect of changing variance on the correlation between baseline and change is illustrated graphically by Fig. 1 of Jamieson (1993, p. 236).

Although this type of descriptive finding is likely familiar to developmental researchers, Jamieson (1993) is informative in noting the great difficulty in applied research of statistically separating out “true LIV” effects from other issues such as skewness and variance change from $X$ to $Y$ for reasons other than the LIV. More recent work by Jamieson (1999) showed that in a multiple-group design such as a randomized controlled trial, the combination of group differences on a baseline measure and skewness allows one to predict, from statistical considerations alone, which group will show greater change, leading to potential dilemmas for applied researchers. As an example, given positively skewed baseline scores and a mean difference in groups at baseline, the group with the higher baseline mean will also show greater increase in response to a stressor (and a greater decrease in response to a “relaxer”); Figs. 2 and 4 of Jamieson (1999, pp. 157–159) present graphical depiction of these effects. Overall, this research suggests that investigators should be mindful of potential consequences of skewed data on their measurements of change: ideally, decisions about transforming skewed data should be done prior to other analyses of reactivity. If not, researchers may be tempted to make transformation decisions based on known implications for finding larger or smaller differences, risking inflated Type I error rates.

Not all developmental psychophysiological researchers will be focused on the Law of Initial Values, and there is debate as to how extensive LIV effects are in practice as well as their theoretical meaning. Our goal here is not to resolve these debates. Instead, we hope that developmental psychophysiological researchers will take the research cited above into account when making decisions about how to measure and interpret reactivity. First, researchers should make decisions about transforming data prior to other analyses. Second, where possible researchers should test for a decrease in variance from $X$ to $Y$; results of such tests suggest that the LIV may be rarer than is commonly appreciated (Geenen & van de Vijver, 1993). At the same time, researchers should note that a statistically significant decrease in variance can result from factors other than the LIV (Jamieson, 1993).

**Independent treatment of baseline and challenge scores**

Recently, developmental physiological researchers have emphasized the importance of examining how children’s baseline or anticipatory levels of arousal—in addition to physiological reactivity—might...
contribute to adaptive functioning. For example, a recent study by Hinnant and El-Sheikh (2009) highlights the importance of examining the interaction between baseline and reactivity scores when examining children’s risk for externalizing and internalizing symptoms. This can be accomplished in a variable-oriented framework by including the baseline variable in a regression equation along with the DS or RS. Note, however, that a “baseline + DS” regression model is mathematically equivalent to a regression model including the baseline and challenge variables as separate predictors of a given outcome (Jamieson, personal communication, January, 2012). More specifically, the regression coefficient for the challenge score in one model equals the regression coefficient for the difference score in the other model; the regression coefficients for the baseline variable differ across models, while the intercept term and overall $R^2$ are the same. Based on the principle of parsimony (Sober, 2006), one might argue that the separate predictors model is preferable.

On the other hand, a “baseline + RS” regression model represents something slightly different, namely partialing out baseline variance from the outcome variable (since the baseline variable and RS are uncorrelated by definition). This approach allows researchers to test the association between reactivity and a third variable free of the influence of individual differences in baseline arousal, given an important assumption of no correlated measurement error between the baseline score and the third variable. Additional theoretical work would be helpful to clarify the particular contexts in which this type of statistical control or joint prediction is desirable.

Turning briefly to applied domains outside of psychophysiology, extensive work from organizational psychology proposing alternatives to DSs has emphasized including the interaction between the components—here, baseline and challenge—as well as squared terms of each in the equation (see Edwards, 1994, 2001, for summaries). While this recommendation was derived from statistical concerns on the interpretability of regression coefficients, it is also of interest given recent developmental theories that emphasize testing quadratic, non-linear associations between physiological reactivity and developmental outcomes (Boyce & Ellis, 2005; Del Giudice, Ellis, & Shirtcliff, 2011; Obradović, 2012).

In a similar vein, work from clinical psychology on informant discrepancies has advocated inclusion of a mean or sum score, representing the combined effect across informants, in regression models that use DSs (Laird & Weems, 2011). As described by Laird and Weems, use of this approach highlights a surprising conclusion: DS effects on outcomes can be conceptualized as a relative difference in the predictive power of the components of the DS, as assessed by the sign of the DS regression coefficient in an equation that includes the mean/sum score. In the psychophysiological context, this represents a differing correlation between baseline and an outcome, on the one hand, and challenge scores and the same outcome, on the other.

**Summary**

An extensive history of psychometric research has outlined the various conditions that are likely to lead to higher versus lower reliability and validity of DSs and RSs, which are summarized in Table 1. Some factors are common to both scores: for example, lower reliability of X and Y will, not surprisingly, serve to lower the reliability of both the DS and the RS. Other factors help to distinguish situations in which one type of change score may be more or less reliable than the other.

Research on the validity of DSs and RSs suggests that close attention needs to be paid to the variability of baseline scores and challenge scores, as well as their association. Regarding the former, the DS seems to fare better in situations where the variance of challenge scores is substantially (i.e., 2x) larger than variation at baseline, in a similar pattern to the findings on reliability. If the variances of X and Y are roughly equal, then higher correlations between X and Y are generally associated with higher DS validity, but lower DS reliability (May & Hittner, 2010). Ultimately, judgments about high versus low validity are dependent on the overall pattern of correlations between X, Y, and the outcome variable of interest (Gupta et al., 1988).

Questions of reliability and validity are inherently linked to measurement error. This is especially important for research using reduced scores such as the DS and RS. The DS combines error from both X and Y, which cannot later be disentangled. On the other hand, the RS does not truly control for initial value, because some of the baseline variance that RS “controls” is error. Thus, it is crucial for
developmental psychophysiology researchers to think carefully about errors of measurement. One way to do so is by specifying latent variable models, to which we now turn.

**Latent variable approaches to the study of reactivity and change**

Latent variable models, including confirmatory factor analysis (CFA) and structural equation modeling (SEM), are statistical approaches in which multiple observed indicators are obtained in order to assess a desired latent construct (Brown, 2006; Kline, 2010). By solving systems of linear equations using iterative methods, latent variable software provides a powerful and flexible method of answering important psychophysiological research questions. A key strength of latent variable methods is their flexibility in integrating modern methods of dealing with missing data, through either multiple imputation or, more commonly, maximum likelihood estimation of model parameters (Schafer & Graham, 2002). Briefly, maximum likelihood estimation involves an iterative process in which computer software attempts to converge on the set of model parameters which, taken together, best reflect the observed variances and covariances in a set of data (see Kline, 2010; Long, 2012). However, we should note that the assumptions of maximum likelihood estimation are sometimes antithetical to small sample sizes, and because latent variable models result from iterative optimization, models can fail to converge. As with any research study, it is important to conduct statistical power analyses prior to collecting data. Power analysis for latent variable methods is an active area of research, with MacCallum, Browne, and Sugawara (1996) providing one potential methodology. Muthén and Muthén’s (2002) simulation approach to statistical power for latent variables is highly recommended, as it encourages the researcher to explore *a priori* the effect of different expected parameter estimates on the power and precision of the overall statistical model.

Importantly, using latent variables allows researchers to both explore the measurement properties of their physiological markers as well as obtain estimates of associations among constructs controlling

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Research recommendation: consider when...</th>
</tr>
</thead>
</table>
| LDS   | Estimates mean change, correcting for measurement error | • Testing indirect effect from baseline → reactivity → outcome  
• Testing whether mean change is greater than zero  
• Desire to interpret based on original metric of assessment |
| LRS   | Estimates baseline-independent change, correcting for measurement error | • Interest is in variance of mean change  
• Differentiating baseline-outcome relations from reactivity-outcome relations  
• Defining reactivity as baseline-independent change  
• Interest is in the variance of baseline-independent change |
| LGC   | Estimates latent intercept and various change factors: e.g., linear slope, quadratic curvature | • Estimating the overall shape of change over time  
• Differentiating between linear, quadratic, and cubic change  
• Using three or more distinct time points or epochs |
| LST   | Estimates latent state and trait factors | • Replicate indicators at each time point are available  
• Interest is in separating out stable trait factor from time-specific state factors  
• Using four or more distinct time points or epochs |

Notes: The methods in the table are not exhaustive, and all methods can be adapted for extensions such as multiple-group comparisons or inclusion of additional exogenous variables. See text for further details.

LDS = latent difference score; LGC = latent curve; LRS = latent residual score; LST = latent state-trait.
for measurement error. We have emphasized the crucial role of measurement error in preceding sections, and a significant advantage of latent variable methods over traditional analyses is their ability to explicitly model measurement error using multiple observed indicators of a variable.

In this section, we provide descriptions of four major approaches to studying reactivity using latent variables: latent difference score modeling, latent residual score modeling, latent growth curve modeling, and latent state-trait modeling. Each of these techniques requires modeling two or more latent variables, including (in some cases) latent baseline variables and latent challenge variables. In turn, each latent variable requires multiple observed indicators—or, in some cases, a single indicator with a specified reliability level. Table 2 summarizes the major distinctions among these models as they relate to the study of physiological reactivity.

**Latent difference score (LDS) modeling**

Latent difference score modeling is conceptually analogous to using a difference score with observed data, but allows the researcher to explicitly model measurement error through SEM. Within SEM, measurement models—the portion of the model relating latent variables to observed or manifest indicators—typically include multiple indicators of each construct. In psychophysiological research, there are several possible options for inclusion of multiple indicators. For example, researchers can include multiple baseline and challenge scores that each represent a task- or context-specific psychophysiological assessment. Conversely, if continuous measurements are made over a period of time, researchers can split their measurement into epochs, treating an average score within each epoch as a separate indicator. When only one score for baseline and challenge is available, researchers can still model measurement error by including a numeric estimate of the reliability of their observed variable, which should be obtained from prior research. However, the complexity of latent difference score models generally requires at least two observed indicators for both baseline and challenge assessments in order to avoid problematic model constraints (Raykov, 1992). The relative strengths and weaknesses of these alternative approaches to measurement models have yet to be systematically examined.

Relative to other latent variable models, the hallmark of an LDS model is the specification of a second-order latent variable that represents change between the first and second time points of assessment. That is, LDS models first specify latent baseline and challenge factors, and then add a second-order latent factor lacking any direct manifest indicators. Instead, this factor is defined by the paths included in the remainder of the model. In a standard LDS model, the latent challenge factor (Y) is regressed on both the latent baseline factor (X) as well as the latent change factor, and the two structural regression weights are set to equal 1.0. This constraint, implicit in Eq. (1) for observed scores, defines the change factor as a true latent difference from baseline to challenge. In addition, the baseline-change association is estimated. Thus, the LDS is equivalent to an observed DS but accounts for measurement error in both X and Y. An example LDS model is shown in Fig. 2, with corresponding syntax presented in the Appendix.

Table 2 lists some situations where the LDS model may be appropriate. For example, researchers may be interested in testing whether true change is different from zero (or any other specific value): that is, what degree of reactivity was observed in the sample as a whole? This is potentially important for work testing new stressors or challenges, as well as developmental work seeking to compare average responses of different age groups to a given task or stressor.

Second, by adding an outcome to the model, the LDS allows one to test mediation of the baseline-outcome variable association by the change factor. Researchers can thus specify a path from baseline, through reactivity, to an outcome of interest and test whether a baseline-outcome association remains after accounting for stress reactivity.

Although LDS applications are visible in many areas of psychology, these models have not been widely incorporated into research on psychophysiological reactivity. One exception using simulated data is work by Cribbie and Jamieson (2000, 2004), extending earlier Monte Carlo research conducted by Jamieson described above. Cribbie and Jamieson contrast an LDS model of change with basic DS and RS analyses while varying the relations between a third variable and simulated change. Their results confirmed that the LDS approach reproduces accurate estimates of the change-third variable relations,
with greater statistical power than an observed DS approach. Although this study used simulated data, it suggests that LDS methods should be more widely adopted in the study of psychophysiological reactivity, since they can counteract pernicious statistical biases in one’s data that lead to spurious relations between a third variable and change when the third variable is also associated with baseline scores.

Latent residual score (LRS) modeling

Specification of a latent residual score (LRS) model is conceptually similar to a latent difference score model, with key differences that reflect the relations between observed DSs and RSs: just as the LDS model can be considered a latent version of the DS, the LRS model can be considered a latent version of the RS. The primary difference between the two models is that whereas the LDS specifies the change factor and baseline to contribute equally to a latent challenge factor \(Y\) by fixing those two structural paths to a value of 1.0, the LRS allows the structural regression path from the baseline factor to the challenge factor to be freely estimated (Little, Bovaird, & Slegers, 2005). This allows the path to take on values other than 1.0, which defines the second-order latent factor as a residual factor rather than a simple difference factor. At the same time, the bidirectional path between the baseline factor and the second-order factor is constrained to zero, which represents the “baseline-free” nature of residual change. The LRS model is depicted in Fig. 3, with the corresponding Mplus input syntax included in the Appendix.

It is important to note that the LDS and LRS are statistically equivalent models (MacCallum, Wegener, Uchino, & Fabrigar, 1993). That is, for any given set of data, they will provide the same reproduced covariance matrix and therefore the same numeric goodness-of-fit indices. Nonetheless, each model partitions the variation related to \(X\) and \(Y\), and their latent means, into distinct second-order change factors. The choice between these two models is likely to depend on whether a given researcher is most interested in modeling change that occurs independent of baseline values (which favors the LRS) versus modeling change that incorporates baseline differences directly (which favors the LDS).

We are not aware of any published substantive applications of LRS models to date. As suggested in Table 2, one example of a situation where the LRS model may be particularly useful is in differentiating baseline-outcome relations from reactivity-outcome relations. That is, researchers may be interested in whether an outcome of interest is more related to resting physiological activity versus physiological reactivity, and whether such associations might vary by age. By defining reactivity as baseline-independent change, a stronger theoretical contrast is made between these two competing associa-

---

**Fig. 2.** Latent difference score (LDS) measurement of reactivity (Little et al., 2005; McArdle, 2001, 2009). \(X_1, X_2,\) and \(X_3\) are observed baseline measurements. \(Y_1, Y_2,\) and \(Y_3\) are observed challenge measurements. \(X\) and \(Y\) are latent baseline/challenge constructs. Delta \(\Delta\) is the latent difference score. The triangle at upper left represents a constant; regressing on a constant estimates means in addition to covariances. Paths denoted “1” or “0” are constrained to that number during estimation. Factor loading and error variance parameters are omitted, as are residual covariances and equality constraints among parallel indicators (e.g., \(X_1 - Y_1\)). Estimated parameters: \(a = \) mean change; \(b = \) variance of change; \(c = \) baseline-change correlation.

---
As an applied example, developmental researchers are finding that baseline arousal may be more predictive of certain outcomes if the baseline scores reflect children’s anticipation of participating in the study or body’s preparedness to respond to challenges (El-Sheikh & Whitson, 2006; Kestler & Lewis, 2009; Obradović, 2012). LRS models provide a means of better differentiating this process from reactivity to the stressor itself. Clearly, much more exploration and testing of these models are warranted. In addition, depending on the exploratory versus confirmatory nature of a given study, researchers may wish to report results of both LDS and LRS models so as to underscore the theoretical implications of defining change as baseline-independent versus baseline-associated in a given substantive domain.

Latent growth curve (LGC) modeling

The approaches detailed above represent situations in which multiple observations are reduced to a single latent score of primary interest. However, these are not the only latent variable models of potential relevance for psychophysiological research. When three or more discrete time points of data are available, growth curve modeling with covariates (Meredith & Tisak, 1990; Rogosa et al., 1982) can be a powerful method for examining theories of psychophysiological reactivity. In a standard growth curve model, latent factors are specified that represent particular aspects of change over time, such as linear change and quadratic or curvilinear change. As noted in Table 2, estimates of starting point (intercept) and change over time can be determined, model fit to data evaluated, and associations of change with outcome variables examined. When continuous or multiple-epoch assessment of physiological data is available, a potential strength of growth curve models is their incorporation of several observed data points into one’s statistical model. Preacher, Wichman, MacCallum, and Briggs (2008) provide an accessible and general introduction to latent growth curve modeling; an example growth model of reactivity, based on a model from Llabre, Spitzer, Saab, and Schneiderman (2001), is shown in Fig. 4, with Mplus input syntax in the Appendix.

LGC models allow one to test relations between change over time and various outcomes of interest. As one example of this technique, Brooker and Buss (2010) examined how linear and quadratic trends in respiratory sinus arrhythmia (RSA) predicted observed positive affect in toddlers during a stranger-interaction task, finding that such trends contributed unique predictive information over observed difference scores. In addition, the flexibility of growth curve models allows for several extensions: for example, Van Ryzin, Chatham, Kryzer, Kertes, and Gunnar (2009) demonstrate semiparametric mixture modeling of cortisol data, in which multiple models are fit to the same dataset representing distinct patterns of change for subgroups within the sample. Conversely, Llabre et al. (2001) present an...
example of piecewise growth modeling in an adult sample, in which separate models are fit to data collected before and during (or after) a laboratory stressor, measuring blood pressure reactivity and recovery periods following a cold pressor task. This particular extension of LGC models has great potential for advancing developmental theory, as it allows one to examine how youth physiological responsivity changes over time as children become better regulators of their broad emotions and behavior. Of interest, although LGC models have been developed in the context of SEM, recent work (see e.g., Curran, 2003) has stressed their statistical equivalence (under many conditions) with multilevel models, also termed mixed models or hierarchical linear models (HLMs). An HLM framework allows for consideration of random effects representing individual variability around a group mean; for example, individual differences in changes in physiological activation over time. Although beyond the scope of the present paper to discuss in detail, HLMs are a powerful method of answering questions closely related to those explored here. A cogent and accessible article describing how to use HLM to answer questions about physiological reactivity over time is provided by Kristjansson, Kircher, and Webb (2007).

Latent state-trait (LST) modeling

Like LDS and LRS modeling, latent state-trait modeling seeks to partition observed variables into latent components that provide meaningful information about change and development (Kenny & Zautra, 2001; Steyer, Schmitt, & Eid, 1999). Table 2 describes the unique contribution of LST models: their focus on the separation between state factors, which influence each time point of assessment, and trait factors, which influence all time points (or a set of linked time points). State factors model any situational influences on the dependent variable: examples include momentary reactions, current mood and health status, responses to experimental manipulations, or even time of day. Because LST models are specified as higher-order factor models, they typically require four or more waves of a dependent variable to achieve statistical identification without imposition of additional constraints. However, these sets of scores can potentially be obtained during a short time interval. An advantage of LST models is the ability to decompose overall variation into state, trait, and error components, allowing researchers descriptive information on what proportion of their measured constructs are stable over time. This explicit focus on stability and change has the potential to inform developmental questions on the changing nature of psychophysiological systems with age; for example, it may be possible to address developmentally-linked artifacts of movement (Bush et al., 2011) via state factors. As with any structural equation model, theoretically meaningful model constraints can be imposed,
such as that of constant state effects. A hypothetical LST model is illustrated in Fig. 5 and corresponding Mplus input syntax is included in the Appendix.

Several psychophysiological applications of LST modeling have been published, with many focusing on the stress hormone cortisol, an index of slower-acting hypothalamic–pituitary–adrenal axis (HPAA) stress response. For example, Shirtcliff, Granger, Booth, and Johnson (2005) assessed cortisol four times over the course of 2 years—twice over 1-day intervals each year—and included two replicate measurements at each point. They used LST modeling to separate out effects of a trait factor (which influences all time points), state factors (influencing a given day) and measurement error (influencing a single replicate). Their findings suggested that while the majority of reliable variance in cortisol assessment was related to state factors, the trait factor was negatively associated with externalizing problems for males. Additional analyses from the same sample found that for adolescent girls, lower levels of trait cortisol were associated with poorer quality social relationships across a variety of relationship types (Booth, Granger, & Shirtcliff, 2008). Kertes and van Dulmen (2012) have also examined cortisol variation using LST methodology, finding roughly equal magnitudes of trait and state factors for both morning and evening cortisol assessments. Of note, Hagemann, Naumann, Thayer, and Bartussek (2002) have shown a similar breakdown of reliable trait and state variance in testing LST models of resting electroencephalograph (EEG) data from adult participants. More broadly, Davey (2001) discusses how LST models can help advance developmental research by focusing attention on better measurement of contextual factors (e.g., short-term interpersonal exchanges) that might help shape state variance. Furthermore, LST models have a potential to better explain the longitudinal stability and change of physiological reactivity across different developmental periods.

**Extensions of latent variable models**

Besides the crucial advantage of correcting for measurement error, the specification of latent variable models allows for important extensions that have promise for advancing developmental theory. First, any structural model can be evaluated across multiple groups, such as gender or age categories, and the equality of various parameters can be tested across said groups (e.g., Farrell, 1994). This allows for powerful tests of developmental theories. For example, an LRS model might be adapted to test reactivity-outcome links across age periods, given meta-analytic evidence of age differences in associations between cortisol reactivity and externalizing behavior (Alink et al.,
Alternatively, an LGC model might be specified to examine developmental differences in slope of a recovery period following a stressor, which may emerge as children develop better self-regulatory skills. Associations between reactivity and a range of outside variables could be tested for invariance across age, lending insight into how reactivity and behavior covary over developmental periods.

Second, interaction terms can be included in latent variable models. Work on latent interaction terms is an active area of research (e.g., Little, Bovaird, & Widaman, 2006; Marsh, Wen, & Hau, 2006), with little application to date to the classes of models reviewed here. However, incorporation and testing of latent interaction effects would allow stronger tests of theories such as biological sensitivity to context (BSC)/differential susceptibility to rearing influence (Belsky & Pluess, 2009; Boyce & Ellis, 2005), which theorize core interactions between physiological reactivity and environmental influences in predicting both positive and negative developmental outcomes. According to such theories, greater reactivity is expected to predict worse outcomes in adverse contexts but better outcomes in thriving contexts (Obradović, Bush, Stamperdahl, Adler, & Boyce, 2010).

Third, all of the aforementioned models can be extended to include multiple core constructs. Bivariate LDS and LRS models examine linked latent differences in two constructs at once: an example is King, King, McArdle, Shalev, and Doron-LaMarca (2009), who examine cross-lagged changes in PTSD and depression (see also McArdle & Hamagami, 2001). Relatedly, bivariate LGC models specify intercept and slope/change factors for two constructs; an example looking at depression and academic achievement that includes comparison to other latent methods is provided by Grimm (2007). Such models have the potential to test theories that propose linked action among multiple, distinct psychophysiological systems. Given growing interest in testing coactivation and/or coinhibition of the SNS and PNS (Alkon et al., 2003; Berntson et al., 1991; Salomon, Matthews, & Allen, 2000) as well as the interaction between ANS and HPAA activity (Bauer, Quas, & Boyce, 2002; Gordis, Granger, Susman, & Trickett, 2006), developmentalists should consider broader use of bivariate LDS/LRS models for operationalizing their theoretical predictions.

The extensions detailed above are not exhaustive. In fact, the model-based analysis of latent variables implies that researchers have a great deal of freedom in testing their theories, with the important caveat that models must converge and should provide reasonable fit to the data. Recently, data analysts have increasingly advocated for approaches to data analysis that focus on comparing a range of competing statistical models using a single dataset (Long, 2012), a general strategy that may be useful for developmental psychophysiology.

**Measurement invariance**

Effective statistical use of multiple-indicator models over time, such as the LDS and LRS, requires longitudinal measurement invariance of the observed indicators (Reise, Widaman, & Pugh, 1993). Longitudinal measurement invariance can be further broken down into three major categories: configural invariance, which assumes a constant factor structure over time (e.g., a single latent factor representing physiological activation), metric invariance, which adds the assumption that the pattern of factor loadings is constant over time (e.g., each observed indicator of activation has the same relationship to the latent factor across time relative to other indicators), and intercept invariance, which adds the assumption that parallel indicators assess the same ranges of the latent construct over time. Invariance can be tested by comparing the fit of two nested models using a chi-square difference test; one model includes assumptions of invariance (e.g., constraining parallel factor loadings over time) and a second model relaxes those assumptions. Both models typically allow the residuals of parallel indicators allowed to correlate over time.

If the invariant model does not show significantly worse fit, the assumption of invariance is supported. This allows interpretation of changes in the latent construct as true change unrelated to patterns of responding for a particular indicator (e.g., LDS as true difference from baseline to challenge). Partial measurement invariance, where some but not all indicators are invariant, is also possible (Byrne, Shavelson, & Muthén, 1989). It is important to note that because of the lack of measurement model applications in the psychophysiological literature to date, the degree to which measurement invariance tends to hold remains an open question, with differing answers possible depending on how manifest indicators are constructed. Ferrer, Balluerka, and Widaman (2008) provide a detailed discussion
of the consequences of lack of invariance for LGC and LDS models. In particular, lacking invariance makes it difficult to interpret the meaning of the change factor, and results can be sensitive to which observed indicator is chosen as a reference indicator in the measurement model. Obradović, Pardini, Long, and Loeber (2007) provide a detailed applied example of longitudinal invariance in developmental psychopathology, and Vandenberg and Lance (2000) give more extensive statistical discussion of the issue. The Appendix syntax for models with multiple observed indicators includes comments indicating how the assumptions of measurement invariance are encoded in the Mplus software program.

Person-centered approaches

In contrast to the variable-centered approaches described in prior sections of this paper, person-centered analysis represents the identification of subgroups of participants who share meaningful profiles (Bergman & Magnusson, 1997). There is a fairly extensive history of person-centered analysis in physiology, often involving cluster analysis of multiple scores from different systems (e.g., Wilson, Lengua, Tininenko, Taylor, & Trancik, 2009) or identification of subgroups based on cut points or patterns of increasing or decreasing response across systems (e.g., Kamarck & Lovallo, 2003).

However, this is also an area where modern latent variable methods, such as latent class analysis (McCutcheon, 1987) and finite mixture modeling (McLachlan & Peel, 2000) can offer advantages over earlier-developed methods, by accounting for error in classification and within-category variability. These methods provide a way to test how patterns of physiological activity across multiple systems might change over time. Applications of these latent models are beginning to appear: for example, Del Giudice et al. (2012) provided a test of the “adaptive calibration model” of individual differences in stress reactivity, examining multiple systems in concert through finite mixture modeling subgroups and finding evidence for different latent classes based on both SNS and PNS activity. The simultaneous focus on multiple systems provides another window to testing Berntson et al.’s (1991) nuanced view of autonomic control described above. Of note, person-centered approaches to physiological reactivity can also be combined with longitudinal assessment if the primary interest is identifying unique patterns of change over time within a sample (Collins & Lanza, 2010; Lubke & Muthén, 2005; Nagin, 2005). This allows identification of, for example, subgroups of participants who show unusually high, low, or fluctuating reactivity over time. These models are conceptually related to the bivariate LGC models described above, with analytic attention focused on latent classes of similar participants rather than latent change estimates for an entire sample.

Summary

Overall, latent variable models are under-utilized in research on psychophysiological reactivity. In particular, their ability to model measurement error and thereby account for unreliability in assessments is a crucial advantage: our review of psychometric issues has highlighted how unreliability can obscure the correct interpretation of observed DSs and RSs, and latent variable methods mitigate this serious problem. Table 2 reviews the basic distinctions among the models described above, and suggests potential research contexts for each. In particular, comparison of LDS and LRS models can highlight the different assumptions made by each class of analysis in partitioning variance over time, which has theoretical implications. For example, adding an outcome variable to the LDS model allows for testing of a latent mediating pathway from baseline through reactivity to the outcome. Conversely, the LRS allows for testing of differential independent relations of baseline and reactivity to the outcome. Unlike the observed RS, the LRS model accomplishes this without biasing the association due to measurement error covariance. LCG and LST models provide their own theoretical tests, the former being particularly useful in contrasting periods of reactivity and recovery from stress.

More general advantages of latent variable methods deserve to be restated. These models can easily take advantage of modern treatment of missing data through estimation methods such as full information maximum likelihood, allowing use of all available information even when some participants are missing data at certain time points (Schafer & Graham, 2002). The flexibility of latent variable models allows for many extensions described above, such as multiple-group analyses and latent interactions. Finally, the
overall approach fits generally with a model-testing philosophy of science—including testing of competing models, where applicable—that is gaining traction in psychology more broadly (Rodgers, 2010).

Conclusion

We conclude with a set of general recommendations for applied psychophysiological researchers based on our review of the statistical and methodological literature on the measurement of change. We divide our recommendations into the major stages of research: planning a study, collecting and analyzing data, and reporting results.

Study planning

We recommend that wherever possible, researchers pay greater attention to the design of baseline assessments in psychophysiological research. For example, some prior work has highlighted the possibility of accounting for individual variability in baseline measurements over time (Malmstrom, 1968). While individual study details will depend on the nature of the system under study and the particular research questions involved, including multiple baseline assessments will almost always strengthen a given study. First, this design allows for formal reliability estimates of the baseline construct. Second, it affords more detailed examination of the role that baseline physiological processes play across different developmental stages and contexts. Analogously, including multiple stressors or challenge periods will generally strengthen a study’s generalizability as well as its feasibility for latent variable analytic methods, providing greater options for analytic approaches.

In addition, it is beneficial for researchers to design and pilot laboratory tasks to ensure that adequate variability of both baseline and challenge scores are obtained. This requires careful attention to the population under study and the expected range of response to selected stressors. In particular, introducing a stressor or manipulation that is too mild or too severe may act to reduce variability in observed reactivity; for example, if nearly all participants’ scores are unperturbed from a biological floor or brought to a biological ceiling (Buss, Goldsmith, & Davidson, 2005; Kamarck, 1992; Obradović, 2012; Quigley & Stifter, 2006). Likewise, pilot data obtained on the correlation between baseline scores and challenge scores can be advantageous, to ensure that this association is not extremely high, impacting the reliability of the DS. More generally, researchers should be mindful of the potential that particular laboratory tasks may elicit particular systems or subsystems, such that lack of observed reactivity may indicate an inappropriate selection of a stressor or a challenge task (Dickerson & Kemeny, 2004; Obradović, 2012). Other classic psychometric principles such as use of multiple “items” (challenges, or epochs) and implementing controls for “noise” or unwanted variance should be employed (Bush et al., 2011; Kamarck, 1992; Kamarck et al., 1993). Finally, plans to transform skewed data should be formed prior to data collection.

Data analysis

If a researcher is making a choice between the observed DS and observed RS, certain conceptual and statistical considerations may serve to influence that choice, and these are summarized in Table 1. For example, research contexts in which reactivity scores relative to a particular sample are desired favor the RS, whereas contexts in which the raw score metric has significance for interpretation favor the DS. Some researchers have explicitly defined reactivity as change that is independent from baseline, which is akin to preselecting the RS (Manuck, Kasprowicz, & Muldoon, 1990). Other things being equal, skewed data favors the RS (Jamieson, 1999), as do contexts in which challenge variance is smaller than baseline variance. Conversely, contexts in which challenge variance greatly exceeds baseline variance may favor the DS, as do studies comparing reactivity across naturally occurring groups (Cribbie & Jamieson, 2000; Jamieson, 1999). Since the RS tends to mask outliers, such scores should be detected prior to core analytic runs and methods taken to analyze them in more depth or otherwise mitigate their influence on the regression model (Cohen, Cohen, West, & Aiken, 2003). Such methods can include the use of robust regression models to calculate RSs (Rogosa et al., 1982). Finally, research-
ers with access to reliability information on their baseline variable may benefit from calculating true difference scores using Eq. (6).

However, despite overwhelming past precedent, it is not clear that the choice between DS and RS is the best way to frame the operationalizing of physiological reactivity. First, work by Edwards (2001) and Laird and Weems (2011), briefly reviewed in the section “Independent Treatment of Baseline and Challenge Scores”, suggests that researchers may be better off including baseline and challenge scores separately, as well as their mathematical sum and/or product, in a single regression equation, with other interaction terms added as necessary if testing complex hypotheses. This approach requires a rearticulation of what is meant by the construct of reactivity, but holds promise for teasing apart thorny issues such as that of differential predictive power of baseline versus challenge scores discussed in a non-psychophysiological context by Laird and Weems (2011). Further applied work using this analytic framework is needed.

In addition, latent variable models should be considered wherever feasible. These methods allow tests of key assumptions regarding psychophysiological research such as the stability of measurement structure over time as well as the relation between starting values and change. For example, researchers can estimate error-corrected associations between baseline values and reactivity across their sample or across subsamples in a multiple-group LDS or LRS analysis. As noted above, Table 2 provides examples of research contexts in which particular latent variable methods may be preferred. Defining reactivity as baseline-related (LDS), baseline-independent (LRS), or as a linear or quadratic change component (LGC) is central to this decision. And while some criticisms of the observed DS also apply to the LDS, the constraints imposed by the model can be tested empirically. Finally, although not yet seen in the applied physiological literature, the independent baseline and challenge model can also be implemented in latent form.

Whether using manifest or latent indices of physiological activity, researchers should aim for a rationale for their choice of including baseline variance directly in their change measure (e.g., DS) versus partialing out such variance (e.g., RS), or perhaps a rationale for testing both models. That is, despite the statistical equivalence of the overall models, LDS reactivity and LRS reactivity may be best viewed as different constructs. Through the whole process, specifying the predicted joint relationships between baseline, reactivity, and outcomes of interest is paramount, as is acknowledging that working with observed scores does not separate true change from measurement error.

**Reporting of results**

Journal reviewers and editors should encourage researchers to provide more detailed information than is typical to date on the descriptive and psychometric properties of their measurements. This information is crucial for several reasons, including allowing accurate meta-analyses and providing basic numeric information that might be leveraged by future statistical and methodological advances.

We recommend that researchers report the means, variances, and skewness of their baseline and challenge assessments, as well as the correlation between baseline and challenge scores and the presence and treatment of any outliers. By including the information above, the potential is increased for detailed comparison of core descriptive information across studies, measures, and tasks. At the same time, researchers should be careful to note the major differences in expected descriptive statistics across different physiological systems. Although not always necessary to report, the ratio of baseline score variance to challenge score variance (lambda) allows for a common metric for interpreting changes in variance over time and provides a convenient shorthand for indexing this change. In addition, as noted above, information on the psychometric reliability of physiological assessments should be reported wherever possible. We stress that baseline reliability may be easier to obtain than challenge reliability, since it is not always clear whether multiple laboratory challenges truly index the same construct (Stemmler, 2003). Reliability assessment can include test–retest measurement, though the temporal stability of psychophysiological measurement has also been conceptualized as an individual differences variable of intrinsic interest (Manuck et al., 1990). Finally, if manifest indices have been used in the analysis, it seems important to include discussion of the possible influence of measurement error on reported findings. Conversely, if latent variable models are used, as we strongly recommend, comprehensive guidelines for reporting the details of SEM results have been published elsewhere (e.g., McDonald &
Ho, 2002) and have proven invaluable for critical evaluation of such models. Researchers should aim to make explicit their hypothesized model for change in all contexts (McArdle, 2009; Stemmler, 1987).

Summary

Building on several decades of psychometric work, the comparison of different methods of measuring change remains an active area of investigation. Almost 50 years ago, Heath and Oken (1965) reviewed some of the issues considered in this paper, comparing the observed DS and the RS against each other and other methods of scoring change. There have been important advances in the intervening years, the development of latent variable methods being particularly noteworthy. Although choice of reactivity operationalization is dependent on the individual researcher, based on the nature of their data, experimental situation, and specific research question, we hope that the preceding discussion and Tables 1 and 2 will provide valuable input to this decision.

Ultimately, we see the statistical methods reviewed in this paper as mechanisms for testing developmental theory. For optimal theory-testing and theory-refinement, it is incumbent on theorists to specify in what ways they expect constructs such as the psychophysiological reactivity of a given system to change over developmental time. Armed with that information, the applied researcher can more easily choose one or more methods of data reduction and analysis that are appropriate for their particular research questions. We hope that by paying greater attention to the methodological and statistical issues reviewed here, applied developmental researchers can advance our general knowledge of psychophysiological processes and increase our understanding of how children’s physiological reactivity relates to other key aspects of development.

Acknowledgments

Jelena Obradović acknowledges the Canadian Institute for Advanced Research (CIFAR), the Jacobs Foundation, and the William T. Grant Foundation for supporting her research. We thank Glenn Roisman for his comments on an earlier draft of this paper.

Appendix

Annotated Mplus input syntax for LDS model

TITLE: Latent Difference Score Model, Figure 2
DATA: FILE IS path-to-data-file.dat;
TYPE IS individual;
VARIABLE: NAMES ARE X1 X2 X3 Y1 Y2 Y3;
MODEL:
! measurement model, including invariance and correlated residuals
X BY X1*(1);
X BY X2(2);
X BY X3(3);
Y BY Y1*(1);
Y BY Y2(2);
Y BY Y3(3);
X1 WITH Y1;
X2 WITH Y2;
X3 WITH Y3;
Annotated Mplus input syntax for LRS model

TITLE: Latent Residual Score Model, Figure 3
DATA: FILE IS path-to-data-file.dat;
TYPE IS individual;
VARIABLE: NAMES ARE X1 X2 X3 Y1 Y2 Y3;
MODEL:
! measurement model, including invariance and correlated residuals
X BY X1*(1);
X BY X2(2);
X BY X3(3);
Y BY Y1*(1);
Y BY Y2(2);
Y BY Y3(3);
X1 WITH Y1;
X2 WITH Y2;
X3 WITH Y3;
[X1 Y1](4); ! invariance of intercepts (here) and loadings (above)
[X2 Y2](5);
[X3 Y3](6);
X@1; ! sets scale of latent baseline factor (Z-score metric)
! core features of LRS model below (some echo Mplus defaults)
OUTPUT:
STAND; CINT; TECH1;
SAMPSTAT; ! optional output, see Mplus manual for details
(continued on next page)
Annotated Mplus input syntax for growth model

TITLE: Growth Model, Figure 4
DATA: FILE IS path-to-data-file.dat;
TYPE IS individual;
VARIABLE: NAMES ARE X1 X2 X3 Y1 Y2 Y3;
MODEL:
! measurement model constrains factor loadings to create growth factors
BASELINE BY X1@1;
BASELINE BY X2@1;
BASELINE BY X3@1;
BASELINE BY Y1@1;
BASELINE BY Y1@1;
REACTIVITY BY X1@0; ! mimics Mplus default (zero), included for explication
REACTIVITY BY X2@0;
REACTIVITY BY X3@0;
REACTIVITY BY X4@0.5;
REACTIVITY BY X5@1; ! can set nonzero loadings based on time intervals of assessment

! factor variances and covariance will be estimated by default
! but we need to estimate factor means and constrain indicator intercepts
(BASELINE);
(REACTIVITY);
[X1-Y2@0];

! constrain error variances equal for identification (can test in simpler models)
X1(1);
X2(1);
X3(1);
Condensed MODEL command (identical to above)

```plaintext
MODEL: i s | x1@0 x2@0 x3@0 y1@0.5 y2@1;
X1(1)
X2(1)
X3(1)
Y1(1)
Y2(1);
```

Annotated Mplus syntax for LST model

```plaintext
TITLE: LST model, Figure 5
DATA: FILE IS path-to-data-file.dat;
TYPE IS individual;
VARIABLE: NAMES ARE t1r1 t1r2 t2r1 t2r2 t3r1 t3r2;
MODEL:
  ! measurement model
  TIME1 BY t1r1(1); ! equal measurement precision of replicates assumed
  TIME1 BY t1r2(1);
  TIME2 BY t2r1(2);
  TIME2 BY t2r2(2);
  TIME3 BY t3r1(3);
  TIME3 BY t3r2(3);
  TRAIT BY TIME1 TIME2 TIME3;
  S1 BY TIME1;
  S2 BY TIME2;
  S3 BY TIME3;
  TRAIT WITH S1-S3@0; ! constrain unwanted paths/residual variances to zero
  S1 WITH S2-S3@0;
  S2 WITH S3@0;
  TIME1-TIME3@0;
OUTPUT:
  STAND; CINT; TECH1; ! optional output, see Mplus manual for details
  SAMPSTAT;
```
References


