Exceptional Children
2019, Vol. 85(2) 248–264

DOI: 10.1177/0014402918802803
journals.sagepub.com/home/ecx
SAGE

Aptitude-by-Treatment Interactions in Research on Educational Interventions

Kristopher J. Preacher[1] and Sonya K. Sterba[1]

Abstract

A common theme uniting articles in this special issue is a focus on aptitude-by-treatment interactions (ATIs). This timely and welcome focus allows the field to synthesize current substantive findings on ATIs in educational intervention research in both reading and math domains. In this methodological commentary, we begin by reviewing traditional approaches for detecting and reporting interactions in single-level and multilevel models. Next, we discuss some limitations of traditional approaches for theorizing about and modeling ATIs, and we suggest some solutions. These solutions include interpreting level-specific (unconflated) ATIs, understanding and ameliorating threats to adequate power for detecting ATIs, expanding focus beyond linear ATIs, and increasing the number of measurement occasions beyond two to allow use of a growth modeling framework for investigating ATIs. Incorporating some of these advances into future research can motivate new research questions about educational interventions and lead to new discoveries in the search for ATIs.

In this special issue of *Exceptional Children* are assembled seven reports of randomized controlled trials (RCTs) of validated instructional interventions that were designed to improve student learning in early education—five in the context of reading (Clemens et al.; Coyne et al.; D. Fuchs et al.; Vaughn et al.; Wanzek et al.) and two in the context of mathematics (Clarke et al.; L. Fuchs et al.).[1] In each study, learners, classrooms, or small groups were randomized to treatment or control conditions.

As noted in the introduction to this special issue, "Moderation analysis can be an important means by which interventionists better understand the nature and effects of their interventions." Indeed, a timely and welcome running theme uniting these articles is a shared focus on differential gains in response to intervention that may be conditional on preintervention aptitudes—in other words, *aptitude-by-treatment interactions* (ATIs)—manifested in different ways across studies. An *aptitude* is any characteristic of a person that forecasts his or her probability of success under a given treatment (Cronbach & Snow, 1977). All seven articles in this special issue deal with ATIs.

We appreciate being asked to provide a commentary on the statistical methodology employed in these studies. We use this commentary to suggest some methods that have the potential to advance the field of special education in the future. Because special education is such a high-stakes issue, it is critical that the best statistical methods be employed to complement rigorous experimental design so that new interventions may be assessed and customized to suit the learning needs of individual learners. Because of the inherent hierarchical structure of most educational interventions,

[1]Vanderbilt University

Corresponding Author:
Kristopher J. Preacher, Vanderbilt University, PMB 552, 230 Appleton Place, Nashville, TN 37203-5721, USA.
E-mail: kris.preacher@vanderbilt.edu

here we focus specifically on methodological issues surrounding the assessment of ATIs in multilevel data. Our commentary is organized into three major sections: (a) an overview of traditional methods for investigating ATIs, (b) an overview of modeling ATIs in multilevel contexts, and (c) a discussion of limitations of traditional approaches for theorizing about and modeling ATIs together with suggested solutions. These solutions include splitting interaction effects into level-specific components, understanding and minimizing threats to adequate power for detecting ATIs, expanding focus beyond linear ATIs to consider the nonlinear aptitude–achievement relationship to be moderated by intervention, and a reconceptualization of ATIs in terms of growth in achievement, with intervention and aptitude as moderators of aspects of change. Incorporating some of these best practices into future research can motivate new research questions about educational interventions and lead to new discoveries in the search for ATIs. Throughout, we make reference to the studies reported in this special issue.

Introduction to Investigating ATIs

Although research on educational interventions aims to shape future instruction to maximize learning, specific goals for accomplishing this vary by study. One goal may be for all learners to achieve a common level of mastery (minimization of individual differences in achievement). Another goal may be for all learners to maximize their achievement by capitalizing on individual aptitudes, which can lead to larger postintervention ability gaps. A third alternative goal may be to enhance the quality of instruction for all learners while deliberately minimizing the achievement gaps that exist among those who are differentially advantaged by differences in aptitudes or socioeconomic circumstances. At some level, all of these goals require research to inform the creation, implementation, and assessment of adaptive instruction tailored to the needs of individual learners.

The first step toward crafting instructional methods that can be tailored to different kinds of learners is to identify the preintervention characteristics that distinguish different types of learners. One common approach is first to rely on theory and cumulative research to create targeted instructional strategies, assess learners' pretreatment aptitudes on the dimension of interest, and then use them to predict postintervention abilities in both the treated and untreated groups. Typically, but not always, the pretreatment aptitude measure is the same as the posttreatment ability measure, administered sometime prior to treatment. An ATI (Cronbach & Snow, 1977) is evident if the interaction effect differs significantly from zero.

A standard tool kit of methods has evolved for specifying, testing, and (if significant) communicating interaction effects. The first step is to specify a model containing a parameter that reflects the presence and magnitude of the interaction, usually a linear regression model. The simplest model takes the form

$$\begin{aligned} y_i &= b_0 + b_1 x_i + b_2 z_i + b_3 x_i z_i + e_i \\ e_i &\sim N(0, \sigma^2), \end{aligned} \tag{1}$$

where y_i is the achievement outcome for individual *i,* b_0 is the intercept, b_1 is the conditional linear effect of the *focal predictor* x on y where the *moderator* $z = 0$, and b_2 is the conditional linear effect of the moderator z where $x = 0$. The interaction effect, b_3, is the amount by which the slope of x is expected to change for a one-unit increase in z (or, symmetrically, the amount by which the slope of z is expected to change for a one-unit increase in x). The points $x = 0$ and $z = 0$ need not be meaningful values. The model is invariant to linear rescalings of x, z, or both; that is, the model-implied values of y and significance of the interaction will not change with such rescalings. In ATI research, aptitude is often treated as a moderator of the intervention effect. Mathematically, it is arbitrary which predictor is treated as the focal predictor and which is treated as the moderator. For consistency, we treat aptitude as the focal predictor (x) and intervention as the moderator (z).

The second step is to fit the model to data, typically by regression analysis. The goal of model fitting is to estimate the regression coefficients, including the key parameter b_3. In the third step, b_3 is tested for significance. If significant, the interaction is termed a bilinear interaction because Equation (1) may be rearranged to show that the intercept and the slope of x are linear functions of z:

$$y_i = (b_0 + b_2 z_i) + (b_1 + b_3 z_i) x_i + e_i. \quad (2)$$

Here, $(b_1 + b_3 z_i)$ is the simple slope of x where z_i is any interesting fixed value of the moderator.

If b_3 is significant, the fourth step is to graphically illustrate the interaction to facilitate interpretation. There are two common strategies. The first is to choose a sequence of benchmark values within the observed range of z and plot the simple regression of y on x at those values. This conditional value method (Aiken & West, 1991; Dearing & Hamilton, 2006) yields a graphical illustration of the predicted values of y as a function of x for hypothetical cases at different points along z, which may be continuous or categorical. If the moderator is categorical with $k \geqslant 3$ groups, then z is replaced with $k - 1$ dummy codes, and the simple regression of y on x is plotted at conditional values of these dummy codes. In ATI research, aptitude is often placed on the horizontal axis, and the conditional regression of y on x is plotted for each treatment condition ($z = 0$ or 1) so that the treatment effect can be discerned as the vertical distance between the lines at any chosen value of x (as in Coyne et al., D. Fuchs et al., and L. Fuchs et al.).

Bilinear interactions can assume any of several distinct patterns. Figure 1 contains exemplar conditional value plots of some of these basic forms, with labels that are sometimes attached to each pattern (e.g., Cronbach & Snow, 1977; McCabe, Kim, & King, 2018). The horizontal axis in each plot in Figure 1 spans the range of the focal predictor x, and the vertical axis represents the outcome y. Each line on a given plot is the regression equation relating y to x at conditional values of the moderator z; these may be any reasonable, interesting values within the observed range of z. In the ATI context, z is the absence or presence of the educational intervention.

The second, and less common, graphical strategy involves plotting the $x \rightarrow y$ simple slope as a function of z in cases where z is continuous. In these marginal-effects plots (Dearing & Hamilton, 2006; McCabe et al., 2018), the plot relating the simple slope to z is depicted with (often) 95% confidence bands, continuously plotted 95% confidence intervals (CIs), around the estimated simple slope across the range of z. The values of z for which the confidence bands exclude zero constitute the region of significance (Preacher, Curran, & Bauer, 2006). Determining this region algebraically is known as the Johnson-Neyman technique (Johnson & Neyman, 1936). In ATI research, because aptitude is continuous and treatment is categorical, the roles of x and z are switched; that is, the horizontal axis represents aptitude and the treatment effect is plotted as a function of aptitude. Clarke et al. and Coyne et al. use this graphical method, plotting the treatment effect as a continuous linear function of aptitude. Clemens et al. report regions of significance but without the optional plots.

Among the special issue articles, L. Fuchs et al. found a main effect of combined treatment conditions but no interaction, implying that the treatments worked well across the observed range of aptitude. Clarke et al. and D. Fuchs et al. found a kind of compensatory interaction, in which most learners benefited from instruction of any sort, but those with relatively low pretest scores especially so. Coyne et al. describe a synergistic interaction, in which learners with relatively larger vocabularies were able to benefit more from an intervention than were those with smaller vocabularies, creating an even larger gap between low- and high-aptitude learners.

Traditional Methods for Investigating ATIs With Multilevel Data

In most ATI research, groups rather than individual learners are randomized to treatments.

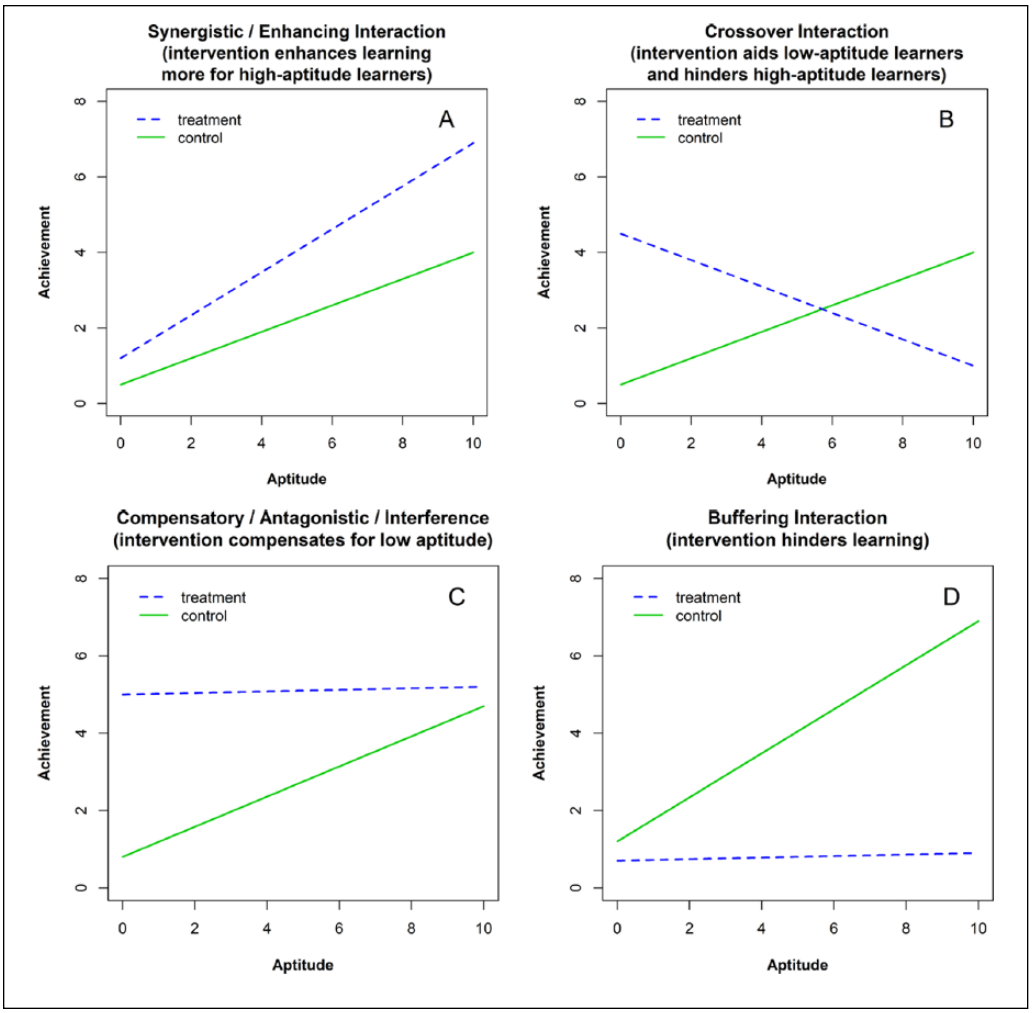

Figure 1. Four prototypical interaction effects. The crossover interaction is an example of a disordinal interaction (the lines cross, showing different treatment effect directions for learners with different aptitudes). The other three are examples of ordinal interactions.

This special case of the RCT is known as a cluster randomized trial (CRT). When faced with multilevel (or clustered) data, there are two main options: standard-error adjustment and multilevel modeling (MLM). Standard-error adjustment is often used when cluster-induced dependency in the data is not of substantive interest, perhaps because it arises as an artifact of data collection but is absent from the population of inference. For example, if data are collected from artificially constructed groups, yet hypotheses refer to a population without such groups, then standard-error adjustment might be appropriate to obtain unbiased estimates of sampling variability for key parameter estimates. However, in most educational research, data are inherently clustered because learners are taught in groups by the same instructors for extended time periods. This situation calls for MLM, also known as random coefficient modeling or hierarchical linear modeling (for an excellent introduction, see Hox, Moerbeek, and van de Schoot (2018).

Equation (1) can be modified to account for clustering by using a multilevel model, such as

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}z_j + \gamma_{11}x_{ij}z_j + u_{0j} + u_{1j}x_{ij} + e_{ij} \quad (3)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}\right), \quad e_{ij} \sim N(0, \sigma^2). \quad (4)$$

Here subscript *i* still references the individual learner and *j* indexes cluster, which may represent school, classroom, or small group, depending on the data. This model permits the intercept and the slope for *x* to vary randomly across clusters. When (as in all the special issue articles) there is no reason to suspect that the aptitude–achievement relationship varies across clusters, the constraints $\tau_{11} = \tau_{10} = 0$ are applied and Equations (3) and (4) reduce to

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}z_j + \gamma_{11}x_{ij}z_j + u_{0j} + e_{ij} \quad (5)$$

$$u_{0j} \sim N\left(0, \tau_{00}\right), \quad e_{ij} \sim N(0, \sigma^2) \quad (6)$$

As in Equations (5) and (6), it is good practice to include the fitted MLM equations in ATI research reports, as in D. Fuchs et al., L. Fuchs et al., Coyne et al., Vaughn et al., and Wanzek et al. Doing so removes ambiguity about the fitted model, facilitates interpretation of model parameters, and enables easier replication of the analyses.

This model in Equations (5) and (6) can be extended in numerous ways. A third level can be accommodated, and any number of predictors—including moderators—may be added at any level. For instance, it is possible to include more than one aptitude measure (as in Vaughn et al.). If the sample is sufficiently large and theory sufficiently rich, the researcher may consider higher-order interactions among several aptitudes (as in Vaughn et al.) or among several aptitudes and treatments.

With the exception of Clemens et al., who used standard-error adjustment, and Vaughn et al., who used a single-level model due to negligible clustering, the special issue articles each used a version of the MLM in Equations (5) and (6). Clarke et al. fit a partially nested MLM (for review, see Sterba, 2017) to accommodate a data structure in which learners were clustered in groups of two or five in the intervention conditions but were ungrouped in the control condition. All learners were further clustered in classrooms. Coyne et al. fit a four-level MLM (learner within subcluster within classroom within school), with intervention as a Level 2 predictor. D. Fuchs et al. report a three-level MLM, with learners within classrooms within schools. In contrast to Equation (5), they considered treatment a Level 1 moderator because the intervention was administered in the form of one-on-one tutoring and different learners in the same classroom could be randomized to one of three conditions. L. Fuchs et al. also used a three-level MLM, in which teachers and classrooms were cross-classified at Level 2 within schools. Finally, Wanzek fit two-level MLMs, with students nested in classrooms.

Typically, methods for evaluating and communicating interactions from single-level analysis are adopted in MLM. That is, coefficient γ_{11} from Equation (5) is tested, and if it is both statistically and practically significant, the interaction is probed and plotted using methods described earlier. In the next section, we highlight further improvements and refinement of this method for multilevel settings.

Limitations of Traditional ATI Research and Best-Practice Solutions

In this section we discuss several limitations of traditional ATI research and we provide best-practice solutions for each.

Limitation: Conflating Interaction Effects Across Levels

Consider the common MLM without an interaction where, for instance, clusters are classrooms:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + e_{ij} \quad (7)$$

$$u_{0j} \sim N\left(0, \tau_{00}\right), \qquad e_{ij} \sim N(0, \sigma^2). \quad (8)$$

In such models, the predictor x_{ij} can be decomposed into uncorrelated within- and between-class components by subtracting the class average $x_{.j}$ from x_{ij} and then using $x_{.j}$ as a separate Level 2 predictor. This yields a model with both within and between effects of x_{ij} (respectively, γ_{10} and γ_{01}):

$$y_{ij} = \gamma_{00} + \gamma_{10}(x_{ij} - x_{.j}) + \gamma_{01} x_{.j} + u_{0j} + e_{ij}. \quad (9)$$

Here, γ_{10} reflects the relationship between the outcome and one's standing on x_{ij} relative to the class mean, and γ_{01} reflects the relationship between the outcome and class averages of the predictor. It can be shown that γ_{10} from Equation 7 is a weighted average of γ_{10} and γ_{01} from Equation (9) and hence may not be interpretable. Many methodologists have echoed this problem (e.g., Burstein, 1980; Kreft, de Leeuw, & Aiken, 1995; Lüdtke et al., 2008; Preacher, Zyphur, & Zhang, 2010), yet applications of MLM in the ATI literature that separate effects of Level 1 predictors remain rare.

The consequences of failing to estimate separate level-specific effects can be serious. If the true within- and between-class effects of aptitude on achievement are, respectively, $\gamma_{10} = .7$ and $\gamma_{01} = .4$, their weighted average of (say) .58 characterizes neither students nor classes. Conversely, a nonsignificant overall slope near zero may mask a significant positive within slope and a significant negative between slope, a doubly tragic loss because two important effects will have been missed. It might be argued that if the level-specific effects are similar, then there is little harm in constraining them to equality (yielding Equation [7]), but we are still left with the problem that the units of analysis for the two conflated effects differ. Hence, it is unclear how to interpret their weighted average.

The problem of conflated effects is even more pernicious in the context of interactions. As Cronbach and Snow (1977) note, "Regressions and interactions divide—at least in principle—into group and individual components that have distinct substantive meanings. Recognition of these distinctions forces a radical change in thinking about ATI" (p. 100). Anderson (1941) identified a significant ATI in a study on arithmetic instruction, showing that drill instruction worked best for learners with high aptitude, whereas instruction emphasizing meaning worked best for learners with low aptitude. Cronbach and Webb (1975) famously reanalyzed Anderson's original data, showing that the within-classroom ATI (the interaction of instructional method with the within-classroom component of aptitude) was nonsignificant, and there were too few classrooms to obtain a reliable estimate of the between-classroom ATI (the interaction of instructional method with the classroom-average aptitude). In other words, the single ATI reported by Anderson likely was illusory. Many methodologists have since recommended that interaction effects should be split into level-specific components, not only for examining ATIs but for any interaction effect involving at least one Level 1 predictor or moderator (e.g., Aguinis, Gottfredson, & Culpepper, 2013; Enders & Tofighi, 2007; Kreft et al., 1995; Preacher, Zhang, & Zyphur, 2016).

Solution: Splitting Interaction Effects Into Level-Specific Components

In ATI research where the treatment variable is typically administered to clusters, there are two ways to model level-specific interactions. The first method for modeling level-specific interactions involves centering aptitude (x) at its cluster mean and fitting the following MLM, where $x_{i.} = x_{ij} - x_{.j}$:

$$\begin{aligned} y_{ij} = {} & \gamma_{00} + \gamma_{01} x_{.j} + \gamma_{02} z_j + \gamma_{03} x_{.j} z_j + \\ & \gamma_{10} x_{i.} + \gamma_{11} x_{i.} z_j + u_{0j} + e_{ij} \end{aligned} \quad (10)$$

The slope of $x_{i.}$ is not treated as random here, but could be. In this model, γ_{03} quantifies the Level 2 interaction, such that the effect of the intervention on classroom-average achievement is conditional on classroom-average aptitude. Further, γ_{11} is a cross-level interaction interpretable (when casting treatment as the moderator) as the difference between

treatment and control groups in the within-classroom aptitude–achievement relationship. Coyne et al. implemented a more complex version of this strategy. They fit a four-level model, with the aptitude measure split into four level-specific components, yielding four level-specific effects plus a cross-level interaction of treatment (Level 2) with within-subcluster aptitude (Level 1).

The second method for modeling level-specific interactions involves using multilevel structural equation modeling (MSEM). The reason to consider MSEM is that the model in Equation (10) yields trustworthy estimates only to the extent that (a) there is sufficient cluster-level variance in aptitude, as reflected by x's intraclass correlation (ICC_x), and (b) clusters are sufficiently large; otherwise, the between effect of x will be biased toward the within effect. The use of MSEM largely eliminates this bias by replacing observed cluster means x_j with latent cluster means (Lüdtke et al., 2008); however, a cost can be reduced power for detecting the between portion of the interaction. Preacher et al. (2016) show how to estimate interaction effects—including cross-level interactions like most ATIs—using MSEM. See Wanzek et al. for an application of MSEM to investigate classroom-level moderation of a treatment effect.

In our view, it is typically worth estimating a few additional parameters in order to decompose main effects and interactions involving Level 1 predictors into within and between components. If effects are not separated, and an ATI is found, we cannot know whether the interaction is driven primarily by student- or cluster-level forces. Or, if an ATI is not found, low power could be the culprit, as we discuss in the next section. But it also could be that either the within or between ATI would be detectable had effects been separated, but they remain obscure when conflated.

If effects are not separated, and an ATI is found, we cannot know whether the interaction is driven primarily by student- or cluster-level forces.

Limitation: Threats to Statistical Power for Detecting ATIs

Substantiated claims of ATIs in practice were rare by the mid-1970s (Cronbach & Snow, 1977). They are still rare. Some of the reasons for their rarity may have nothing to do with statistical power. For instance, theoretical and mathematical constraints on the possible forms of the interaction can make some interaction hypotheses highly unlikely or impossible to support (Rogers, 2002). For example, most ATI research concerns ordinal interactions with one continuous predictor, aptitude (because it is rare to expect an intervention to enhance learning for those with aptitudes in one range but actually suppress learning for others). However, it is especially difficult to detect ordinal interactions when at least one predictor is continuous (Rogers, 2002).

Part of the apparent scarcity of ATIs very likely is due to low statistical power. Even in the best of circumstances, interaction effects are notoriously difficult to detect (Aguinis, 1995; Zedeck, 1971). A variety of reasons have been suggested for the depressed power seen in interaction research. Some reasons include error variance heterogeneity across groups (Alexander & DeShon, 1994), measurement error (Busemeyer & Jones, 1983; Mathieu, Aguinis, Culpepper, & Chen, 2012), information loss due to using coarse measurement (Russell & Bobko, 1992), and unbalanced sample sizes across groups (Stone-Romero, Alliger, & Aguinis, 1994). For instance, Mathieu et al. (2012) found that low sample size at either Level 1 or Level 2 can undermine power for tests of cross-level interactions. Often several of these issues coexist in a given study, compounding the problem.

Part of the apparent scarcity of ATIs very likely is due to low statistical power.

Another reason for low power to detect interactions is particularly common in ATI research: restriction of range (e.g., Smith &

Sechrest, 1991; Stone-Romero & Anderson, 1994). All but one of the studies in this special issue (Wanzek et al.) restricted the range of aptitude or pretest scores in some way by design. Coyne et al. restricted focus to learners below the 30th percentile and between the 37th and 67th percentiles on the Peabody Picture Vocabulary Test. Clemens et al. examined learners assigned to reading intervention classes based on poor prior test performance. D. Fuchs et al. chose learners based on teacher nominations of lowest-performing learners, followed by elimination of 40% of the remaining learners based on rank-ordered factor scores. Clarke et al. selected the lowest-performing learners based on a composite of two number-sense instruments. L. Fuchs selected learners with math factor scores below the 40th percentile. Vaughn et al. restricted attention to learners with a standard score below 85 on the Gates-MacGinitie Reading Comprehension subtest. A focus on learners with low initial aptitudes (e.g., learning-disabled learners) seems reasonable because they are often the targets of novel instructional interventions. However, restriction of range to one or two narrow bands within the full range of aptitude scores can seriously undermine power for detecting an ATI. Ironically, researchers intent on creating interventions targeted to low-performing learners may be sacrificing the statistical power they need to identify those populations.

An additional reason for persistent low power for detecting ATIs is that, commonly, ATIs are omitted from the a priori power analyses but are nonetheless later investigated in an exploratory analysis once main effects analyses have been conducted. Such tests of interactions are inevitably underpowered. When they are not detected, researchers will sometimes conclude that an intervention worked well (or did not work) for all learners, regardless of aptitude, but low power may instead be responsible.

Solution: Methods for Assessing and Increasing Power for Detecting ATIs

We echo Mathieu et al. (2012) in strongly recommending that researchers conduct an a priori power analysis, perhaps informed by a pilot study, that includes not only the anticipated main effects but also anticipated interactions. Not only is it important to choose a minimally acceptable sample size on the basis of power analysis, but it is also advisable to exceed it by as much as possible, for two main reasons. First, the minimum sample size necessary to detect a given effect size with (typically) a .80 probability does not, however, guarantee narrow CIs. Many methodologists have emphasized that obtaining a precise parameter estimate (a narrow CI) is probably more important than rejecting a point null hypothesis known a priori to be false (Kelley & Maxwell, 2003). Obtaining a usefully narrow CI typically requires a larger sample than does rejecting a null hypothesis. Second, power analyses are usually conducted under idealized circumstances conducive to higher power—normality, homoscedasticity, linearity, no attrition, and so on—which are never fully realized in practice.

Software options exist for conducting simulation-based power analyses for cross-level interactions, like ATIs. For instance, Mplus 8's (Muthén & Muthén, 1998–2018) customizable model simulation capabilities permit power analysis for a large variety of models and can incorporate attrition and assumption violation (Muthén & Muthén, 2002). Another example is MLPowSim (Browne, Lahi, & Parker, 2009), which can be run as a standalone program to generate R code or in conjunction with MLwiN. It handles a much smaller range of models than Mplus; however, it is free and can be used to conduct multiple power analyses in one run. A third example, ML Power Tool (Mathieu et al., 2012) can be run directly in R but, again, handles a much smaller range of models than Mplus. A suite of Excel-based tools (PowerUp!) is also available (Spybrook, Kelcey, & Dong, 2016).

Beyond conducting a priori power analyses that include ATIs, there are additional strategies that can be used to mitigate the low power problem. First, strategically including covariates that are correlated with the outcome but not with predictors will lower standard errors for interactions without altering effect sizes (Moerbeek & Teerenstra, 2016). With random assignment, any

preintervention covariate will tend to be uncorrelated with treatment, but it may be difficult to identify covariates that are correlated with achievement yet uncorrelated with aptitude. Second, taking repeated measurements of the outcome can greatly enhance power. We return to this topic in a later section. Third, using latent variables with multiple indicators (as in MSEM) can mitigate the effects of unreliability in measures of aptitude, achievement outcomes, and covariates. The gain in power obtained by removing the attenuating effects of unreliability can greatly outweigh the increase in sampling error incurred by estimating more parameters. Fourth, careful consideration of the chosen comparison standard is warranted due to its large influence on effect size, which in turn influences power. Choices for the comparison standard include (a) the complete absence of an instructional method, (b) the business-as-usual (BAU) scientifically informed instructional method already in place, or (c) a novel competing intervention. The effect size could appear lowest under (c) but highest under (a). However, we caution that (a) usually is not feasible, realistic, or informative. We refer readers to Moerbeek and Teerenstra (2016) for additional design-based strategies to enhance power.

In some situations, the goal or hope of the researcher is to *fail* to find evidence for an ATI. In this event, if ordinary hypothesis-testing procedures are followed, there is a perverse incentive to seek *lower* power because failure to reject the null hypothesis is seen as providing support for the researcher's predictions. If the theoretical hypothesis is that an interaction does not exist or is too small to be practically relevant, then an alternative approach to hypothesis testing is more appropriate: equivalence testing (Walker & Nowacki, 2010). First the researcher establishes a practical definition of a "not meaningfully large" ATI—a range of values for the interaction effect that would be considered too trivial to be of consequence (say, the values in the interval ± .12). The null hypothesis is that the interaction effect lies *outside* this interval. Rejection of the null hypothesis implies that the entire $100(1-\alpha)\%$ CI is inside the interval ± .12. With equivalence testing, because narrow CIs are consistent with the desired outcome, the researcher is once again rewarded for using large samples, reliable measurement instruments, and other factors that traditionally improve power.

Limitation: Considering Only Linear Interactions

All of the articles in this special issue explored bilinear interactions, in which the linear slope of intervention was modeled as a linear function of aptitude. Assume for simplicity there are only two intervention groups—treatment and control. The effect of intervention must necessarily be linear—it is binary, so its slope is interpreted as a mean difference in y. Linear moderation of this treatment effect is usually the default choice for a variety of reasons, including tradition, expedience, and unawareness of alternative options. However, the outcome, intercept, and slope in such a model need not be linear functions of the moderator(s). There are theoretical reasons for considering nonlinear ATIs. Indeed, we believe that linearity is the exception rather than the rule.

There are theoretical reasons for considering nonlinear ATIs. Indeed, we believe that linearity is the exception rather than the rule.

First, consider again the plot in Figure 1, Panel A, which represents a typical example of an ATI model. There are three variables involved: intervention (0,1), postintervention achievement (y), and preintervention aptitude (x). Aptitude assumes the role of the focal predictor, achievement is considered the dependent variable, and intervention is treated as the moderator. Traditionally, aptitude is used as the horizontal axis, and each intervention condition is represented by a conditional regression line. The entire plot is a snapshot in time with respect to

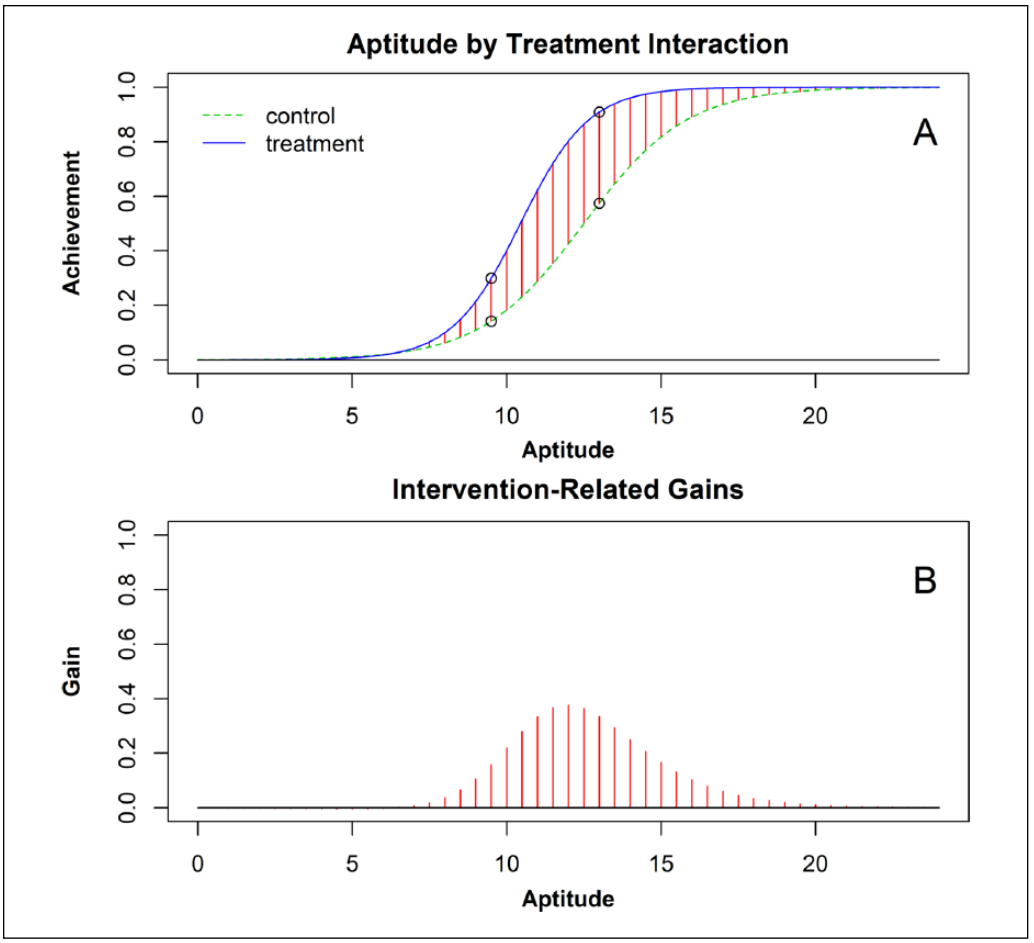

Figure 2. (A) A nonlinear Aptitude × Treatment interaction. (B) Intervention-related gain scores from Panel A.

achievement. Aptitude is assessed just prior to, or concurrent with, the initiation of the intervention. In this example, most learners are expected to benefit from treatment regardless of their aptitudes, but learners with higher aptitudes are expected to benefit more (a synergistic interaction, as in Coyne et al.). Linearity is evident in three aspects of this plot: (a) a clear overall main effect of z on y (the average vertical distance between the lines), (b) a clear overall main effect of x on y (the average slope of y regressed on x), and (c) a bilinear ATI. An ATI is evident in the form of different x slopes for $z = 0$ and $z = 1$. In this plot it is clear that learners with higher aptitudes benefit more from treatment, a common finding in practice.

Solution: Also Considering Nonlinear Interactions

The foregoing represents common practice in modeling ATIs. However, Smith and Sechrest (1991) note that when modeling ATIs, nonlinear models may be more appropriate. Also in the ATI context, Cronbach and Snow (1977) explained that "floor and ceiling effects inevitably place limits on the validity of a linear hypothesis. Data from the extremes of a scale often depart from a trend found in the middle range. Nor need causal relations be linear" (p. 31). For an example of how such floor and ceiling effects could occur, consider that learners with low aptitudes may not have sufficient foundational skills or knowledge to

benefit from treatment (a floor effect), but learners with high aptitudes cannot benefit from treatment because they easily master the material with or without the intervention (a ceiling effect). However, learners in the middle of the distribution may gain from treatment, but to a degree conditional on their aptitudes. Floor and ceiling effects can also occur when the measurement instrument does not measure past a certain level of mastery or below some minimal level of understanding. We agree with D. Fuchs et al. that many tests are not designed to distinguish well within extremely low or extremely high achievers, even when these ability ranges are quite wide.

Using Figure 1, Panel A, as a springboard, consider the plot in Figure 2, Panel A, which may be a more realistic depiction of the relationship between x and y if a sufficiently full range of x were to be examined. In this hypothetical example, the function linking x to y is logistic, one kind of intrinsically nonlinear function consistent with floor and ceiling effects like those suggested by Cronbach and Snow (1977). For instance, the model yielding the plot in Figure 2, Panel A, might be a nonlinear MLM—specifically a moderated logistic function:

$$y_{ij} = \frac{1}{1+\exp((\gamma_{00}+\gamma_{02}z_j+u_{0j})-(\gamma_{10}+\gamma_{11}z_j)x_{ij})} + e_{ij} \quad (11)$$

$$u_{0j} \sim N(0,\tau_{00}), \qquad e_{ij} \sim N(0,\sigma^2) \quad (12)$$

The *simple logistic intercept* ($\gamma_{00} + \gamma_{02}z_j + u_{0j}$) controls the horizontal position of the curve, with γ_{02} specifically controlling the horizontal distance between treatment and control curves. The *simple logistic slope* ($\gamma_{10} + \gamma_{11}z_j$) controls the rate at which the curve approaches the upper asymptote (mastery), with γ_{11} specifically controlling the treatment difference induced in this rate by intervention. Possibly only a segment of a curve like this describes the aptitude–achievement relationship in a given population. The error variance in such a model can also be modeled as heteroscedastic—for instance, if there is lower residual variability for learners who test extremely poorly or extremely well because of floor and ceiling effects (Davidian & Giltinan, 1995).

The standard practice of selecting learners on aptitude and fitting only a bilinear interaction model can be seen as fitting an overly simple model to a selected subset of the population. This can lead to models with reduced potential for generalizability. The four circled points in Figure 2 represent the synergistic bilinear interaction commonly seen when learners are selected based on, say, low and medium aptitude scores and a linear ATI model is fit. Had another range of aptitudes been selected, no interaction or even an antagonistic interaction might be observed. The *S*-shaped trends, and the distribution of intervention-related gains for learners at different points along the aptitude distribution, would remain obscured.

Employing nonlinear functions allows the goal of an educational intervention to be stated more precisely or even conceptualized entirely differently. Consider for example the following three ways that the goal of the intervention could be conceptualized using a moderated nonlinear model, none of which would be possible if a moderated linear model were fit. For example, one possible way to conceptualize an instructional intervention's goal would be to shift the curve describing the treated learners' aptitude–achievement relationship to the left relative to that of controls; this would in essence amount to an ATI because the vertical distance between the treatment and control curves changes in size across the range of aptitude (see Figure 2, Panel A, for illustration). A second way to conceptualize an intervention's goal is to make the curve for treated learners rise more steeply than that for controls (again see Figure 2, Panel A, for illustration). The vertical bars between the curves in Figure 2, Panel A, represent the gain due to intervention conditional on several selected aptitudes; these gains are plotted in Figure 2, Panel B, where the greatest gains from intervention are enjoyed by learners in the middle one third of the aptitude distribution. A third possible way to conceptualize an intervention's goal is to

raise the ceiling on learner understanding. Any of these criteria for success can be represented in nonlinear MLMs as free parameters or as random effects to be predicted.

It might be argued that models like that in Equation (11) are too complex to be practical. However, Equation (11) has just as many free parameters as the model in Equation (3)—perhaps one more if the error variance is modeled to be conditional on aptitude. In addition, several software options are available to help in fitting nonlinear models to individual learners' data (e.g., PROC NLIN in SAS or NLR in SPSS) as well as for many learners simultaneously, with or without nesting in higher-level units, like classrooms or schools (e.g., PROC NLMIXED in SAS, nlme and lme4 in R, or Mplus). Interactions can be plotted and probed using procedures that extend those used in linear regression.

Limitation: Two-Occasion Designs for Investigating ATIs

Many examples of ATI research (e.g., Clarke et al.; Coyne et al.; D. Fuchs et al.; L. Fuchs et al.; Wanzek et al.) use the same instrument to measure both aptitude and achievement. Regressing the latter on the former is an example of assessing *residualized change*, one of two classic methods for gauging change using pre-post data. The other such classic method is the analysis of difference scores, in which the benefit of receiving intervention is quantified by subtracting aptitude from achievement ($\Delta y = y_2 - y_1$) and treating Δy as the variable of interest. In traditional ATI research, the residualized change method is implemented using a two-occasion design.

Smith and Sechrest (1991), however, raise concerns about the use of two-occasion designs in this context—making the sobering point that some apparent ATIs may reflect not qualitative treatment differences in the outcome but, rather, differences in the rate of change as a function of learners' initial aptitudes. That is, all learners may eventually arrive at the same level of mastery as a result of the intervention, but at different speeds depending on their aptitudes. If the outcome is assessed prior to that point of eventual convergence, the pattern will resemble a synergistic ATI. Or the achievement outcome may be measured too early or too late to notice any treatment differences in learning as a function of aptitude at all. In general, the limitations of using only two occasions can be severe, as different learners' trajectories may follow distinct, nonlinear functional forms, as in Figure 3. Although linear growth may be sufficient to approximate change over short periods of time, it may be a gross oversimplification for trajectories of change over the time spanned by most educational interventions.

Solution: Conceptualizing ATIs as Aspects of Change Moderated by Intervention Using Multiple Repeated Measure Designs

To help guard against these possibilities, it is beneficial to assess learners at multiple occasions between the initial assessment of aptitudes and the final assessment of achievement and to extend the final assessment or follow-up period as much as is practical given constraints of the school year. Ideally each learner would be measured at least three times but preferably more (Smith & Sechrest, 1991; Willett, 1988–1989) to be able to model the learning trajectory for each individual. We advocate approaching the question of ATI from the perspective of longitudinal modeling—examining learning as a function of time and instructional intervention (Rogosa, 1991).

We advocate approaching the question of ATI from the perspective of longitudinal modeling—examining learning as a function of time and instructional intervention

Unlike the residualized change score method, the difference score method extends naturally to any number of occasions (Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1985). The residualized change model favored in classic ATI research considers between-subjects individual differences in achievement at a specific occasion (e.g., the end of the school

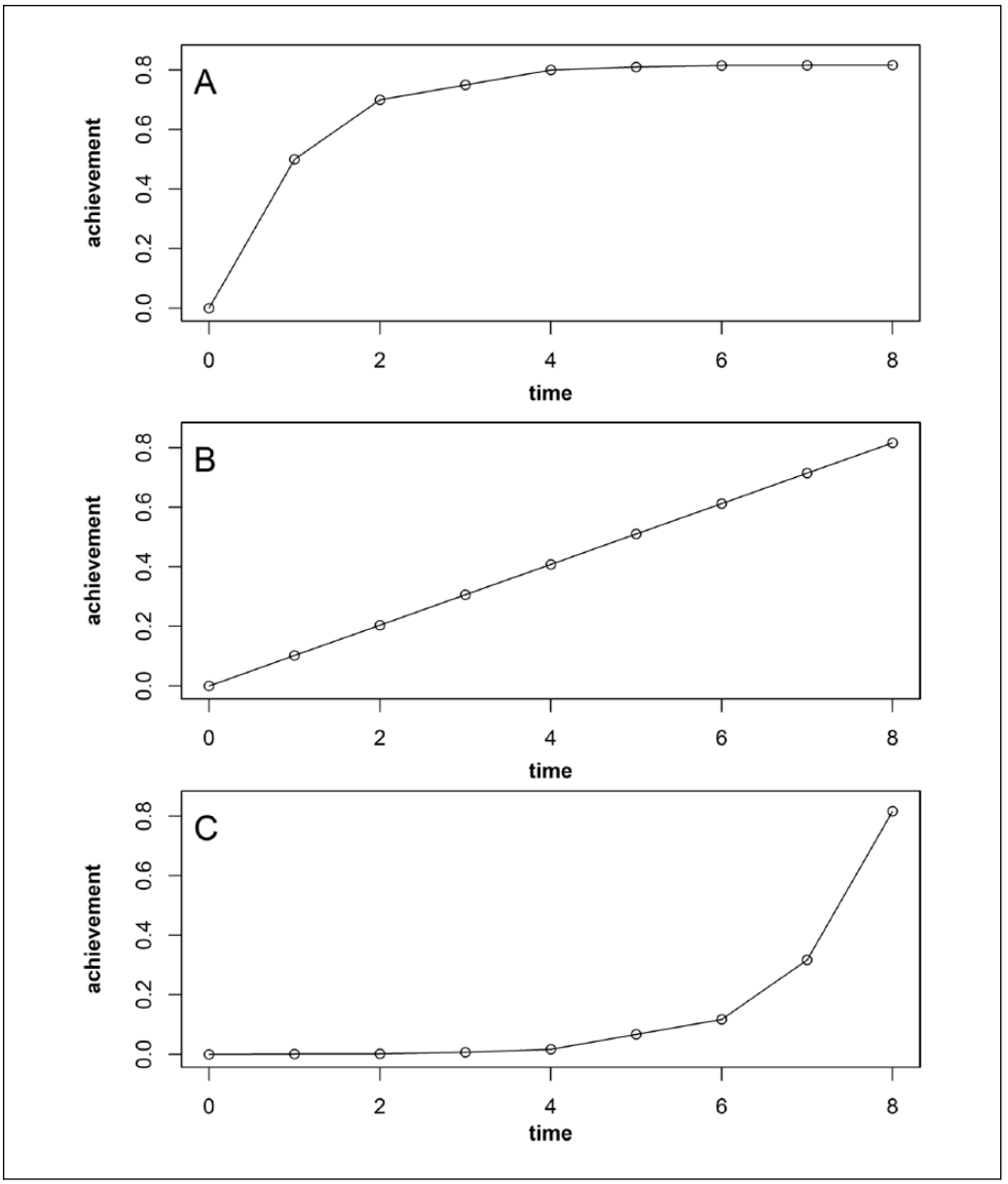

Figure 3. Achievement as a function of time for three hypothetical learners.

year). However, the use of difference scores and their logical extension—growth curve modeling (e.g., Bollen & Curran, 2006; Preacher, Wichman, MacCallum, & Briggs, 2008)—represents a shift from between- to within-subjects models of change, with all the usual benefits of within-subjects designs.

Given availability of multiple repeated measures, the next questions concern how to model change within a growth modeling framework, how to conceive of individual differences in aspects of change that might be amenable to intervention, and how to investigate an ATI in a growth model. First, it is sensible to determine an appropriate function relating achievement to time using an unconditional growth curve model. That is, a growth function needs to be identified that tracks

achievement over time for a typical learner from the population of interest. This function should fit the data well but should otherwise be as parsimonious as possible—that is, it should be a simple function with relatively few parameters.

Besides the traditional criteria of model fit and parsimony, another relevant dimension is the degree to which the growth function's parameters are substantively interpretable. For example, if an intervention targets cumulative learning rate (the speed with which the acquisition of, say, new math concepts depends upon previously acquired skills taught in the same intervention), it is crucial that the growth function include a rate parameter that can be conditioned on intervention status. If the intervention is designed to raise the ceiling on maximum vocabulary knowledge, the growth function applied to the data needs to have an upper asymptote parameter. Parameters that govern the behavior of growth functions, like asymptotes and rate parameters, are called *aspects of change*. In many cases relevant for ATI, an appropriate growth curve model would be a three-level nonlinear mixed model (occasion nested within learner nested within, say, classroom or some other grouping that is randomized to treatment). Time becomes a Level 1 predictor, person-level characteristics become Level 2 predictors of aspects of change, and classroom-level predictors (like intervention status) become Level 3 predictors of aspects of change and potential moderators of the effects of person characteristics.

Second, some thought should be given to how to incorporate individual differences in aspects of change. There are generally three ways to do this. The aspect of change can be fixed to a constant value if there is sufficiently strong theory to warrant such a constraint. For example, the lower and upper asymptotes of an *S*-shaped curve might be fixed to, say, 0.2 and 1, respectively, to span the range from chance levels of performance to complete mastery on a multiple-choice test. Or the aspect of change can be freely estimated, in which case the data help determine the best estimate for the parameter. Or the aspect of change can be treated as a random coefficient, allowed to vary normally across learners. If the aspect of change is treated as a free parameter or as a random coefficient, it can be moderated by treatment status to see how the intervention influences individual differences in aspects of change.

Third, how might aptitude be brought into the model? If aptitude and achievement are measured using different instruments, then aptitude can be included as a person-level predictor of any aspect of change. It is then possible for treatment status to moderate the extent to which aptitude predicts various aspects of change (an ATI). For example, perhaps aptitude is highly predictive of the rate of approach to mastery for BAU learners but not predictive of rate for those in the intervention. Or perhaps low-aptitude intervention learners absorb knowledge at a slower rate than their BAU counterparts but retain it better, setting them up for achieving a higher asymptote later, whereas high-aptitude learners are not as responsive to treatment but also tend to have high asymptotes.

If the same instrument is used to measure both aptitude and achievement, then aptitude is simply the initial measurement of achievement, at a point in time when no treatment effect is hypothesized because the intervention has not yet occurred. Theoretically it can still be used to predict aspects of change in individual learners' trajectories when initial status (the intercept or lower asymptote, or some other aspect of change that codes individual differences in aptitude just prior to intervention) is a random coefficient, capable of serving as a predictor of other aspects of change (Muthén & Curran, 1997).

Discussion

We applaud the authors in this special issue for bringing a timely and welcome focus to the influence of preintervention academic performance on the effectiveness of interventions—the classic yet elusive ATI. Taken together, articles in this special issue allow the field to synthesize current substantive findings on

ATIs in educational intervention research in both reading and math domains. In this commentary, our goal was to suggest ways to improve traditional ATI research, and we chose to highlight solutions to four limitations that have persisted across the past 40 years in the pursuit of ATIs in educational intervention research: (a) the separation of ATI effects into level-specific components; (b) understanding threats to, and improving statistical power for, detecting ATIs; (c) considering nonlinear aptitude–achievement models moderated by treatment; and (d) reconceptualizing ATIs using longitudinal time–achievement models moderated by treatment and aptitude. One can imagine an ideal educational intervention trial that combines elements of traditional ATI approaches with some of the suggestions we outlined here. Such a study would be preceded by an a priori power analysis and perhaps a pilot study to establish the likely functional form of learning across a wide range of aptitude scores. The model would involve separate estimation of within- and between-cluster effects and would not necessarily be limited to linear effects of aptitude and time. Although including all of these elements in a single study would likely be infeasible, it would nonetheless be profitable to consider some elements of what we have discussed in future studies and recognize the limitations inherent in traditional approaches. For example, fitting a linear model when the process under study is almost certainly nonlinear may be expedient but may destroy the model's fidelity to the process underlying the data. Multiple factors conspire to undermine statistical power; thoroughly understanding those factors can help researchers take steps to avoid them.

Greater progress can also be made by combining some of the strategies we suggest. For example, the logic of separating within and between effects still applies when nonlinear models are used. If one has repeated measures nested within students, who in turn are nested in small groups, it is plausible for learners within groups to follow an *S*-shaped curve as a function of time (at Level 1), and for the learners' average or terminal achievement (at Level 2) to follow a different function of aptitude, and for treatment to moderate parameters at either or both levels.

There are also many opportunities to go beyond the topics discussed here to advance the frontier of ATI research. For example, parametric models, whether linear or nonlinear, constitute only one method to investigate interactions, and their limitation is that they require specifying the functional form of the interaction in advance of fitting the model. When the functional form of the interaction is unknown, however, semiparametric latent class models can be used to identify subgroups with distinct aptitude–achievement or time–achievement relationships, with treatment condition predicting latent class membership (e.g., Sterba & Bauer, 2014), or local structural equation modeling (e.g., Hildebrandt, Wilhelm, & Robitzsch, 2009) can be used to identify varying patterns of effects across the range of a continuous moderator. These and other techniques bear consideration in future ATI studies.

References

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management*, *21*, 1141–1158. doi:10.1177/014920639502100607

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*, 1490–1528. doi:10.1177/0149206313478188

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

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, *115*, 308–314. doi:10.1037/0033-2909.115.2.308

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley. doi:10.1002/0471746096

Browne, W. J., Lahi, M., & Parker, R. M. A. (2009). *A guide to sample size calculations for random effect models via simulation and the MLPowSim software package*. Bristol, UK: University of Bristol.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, *8*, 158–233. doi:10.3102/0091732X008001158

Busemeyer, J. R., & Jones, L. E. (1983). Analyses of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *92*, 549–562. doi:10.1037/0033-2909.93.3.549

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on aptitude-treatment interactions*. New York, NY: Irvington.

Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported Aptitude × Treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology*, *67*, 717–724. doi:10.1037/0022-0663.67.6.717

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. Boca Raton, FL: CRC Press.

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.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138. doi:10.1037/1082-989X.12.2.121

Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology*, *16*, 87–102.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). New York, NY: Routledge.

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their applications to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*, 305–321. doi:10.1037/1082-989X.8.3.305

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effects of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1–21. doi:10.1207/s15327906mbr3001_1

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. doi:10.1037/a0012869

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, *97*, 951–966. doi:10.1037/a0028380

McCabe, C. J., Kim, D. S., & King, K. M. (2018). Improving present practices in the visual display of interactions. *Advances in Methods and Practices in Psychological Science*, *1*, 147–165. doi:10.1177/2515245917746792

Moerbeek, M., & Teerenstra, S. (2016). *Power analysis of trials with multilevel data*. Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371–402. doi:10.1037/1082-989X.2.4.371

Muthén, L. K., & Muthén, B. O. (1998–2018). *Mplus user's guide: Statistical analysis with latent variables* (8th ed.). Los Angeles, CA: Author.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620. doi:10.1207/S15328007SEM0904_8

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448. doi:10.3102/10769986031004437

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage. doi:10.4135/9781412984737

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, *21*, 189–205. doi:10.1037/met0000052

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209–233. doi:10.1037/a0020141

Rogers, W. M. (2002). Theoretical and mathematical constraints of interactive regression models. *Organizational Research Methods*, *5*, 212–230. doi:10.1177/1094428102005003002

Rogosa, D. R. (1991). A longitudinal approach to ATI research: Models for individual growth and models for individual differences in response to intervention. In R. E. Snow, & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 221–248). Hillsdale, NJ: Lawrence Erlbaum.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726–748. doi:10.1037/0033-2909.92.3.726

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203–228. doi:10.1007/BF02294247

Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology*, *77*, 336–342. doi:10.1037/0021-9010.77.3.336

Smith, B., & Sechrest, L. (1991). Treatment of Aptitude × Treatment interactions. *Journal of Consulting and Clinical Psychology*, *59*, 233–244. doi:10.1037/0022-006X.59.2.233

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two- and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics*, *41*, 605–627. doi:10.3102/1076998616655442

Sterba, S. K. (2017). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research*, *27*, 425–436. doi:10.1080/10503307.2015.1114688

Sterba, S. K., & Bauer, D. J. (2014). Predictions of individual change recovered with latent class or random coefficient growth models. *Structural Equation Modeling*, *21*, 342–360. doi:10.1080/10705511.2014.915189

Stone-Romero, E. F., Alliger, G. M., & Aguinis, H. (1994). Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management*, *20*, 167–178. doi:10.1177/014920639402000109

Stone-Romero, E. F., & Anderson, L. E. (1994). Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderating effects. *Journal of Applied Psychology*, *79*, 354–359. doi:10.1037/0021-9010.79.3.354

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*, 192–196. doi:10.1007/s11606-010-1513-8

Willett, J. B. (1988-1989). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422. doi:10.3102/0091732X015001345

Zedeck, S. (1971). Problems with the use of "moderator" variables. *Psychological Bulletin*, *76*, 295–310. doi:doi.org/10.1037/h0031543

Note

1. In the following, citations without dates are understood to refer to articles in this special issue.

Authors' Note

This research was supported in part by grant #R23413D0003 to Vanderbilt University from the National Center for Special Education Research in the Institute of Education Sciences in U.S. Department of Education). This content does not necessarily reflect positions or policies of the funding agency and no official endorsement by them should be inferred.

Manuscript received June 2018; accepted August 2018.