Journal of Pediatric Psychology Advance Access published November 26, 2013

Aggregated N-of-1 Randomized Controlled Trials: Modern Data Analytics Applied to a Clinically Valid Method of Intervention Effectiveness

Christopher C. Cushing,[1] PhD, Ryan W. Walters,[2,3] MS, and Lesa Hoffman,[3] PhD

[1]*Department of Psychology, Oklahoma State University,* [2]*Department of Medicine, Division of Clinical Research and Evaluative Sciences (CRES), Creighton University, and* [3]*Department of Psychology, University of Nebraska–Lincoln*

All correspondence concerning this article should be addressed to Christopher C. Cushing, Department of Psychology, Oklahoma State University, 116 North Murray, Stillwater, OK 74078, USA.
E-mail: christopher.cushing@okstate.edu

Received June 14, 2013; revisions received October 11, 2013; accepted October 21, 2013

Objective Aggregated N-of-1 randomized controlled trials (RCTs) combined with multilevel modeling represent a methodological advancement that may help bridge science and practice in pediatric psychology. The purpose of this article is to offer a primer for pediatric psychologists interested in conducting aggregated N-of-1 RCTs. **Methods** An overview of N-of-1 RCT methodology is provided and 2 simulated datasets are analyzed to demonstrate the clinical and research potential of the methodology. **Results** The simulated data example demonstrates the utility of aggregated N-of-1 RCTs for understanding the clinical impact of an intervention for a given individual and the modeling of covariates to explain why an intervention worked for one patient and not another. **Conclusions** Aggregated N-of-1 RCTs hold potential for improving the science and practice of pediatric psychology.

Key words methodology; multilevel modeling; data simulation.

Introduction

Techniques for producing behavior change in pediatric psychology include stimulus control (e.g., keeping pill bottles on the kitchen counter), operant conditioning (e.g., labeled praise for complying with treatment), and self-regulation (e.g., self-monitoring, goal setting, and feedback). Each of these interventions represents a discrete variable in a child's treatment that is believed to modify a behavior when present and should also produce a behavior change in the opposite direction when withdrawn (e.g., patients who stop self-monitoring are less likely to adhere to a regimen). Indeed, several meta-analytic studies in pediatric psychology highlight the importance of behavioral self-regulation strategies for improving health behavior (Cushing & Steele, 2010; Kahana, Drotar, & Frazier, 2008).

Moreover, behavioral intervention targets are likely to change more quickly in the context of treatment than would cognitions, and as such are a good fit for small-*n* research designs that capitalize on within person variability. Consider for a moment the aforementioned techniques necessary to encourage a child to take a pill tomorrow. These are markedly different than the techniques necessary to change the child's belief that taking a medication is not important (e.g., cognitive modification). For this reason, behavioral foci of treatment may deserve special attention in the intervention design and evaluation phases. By focusing on behavioral targets early in the design phase, it may be possible to streamline treatments by elucidating the strategies most likely to promote behavioral compliance when cognitive factors are ideal and include only treatment components empirically demonstrated to have an effect on the dependent variable (DV) of interest in the final treatment package.

Relatedly, the relatively rapid modifiability of behavioral targets in pediatric psychology interventions (e.g.,

doi:10.1093/jpepsy/jst083

pill taking, physical activity, diet, treatment adherence) is amenable to small-*n* research designs. Rigorous small-*n* research methodologies create potential for establishing the preliminary efficacy of behavioral intervention strategies or for pursuing an innovation within an established approach. Most small-*n* research in pediatric psychology relies on the behavioral research methodologies of the 1970s and 1980s (Rapoff & Stark, 2008). Such approaches can determine the effect of a given intervention, but may fail to provide meaningful information when an intervention is not uniformly effective across participants. Herein the question is no longer *does the intervention work* but rather *for whom did the intervention work and why* (Kazdin, 1997).

Advancements in data analytic methods that allow accurate modeling of nested sources of variability represent a tremendous opportunity for the convergence of clinical practice and rigorous research methods; they may also lead the field to an approach that more closely ties science to practice. The aims of the current article are to: (1) provide an overview of N-of-1 randomized controlled trial (RCT) methodology as a rigorous strategy for testing the efficacy of an intervention within a given individual, (2) present a clinically feasible method of analyzing the data yielded from multiple N-of-1 RCTs to answer the question *for whom does the intervention work?*, and (3) present statistical models for answering the question *why does the intervention work for some participants and not others*?

In the next section, we will discuss the methodological requirements of an N-of-1 RCT, its strengths, and some limitations. Following this discussion, we will present a flexible and powerful method of conceptualizing and modeling the information yielded by this methodology using multilevel modeling, complete with a simulated data example. Our aim with this presentation is to provide sufficient explanation, data, and syntax so that a reader experienced with multilevel modeling can practice conducting data analysis of N-of-1 RCTs using the practice case in the current article. For a reader less familiar with multilevel modeling, we present the current article as a primer to demonstrate the use of N-of-1 RCTs and a corresponding data analytic procedure that might inspire further analysis in collaboration with others.

Overview of N-of-1 RCTs and Methodological Considerations

N-of-1 RCTs differ from other randomized designs in that randomization occurs at the level of measurement occasions rather than participants (Keller, Guyatt, Roberts, Adachi, & Rosenbloom, 1988). Presumably, the most logical study interval within pediatric psychology would be to randomize at the level of study days. However, the length of measurement occasions can be set by the interventionist based on the hypothesized length of treatment required to detect an effect and could be as short as minutes or as long as is reasonable given the context.

We note that readers familiar with the behavioral tradition in pediatric psychology will be aware of the methodological similarity between the N-of-1 design and a reversal design (Rapoff & Stark, 2008). In fact, the reasons that one would choose an N-of-1 RCT over other small-*n* designs are the same as a reversal design: (1) it is safe to withdraw treatment, (2) information about within subject variability is of benefit to the research question, and (3) clear benefit/feasibility of running a similar trial with a large-*n* has not yet been established.

Methodological Considerations and Requirements for N-of-1 RCTs

Requirement of Randomization

As noted previously, a key feature of N-of-1 RCTs is the use of randomization. Random assignment of days to conditions allows causal inferences to be drawn about the efficacy of a treatment for a given individual. Here we recommend block randomization as is common in traditional RCT protocols (see Altman & Bland, 1999). For example, if a study period included 30 days, then 15 days would be randomized to intervention and 15 to the control condition. By using block randomization, the number of intervention days is held constant across individuals to ensure an equal ‘‘dose’’ of the procedure. For additional reading on this topic as well as a useful randomization tool, see the R package (SCRT-R) developed by Bulté and Onghena (2008). If randomization does not occur, it may introduce systematic variability into the design that cannot be explained statistically due to time-based confounds. Such threats to internal validity in N-of-1 RCTs are the same as those in larger RCTs and include history and maturation effects, as well as effects of particular treatment days (e.g., a child attends an exercise group on Saturday).

Block randomization is not the only option in the context of N-of-1 RCTs. Investigators may also choose Latin-square or counterbalanced designs (*see* Brooks, 2012 *for review*). For instance, a traditional counterbalanced design involves two groups with different treatment orders such as Group 1 order AB and Group 2 order BA. In the case of an N-of-1 design the ‘‘grouping’’ variable is within persons, such that an individual patient would receive two treatment sequences back-to-back (e.g., ABBA). This technique may be appropriate when practical limitations prohibit block randomization. In contrast, a Latin-square design may be more useful when a researcher needs to control for order effects across more than two conditions (Brooks, 2012). For example, a researcher may create a 3 × 3 matrix to represent two treatment conditions compared against one control condition. The first row and column of the 3 × 3 matrix would each be ordered ABC; the other cells would be ordered such that each condition appears once per row and once per column.

Requirement of Rapid Onset and Abrupt Reversibility

One of the first requirements for an N-of-1 RCT is that the treatment *must have an immediate and abrupt effect that will be reversed when withdrawn* (Guyatt et al., 1988). Here reversibility applies to the quantitative effect on the DV, but also implies that the treatment is being reversed/withdrawn. This requirement means that N-of-1 RCTs are most commonly used for pharmacological trials in medicine (Tsapas & Matthews, 2008); however, many behavioral interventions are also amenable to such techniques. For example, Sniehotta, Presseau, Hobbs, and Araújo-Soares (2012) hypothesized that access to a pedometer for self-monitoring would increase physical activity in a group of adult office workers. Participants were randomly assigned to review self-monitoring data on treatment days and were blind to the data on control days. Because exercise behavior occurs in bouts and the experimenters chose a 24-h period as their study interval, it was plausible that participants would return to baseline when the intervention was withdrawn.

At first blush it may seem as though pediatric psychology interventions are not well-suited to N-of-1 RCTs precisely because the goal of the field is to develop lasting changes in health behavior that are not reversible (*à la* the Society of Pediatric Psychology vision statement: Healthier children, youth, and families). Indeed, this might seem to be the case if one *only* considers entire treatment *packages* within pediatric psychology. Herein lies what we believe is the major contribution of N-of-1 RCTs in pediatric psychology. That is, they are not useful for evaluating a treatment *package* (in which case a full-scale RCTs would be needed); rather, they are most useful for testing whether a treatment *component* has an effect on a specific target behavior over a discrete time interval. The difference is that a treatment package confers a lasting effect (ideally) over a long period of time and cannot be withdrawn. In contrast, a child who puts their medication bottle on the kitchen counter on Monday can put it back in the cupboard on Tuesday (i.e., application and withdrawal of stimulus control) to demonstrate what effect this change has.

The adjacent field of health psychology is becoming more and more concerned with treatment *components* that lead to outcomes at the level of systematic review (Abraham & Michie, 2008). We believe that N-of-1 RCTs may give the field of pediatric psychology a tool for investigating *components* in the stream of treatment rather than waiting for a large and expensive body of literature to review after treatment *packages* have been shown to be effective for some but potentially ineffective for others.

In our view, behavioral self-regulation strategies (i.e., components) for adherence to medication or behavioral recommendations (e.g., diet, exercise, sleep) are the types of pediatric psychology interventions likely to meet the requirement of rapid onset and reversibility. However, such assumptions are in and of themselves testable. In a pilot phase for an N-of-1 RCT, an investigator could easily use a few participants in a traditional ABAB design to determine the reversibility of an intervention component. Similarly, if the initial pilot raised suspicions that a carryover effect might exist in the data, then the pilot phase could inform the specification of covariates to account for such variability.

Carryover Effects

Related to the concepts of rapid onset and reversibility are carryover effects. In some cases, the withdrawal of an intervention may result in reduction in the DV that is lower than the value observed in the intervention condition but higher than baseline. In this case, it may be that the intervention imparted a weak, but permanent, change to the DV. In the pill-taking example in the introduction, the child that takes the pill on Monday because the bottle is on the counter may be more likely to take a dose on the next control day than she was on the previous control day, but less likely than on the next intervention day. This is a positive finding from a clinical perspective, but it does pose a problem within the N-of-1 context. In general, the N-of-1 RCT is simplest if the intervention under investigation does not have carryover effects from the intervention to control condition.

Carryover effects can lead to both Type I and Type II errors if the effect is not included in the statistical model. The potential for carryover effects can be minimized by using a Latin-square or some other counterbalancing design described in the research methods literature (e.g., crossover design with appropriate washout period between treatments; see Shadish, Cook, & Campbell, 2002 for review). However, following data collection, whether or not a carryover effect occurred is a testable hypothesis by including whether *the day prior to a treatment day* was a treatment or control day as an additional predictor variable in the analysis. The procedure for testing carryover effects is presented for both simulated datasets.

Suitability of Outcome Measures

We expect that most interventionists interested in N-of-1 methodology will conduct studies with participants, spending a relatively short time in each condition (e.g., 1 day). As such, we do not recommend the use of measures that are likely to change slowly over time, such as quality of life, self-efficacy, or measures of psychological symptoms such as a narrow-band internalizing measure or broad-band screener. Instead, the ideal measure is an objective indicator that provides no feedback to the participant. Common examples in pediatric psychology include accelerometers for sleep and physical activity (e.g., Meltzer, Montgomery-Downs, Insana, & Walsh, 2012) or electronic measures of medication adherence (Ingerski, Hente, Modi, & Hommel, 2011). If information about a subjective state or self-reported symptoms is desired, then daily diary methods may be appropriate. For example, patients with chronic pain might be asked to rate their daily pain and to rate their mood.

Blinding to Condition

In the ideal RCT, both the participant and experimenter are blind to condition. This is consistent with all experimental work in which double blinding is best, single blinding is second best, and unblinded is the worst with regard to controlling the threats to internal validity. Although double blinding may be difficult in many interventions within pediatric psychology, at least single blinding may be possible in many instances. Creative methodologists can likely think of a way to incorporate single or double blinding into a trial, and should do so if the knowledge gains outweigh the risks to the participants. However, in cases where blinding is not possible, the researcher should be aware that failure to blind does open the study to a risk of bias and problems with internal validity.

Ethical Considerations

Because treatment will be withdrawn and restarted in the course of the N-of-1 RCT, it is critical that the experimenter guard against any possible harm that could come to the participant as a result. Perhaps of greater concern is whether the treatment can be reasonably withdrawn if it is suspected that the treatment had a significant effect for a serious health behavior. For instance, adherence to medication is critical in many pediatric conditions. For a patient highly sensitive to the effects of a medication, it may not be acceptable to withdraw an intervention if a return to baseline would represent a significant decrease in the participant's overall health. In this case, the experimenter may set an *a priori* criterion for determining that an effect of treatment has occurred. That is, even a marked difference between the first two conditions may constitute a treatment effect that can be agreed on by the patient and provider (Guyatt et al., 1988). However, if the trial is discontinued to allow the patient to continue the therapeutic effect, then the multilevel modeling procedures described in this manuscript will not be possible.

Data Analysis

Uncertainty about how to analyze data yielded by N-of-1 RCTs exists in the methodological literature. Bayesian statistics have been suggested as one method for using aggregated data to understand a single case (Schluter & Ware, 2005). However, this approach may be relatively new to most pediatric psychologists more familiar with linear regression, and may be insufficient if the investigator is interested in explaining sources of variability observed between participants. For these reasons, a regression-based technique such as multilevel modeling may be both more approachable and more useful within pediatric psychology. In the current primer, we use multilevel modeling as a method of aggregating and analyzing N-of-1 RCT data. We describe this approach in detail in the context of the case example. While discussions of power and sample size are beyond the scope of this article Chapter 11 of Snijders and Bosker (2011) and the Optimal Design freeware from Raudenbush, Spybrook, Congdon, Liu, and Martinez (2011) are useful resources for the interested reader.

Illustrative Case Example

For the purposes of illustration, imagine that a clinician delivering empirically supported family-based behavioral treatment for pediatric obesity is interested in using mobile health (mHealth) technology to enhance an evidence-based treatment protocol, such as the *Positively Fit* program (Steele et al., 2012). This program involves 10 weeks of group-based sessions, with content tailored to both parents and their child or adolescent. One of the program goals is to increase the amount of physical activity performed by children, adolescents, and their families. The protocol has previously been enhanced with mHealth technologies on a small scale (Cushing, Jensen, & Steele, 2011), suggesting that adolescents are willing to use these technologies in the context of their clinical care.

One weakness of *Positively Fit* and other weight-management protocols is the lack of intervention content delivered *in vivo* for individual families. In this hypothetical example, imagine that a provider is interested in delivering a brief standardized text messaging intervention to eight adolescents enrolled in their hospital-based iteration of the *Positively Fit* program (which would typically constitute a maximally full group for *Positively Fit*). A strength of an N-of-1 approach is its ability to yield clinically meaningful results *in the course of treatment* rather than requiring that the interventionist wait until termination to evaluate the effectiveness of the intervention. As a consequence, the provider can make additional adjustments or perhaps even conduct another N-of-1 RCT if the first trial is ineffective for a given patient. For example, in the text message adjunct to the *Positively Fit* program, a 30-day study period—in which the eight patients each participate on half of the study days—could yield a great deal of information about the utility of the intervention, and constitutes less than half of the overall treatment period.

Simulation Description

For the current example, we imagine a group of adolescent females enrolled in the *Positively Fit* program. Minutes spent in moderate to vigorous physical activity (MVPA) as measured by accelerometry will be the DV. For demonstration purposes, two simulated datasets were created to evaluate participants' physical activity for 30 days. One dataset includes eight simulated participants to demonstrate simple individual-level effects while a 30-person dataset was simulated to demonstrate multilevel interactions. Study days were block randomized within participants such that each participant received 15 intervention and 15 control days. Data were generated using a multilevel linear model; associated SAS and SPSS syntax for all analyses are provided in the online Supplementary Materials.

Fixed or Random Effects?

Individual differences in a treatment effect can be evaluated using fixed-effects models or random-effects models. Fixed-effects models are most useful to practitioners wanting to know *for whom did the intervention work*. In a fixed-effects model, the *N* individual participants are distinguished by creating $N-1$ dummy-coded variables that are included as main effects and in interactions with the treatment variable.

The primary limitation of using a fixed-effects model is that it is then impossible to explain *why* some participants respond better than others. By contrast, in a random effects model, the individual variability in the treatment effect is partitioned explicitly so that it can be explained by predictors at the level of the individual (e.g., self-efficacy, social support). Thus, a random-effects model is most useful to researchers who are interested not only in for whom did the intervention work but also *why it worked* for some persons but not others, thereby making the answers to both of Kazdin's questions (that have characterized most of the past 2 decades of large-sample multi-site research) available to clinicians with few participants. Here we are simply echoing the previous examples in the pediatric psychology literature (e.g., Nelson, Aylward, & Rausch, 2011), which demonstrate convincingly that answering complex research questions may not *always* require large samples.

Study I: Evaluating for Whom the Intervention Worked

Examining the Treatment Effect for Each Participant

The simulated data for Study I involve eight participants measured across 30 days, in which 15 treatment days and 15 control days are randomly assigned within each participant. This treatment effect is a binary variable in which 0 indicated a control day and 1 indicated a treatment day. The response to the intervention within each of our eight participants was examined using a fixed-effects model. Accordingly, in creating $N-1$ dummy-coded variables to represent the *N* participants, one participant must serve as the reference (i.e., as the model intercept). We arbitrarily chose participant 8, such that the dummy-coded variables then represent the difference between participant 8 and each other participant. In actual clinical care, the choice of a reference participant should be tied to program goals. For instance, a clinician could choose the patient that was closest to meeting a program goal (e.g., 60 min of physical activity). Alternatively, the reference participant could be the patient who performed the most or least of a given variable. In any case, all other patient's values will be relative to the reference participant, but it is important to note that the choice of reference participant will not change model predictions or model fit.

The overall differences across participants in fixed effects can be evaluated with multiple degree-of-freedom F-tests, as routinely provided by many software programs. For our simulated data, there were significant differences across participants in their average minutes of physical activity across days, $F(7{,}224)=25.95$, $p<.05$, as well as in their intervention effects (i.e., differences between control and treatment days), $F(7{,}224)=2.92$, $p<.05$. As an effect size estimate, the intervention effect explained 25.62% of the variability in minutes of physical activity.

The regression effects—as given by default in any statistical package—are presented in the first set of columns of Table I. Given the dummy coding for participant, the intercept estimate of 54.80 represents the average minutes of physical activity *specifically for participant 8 on control days*. The intervention main effect indicates that *participant 8* had a nonsignificant 9.60-min increase in physical

Table I. *Final Results Using Fixed-Effects Model With Constant (Homogeneous) Variance*

	Differences relative to participant 8				Unique effects for each participant			
Fixed effects	Estimate[a]	*SE*	*t*	*p*	Estimate[b]	*SE*	*t*	*p*
Intercept	54.80	3.60	–	–	–	–	–	–
Intervention day	9.60	5.09	1.89	.06	–	–	–	–
Control day								
Participant 1	28.93	5.09	5.68	<.05	83.73	3.60	23.26	<.05
Participant 2	.40	5.09	.08	.94	55.20	3.60	15.33	<.05
Participant 3	10.60	5.09	2.08	<.05	65.40	3.60	18.17	<.05
Participant 4	−12.73	5.09	−2.50	<.05	42.07	3.60	11.69	<.05
Participant 5	41.13	5.09	8.08	<.05	95.93	3.60	26.65	<.05
Participant 6	−6.60	5.09	−1.30	.20	48.20	3.60	13.39	<.05
Participant 7	16.47	5.09	3.23	<.05	71.27	3.60	19.80	<.05
Participant 8	–	–	–	–	54.80	3.60	15.22	<.05
Intervention day increase								
Participant 1	6.07	7.20	.84	.40	15.67	5.09	3.08	<.05
Participant 2	16.13	7.20	2.24	<.05	25.73	5.09	5.05	<.05
Participant 3	−1.53	7.20	−.21	.83	8.07	5.09	1.58	.12
Participant 4	5.80	7.20	.81	.42	15.40	5.09	3.02	<.05
Participant 5	19.07	7.20	2.65	<.05	28.67	5.09	5.63	<.05
Participant 6	2.47	7.20	.34	.73	12.07	5.09	2.37	<.05
Participant 7	−6.53	7.20	−.91	.37	3.07	5.09	.60	.55
Participant 8	–	–	–	–	9.60	5.09	1.89	.06
Model fit								
−2LL	1859.5				1859.5			

[a] Default output given by SPSS or SAS.

[b] Results requested by either ESTIMATE or TEST statements.

activity on intervention days, in which his or her expected physical activity would be 54.80 + 9.60 = 64.40 min. The Control Day main effects for participants 1–7 indicate the difference in physical activity on control days between each participant and participant 8. For example, on control days, participant 1 averaged 28.93 *more* minutes of physical activity than participant 8; thus, participant 1 was expected to have 54.80 + 28.93 = 83.73 min of physical activity on control days (i.e., 83.73–64.40 = 28.93). The Intervention Day effects for participants 1–7 are intervention-by-participant interactions and indicate the difference in the treatment effect between each participant and participant 8. For example, the average increase in simulated physical activity on treatment days for participant 1 was 6.07 min *larger than participant 8*, such that the simulated intervention day increase for participant 1 would be 9.60 + 6.07 = 15.67 min. Taken together, participant 1 would be expected to have 99.40 min of physical activity on treatment days (i.e., 54.80 + 28.93 + 9.60 + 6.07).

Although perhaps useful for didactic purposes, these regression results do not indicate for whom the intervention worked *directly*. However, this information can be found by requesting the necessary linear combinations of the model effects using the post-estimation commands available in most software packages (e.g., ESTIMATE in SAS, TEST in SPSS, LINCOM in STATA, or NEW in M*plus*). Critically, these commands provide not only the requested linear combinations of model effects, but also the appropriate standard errors with which to assess their significance. The model-implied results for the current example are presented in the second set of columns of Table I and reveal that only five of the eight participants had a statistically significant increase in physical activity on their treatment days relative to control days.

Examining for Carryover Effects

To evaluate for carryover effects, a new binary predictor variable was created that represents whether or not the previous day was a treatment day. The main effect of this new carryover predictor is added to the previous model, as well as all two-way interactions between the carryover predictor and the participant dummy codes, and all the three-way interactions among the dummy codes, treatment predictor, and the carryover predictor. Carryover effects are indicated by any three-way interaction being statistically significant. Alternatively, an omnibus *F*-test for the three-way interactions can be requested using CONTRAST in SAS, SPSS, or STATA. In our data, the omnibus *F*-test was not statistically significant, $F(7,200) = 0.41$, $p = .90$, indicating the absence of carryover effects across participants in the current simulated data.

Examining the Variability in Physical Activity Within Each Participant

Up until this point, our model has assumed that each person has the same error variance (i.e., homogeneity) on minutes of physical activity across treatment or control days. The advantage of using a common error variance is the gain in statistical power that results from estimating fewer parameters. In reality, however, some participants will have more (or less) variability in their minutes of physical activity within treatment or control days compared with other participants. Fortunately, whether or not a participant needs his or her own error variance is a testable hypothesis by estimating a *heterogeneous variance model*, which estimates a separate error variance for each participant (see Bryk & Raudenbush, 1988; Hoffman, 2007).

Heterogeneous variance models produce results in a single analysis akin to the estimates provided by conducting individual regression analyses using each participant's data (e.g., Sniehotta et al., 2012). The fit of the heterogeneous variance model is then compared with the fit of the homogeneous variance model via likelihood ratio test (-2ΔLL), with degrees of freedom equal to the difference in the number of estimated parameters. In our simulated data, this test was not statistically significant, indicating the heterogeneous variance model did not fit better than the homogeneous variance model, $-2\Delta\text{LL}(7) = 1859.5 - 1849.1 = 10.4$, $p = .17$, and that using a common error variance across participants was sufficient for these data.

Study IIa: Evaluating Individual Differences in the Treatment Effect

From Fixed to Random Effects

As noted previously, the purpose of random-effects models (otherwise known as multilevel or hierarchical linear models) is to answer the question of *why the intervention works for some participants and not others*. We note that the text that follows is intended to be only a primer; readers interested in more detail can consult Hoffman and Stawski (2009) for more detail.

To understand random effects, we first distinguish the two sides of any linear model. First, *the model for the means* considers how minutes of physical activity vary as a function of predictor variables and includes a fixed intercept, fixed main effects, and fixed interaction effects between predictors used to predict the average minutes of physical activity for each participant on each day. Second, *the model for the variance* considers how the model residuals are related across observations. In a single-level linear regression model for cross-sectional data, there is only one residual for each participant (i.e., the participant-specific difference between the outcome predicted by the model and the participant's actual outcome value), and thus all residuals are assumed uncorrelated. In contrast, a multilevel model assumes a nested data structure (noted using the term *levels*), which for longitudinal data indicates that the repeated observations across days at level 1 are nested within individual participants at level 2. This nested structure creates correlation (or dependency) that we will quantify using random effects. More specifically, random intercepts will be used to describe mean outcome differences across participants, and random slopes will be used to describe differences between participants in the effects of longitudinal predictors, such as the difference between control and treatment days.

Decomposing Variability via Intraclass Correlation

The extent to which residuals from the same person are correlated on average over time can be quantified explicitly using an intraclass correlation once a random intercept is included in the model for the variance (i.e., in moving from a single-level model to a two-level model). The average correlation across occasions from the same person is known as the *intraclass correlation* (*ICC*), which is formed as a ratio of between-person variability to total variability (i.e., in which total variability = between-person + within-person variability) and thus indicates the proportion of variance between participants at level 2. The ICC from our simulated data was .35, indicating that ∼35% of the variance in physical activity was due to individual mean differences between participants, and 65% was due to within-person variation over time. To further quantify the extent of these individual mean differences, we used the random intercept variance to calculate a 95% random effects confidence interval (see Snijders & Bosker, 2011 and the Excel spreadsheet provided in the online Supplementary Materials) around the fixed intercept value of 64.97 min. We found that 95% of our simulated sample was expected to average between 15.99 and 113.94 min physical activity across days.

Evaluating Individual Differences in the Intervention Effect

To evaluate individual differences in the treatment effect, we first estimated a model with only a fixed effect for treatment, in which the average difference in physical activity between control and treatment days is assumed to be the same for all participants. Our results indicated that participants engaged in significantly more physical activity—an average of 26.38 min more—on treatment days compared with control days. However, given that the simulated results in Study I suggested varying increases in simulated physical activity on treatment days across participants, we estimated a second model that also included a random treatment slope for the participant-specific deviation from the fixed treatment effect. As with the random intercept, a random treatment slope variance is estimated to represent all individual differences in the treatment effect, rather than the individual-specific deviations. A covariance between the random intercept and treatment slopes is also estimated to allow a relationship between the amount of physical activity on control days and the treatment effect across persons.

In using a likelihood ratio test to compare these fixed and random treatment slope models, we found that allowing individual differences in the treatment effect (i.e., random slope variance and intercept variance) resulted in a significant improvement in model fit, $-2\Delta LL(2) = 8845.1 - 8795.4 = 49.7$, $p < .05$. To quantify these individual slope differences, we calculated a 95% random effects confidence interval around the treatment effect that indicated 95% of our simulated sample was expected to have a treatment slope between −9.96 and 62.71 min (see Excel spreadsheet provided in the online Supplementary Materials). Clearly, the intervention was not predicted to be effective for every participant in the simulated data—some participants were actually predicted to decrease their physical activity on treatment days! The next step is to include additional predictors to explain why the intervention worked for some participants but not others. This is the focus of Study IIb.

Study IIb: Evaluating why the Intervention Worked

Effect Size as Variance Explained

In a single-level regression model, residual variance is the only variance component and effect size describes the proportion of residual variability explained by the predictors (i.e., R^2). An analog to a single R^2 can be created in multilevel models by squaring the Pearson correlation between the model-predicted outcome and the actual outcome for each participant, which then provides the total outcome variability explained by predictors at all levels of analysis. Alternatively, effect size estimates for the specific variance components estimated are called *pseudo-R^2*, which can be calculated as the proportion reduction in the variance component before and after inclusion of predictors. An Excel spreadsheet provided in the online Supplementary Materials shows the calculation of all pseudo-R^2 values for the multilevel linear model.

Time-Invariant Predictors

Given the presence of significant individual differences in treatment effect, the final step is to attempt to identify why the intervention was more effective for some participants than for others. To do so we will examine two variables that are commonly related to physical activity: self-efficacy and social support (Motl, Dishman, Saunders, Dowda, & Pate, 2007). Self-efficacy was simulated to be a continuous time-invariant predictor. In clinical practice, time-invariant predictors are likely to be collected at baseline before initiating treatment. In our study, the time-invariant predictor of self-efficacy can contribute to the level-2 model by potentially explaining both between-person differences in mean physical activity on control days (i.e., random intercept variance through its main effect) and between-person differences in the treatment effect (i.e., random treatment slope variance through its cross-level interaction with treatment). To maintain a meaningful intercept and treatment main effect, self-efficacy was centered near the grand mean at 25, such that a value of 0 on the new variable indicated a self-efficacy of 25. As with any regression model, centering is a method of improving the ability to interpret intercept and main effects by providing a meaningful zero point, but will not change model fit or model predictions.

Time-Varying Predictors

The treatment variable and social support were simulated as measured daily and are thus considered *time-varying* predictors. In clinical practice, time-varying predictors would likely be recorded as daily diary data, which could be reviewed in session and inform statistical models such as the one presented here. An important caveat to note is that even though they are often thought of as level-1 (time-level) variables, time-varying predictors generally contain both within-person (WP) and between-person (BP) variability. Because we used block randomization in our simulated data, all participants received exactly 15 treatment days. As such, the treatment variable had no BP variability. However, social support, measured as a continuous variable, contained both WP and BP variability. That is, while a participant's level of social support may fluctuate *daily* (at level 1, WP), it may also be the case that some participants may have higher levels of social support than other participants *on average* (at level 2, BP). As a result, social support could have differential effects on physical activity across levels of analysis, and so it is important to specify its effects in the model allowing these level-specific effects to be observed.

The Convergence Effect and Variable Partitioning

It is important to note that because time-varying predictors typically contain level-1 and level-2 variability, they are really two variables instead of one. Thus, inclusion of a time-varying predictor directly into a model without first partitioning its WP and BP sources of variance into separate predictions will result in a *convergence* slope (aka, conflated or composite slope; a weighted combination of the WP and BP slopes) that will ultimately lead to uninterpretable and incorrect parameter estimates (see Raudenbush & Bryk, 2002, pp. 138–139 for full description). Thus, to avoid such model mis-specification, we recommend explicitly partitioning the WP and BP portions of any time-varying predictor that has nonzero BP variability (as determined by the ICC for the predictor from an empty model). For our simulated data, social support had an ICC of 0.41, indicating that 41% of the variability in social support was between participants. As a result, we partitioned the WP and BP variance in social support into two uncorrelated predictors using *person-mean-centering* for the level-1 predictor, as described by Hoffman and Stawski (2009).

In person-mean centering (also known as group-mean centering in clustered models, e.g., persons nested within groups), each person's mean social support across days (i.e., person-mean social support) was subtracted from the time-varying social support variable at each occasion. This new WP social support predictor—composed only of level-1 variability—represents a participant's time-specific deviation from their *usual* level of social support. Specifically to our study, WP social support examined whether a participant increased physical activity from their usual level if they had more social support than usual on a given day.

The second predictor, person-mean social support—composed only of level-2, BP variability—was then centered at a constant of 3, which was near its grand mean. Person-mean social support was included as a predictor in the level-2 model to explore how between-person differences in *average* levels of social support may predict between-person differences in mean physical activity on control days (i.e., random intercept variance through its main effect) and between-person differences in the treatment effect (i.e., random treatment slope variance through its cross-level interaction with treatment).

Model Results

The final model provides interpretations for the most common effects considered in a multilevel linear model. Specifically, we model two time-invariant predictors, two time-varying predictors (one with no BP variability, one with some BP variability), and three cross-level interactions. While not exhaustive of all possible effects, additional time-invariant variables, time-varying variables, and interaction effects can be added sequentially to the model to evaluate their importance. For example, if two time-varying predictors (both with some BP variability) are being considered, their WP and BP variance should be partitioned and added separately to the model to evaluate their statistical significance via *p*-values as well as their unique contribution to the model via pseudo-R^2. Similar to a single-level linear regression, the variable that explains the largest proportion of variance can be considered the most important.

The final model results are presented in Table II and illustrated in Figure 1; syntax, results, and pseudo-R^2 estimates for the intermediate model building process are provided in the online Supplementary Materials. The predictors, as a set, explained ∼23.89% of the *total* variability in physical activity. We interpret each of the final model parameters, beginning with the fixed intercept, which indicates average minutes of physical activity when all predictors are centered at zero: Specifically on control days, for participants with self-efficacy of 25, with person-mean social support of 3, who were at their usual daily level of social support (within-person = 0) averaged 52.01 min of physical activity.

Likewise, because all predictors are involved in one or more interaction effects, their main effects are interpreted as simple main effects, which are conditional on the interacting predictor(s) = 0. Accordingly, the treatment effect can be understood as a simple main effect conditional on self-efficacy, within-person social support, and person-mean social support = 0. Thus, specifically for a participant who has self-efficacy of 25, an average social support of 3, who is also at their usual daily level of social support, physical activity was significantly higher by 27.74 min on treatment days than on control days. This within-person treatment effect explained ∼15.48% of the level-1 residual variance (i.e., the remaining within-person daily fluctuation in physical activity).

During the model-building process, we evaluated for carryover effects by examining whether a binary predictor for whether the previous day was a treatment or control day significantly moderated the treatment effect. Similar to the fixed effects described in Study I, this new predictor

Table II. *Final Results Using Random-Effects Model*

Fixed effects (model for the means)	Estimate	*SE*	*t*	*p*
Intercept	52.01	3.90		
Intervention day	27.74	3.76	7.38	<.05
WP social support on control day	3.10	1.95	1.59	.11
WP social support on intervention day	11.67	1.95	5.99	<.05
PM social support on control day	13.40	6.00	2.23	<.05
PM social support on intervention day	22.33	7.73	2.89	<.05
Intervention day-by-WP social support	8.57	2.76	3.10	<.05
Intervention day-by-PM social support	8.93	5.79	1.54	.13
Self-efficacy on control day	.63	.48	1.32	.20
Self-efficacy on intervention day	1.67	.61	2.74	<.05
Intervention day-by-self efficacy	1.05	.46	2.29	<.05
Random effects (model for the variances)	**Estimate**	***SE***	***z***	***p***
Intercept	367.36	109.60	3.35	<.05
Intervention day slope	280.92	102.12	2.75	<.05
Intercept–intervention day covariance	−.80	75.56	−.01	.99
Residual	852.69	41.61	20.49	<.05
Model fit				
−2 Log Likelihood		8745.5		
AIC		8769.5		
BIC		8786.3		

Note. WP = within-participant; PM = person-mean.
PM social support was centered at 3.
Self-efficacy was centered at 25.

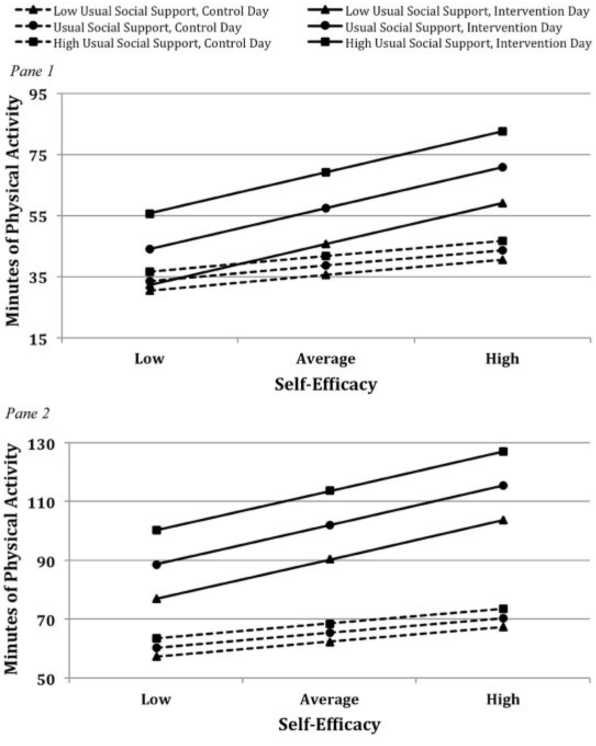

Figure 1. Pane 1: Low person-mean social support (i.e., person-mean social support = 2). Pane 2: High person-mean social support (i.e., person-mean social support = 4).

was then included in the model alongside to the treatment variable and their interaction, with a statistically significant interaction effect indicating a carryover effect. Results indicated no significant carryover effects, $t(813) = -0.38$, $p = 0.70$. However, we suggest that other pediatric psychologists using this technique should not expect a null finding and testing for carryover effects is likely to be an important part of the model-building process.

Next, we consider the WP effect of social support. Because WP social support interacts with the treatment variable, the simple main effect of WP social support indicated that specifically on a control day a one-point increase beyond a participant's *usual level of social support* was related to a nonsignificant increase in physical activity *on that day* by 3.10 min. This level-1 simple main effect—as evaluated specifically for control days—explained an additional 3.49% of daily fluctuation in physical activity (i.e., residual variance) and is seen in Figure 1 as the difference between the dashed lines. The treatment-by-WP social support interaction can then be interpreted as how the effect of greater than usual levels of social support differed between control and treatment days. Thus, when a participant's social support was rated one-point higher than their usual levels, physical activity increased significantly by an *additional 8.57 min on that treatment day*, for a total significant effect of 3.10 + 8.57 = 11.67 min (as shown by the difference between the solid lines in the figure). This level-1 interaction effect explained an additional 1.08% of the daily fluctuation in physical activity (i.e., residual variance) and is shown as the increasing difference of the respective markers between the dashed and solid lines in each panel of Figure 1. For example, in panel 1 at an average level of self-efficacy, the small difference in physical activity between intervention and control days at lower than usual social support level (i.e., triangle markers) increased to a larger difference at higher than usual social support level (i.e., square markers).

Now we consider the effect of person-mean social support. Because person-mean social support interacted with the treatment variable, the simple main effect of person-mean social support indicated that specifically on a control day a one-point increase in the person-mean of social support indicated that physical activity increased significantly by an average of 13.40 min. This level-2 simple main effect—as evaluated for control days—explained 12.40% of mean differences in physical activity across participants on control days (i.e., random intercept variance), shown by the increase in the dashed lines from panel 1 to panel 2. The treatment-by-person-mean social support interaction effect can then be interpreted as how the effect of greater person-mean levels of social support differed between control and treatment days. Specifically, a one-point increase in person-mean social support was related to a nonsignificant 8.93-min increase in physical activity *on treatment days* compared to control days, for a total significant effect of 13.40 + 8.93 = 22.33 min (as shown by the increase in the solid lines from panel 1 to panel 2). This cross-level interaction explained ~6.81% of the individual differences in the treatment effect (i.e., random treatment slope variance), and is shown by comparing the differences between the solid and dashed lines across panels 1 and 2. For example, consider the markers at high levels of self-efficacy: The differences in physical activity between intervention and control days in panel 1 were more pronounced at higher person-mean social support in panel 2.

Next, we consider the effect of time-invariant self-efficacy, which also interacted with the treatment variable. The simple main effect of self-efficacy indicated that, specifically on a control day, a one-point increase in self-efficacy resulted in a nonsignificant increase in physical activity by an average of 0.63 min. The level-2 simple main effect—as evaluated for control days—explained an additional 5.89% of mean differences in physical activity across participants on control days (i.e., random intercept variance), as seen in panels 1 and 2 by the slightly positive slope of the dashed lines. The cross-level treatment-by-self-efficacy interaction effect can then be interpreted as how the effect of self-efficacy differs between control and treatment days: A one-point increase in self-efficacy was related to a significant 1.05-min increase in physical activity on treatment days compared to control days, for a total significant effect of 0.63 + 1.05 = 1.67 min within rounding error (as shown by the positive slope for the solid lines in each panel). This cross-level interaction explained an additional 19.73% of the individual differences in the treatment effect (i.e., random treatment slope variance) and is shown in either panel by the greater difference between the dashed and solid lines at higher levels of self-efficacy.

Finally, we note that our random-effects model could be extended to include differential within-person variability (i.e., heterogeneity in level-1 residual variance), as we had examined in the previous fixed-effects model in Study I. However, just as we switched from a focus of *who improves* in Study I to *why do some participants improve more than others* in Study IIa and IIb, here the emphasis would be on *predicting* between-person differences in within-person variability, rather than simply allowing different residual variances by participant ID. But, given their additional computational complexity and lack of availability in some popular software programs (such as SPSS), we do not pursue random-effects models with heterogeneous residual variances here (but see Snijders & Bosker, 2011 for more details).

Limitations of N-of-1 RCTs and Alternative Methodologies

As noted previously, N-of-1 RCTs are not suitable for all research questions. Specifically, research questions that pertain primarily to external validity or generalizability are likely to be best addressed using traditional fully powered between-groups designs. Moreover, there may be instances where an investigator is interested in a research question that does not meet the criteria of reversibility. Indeed, many interventions in pediatric psychology are aimed at teaching problem solving or coping skills that the investigator undoubtedly hopes will be sustained beyond the treatment phase (e.g., Wysocki et al., 2006). When this is the case *and* the research question is early in its stage of development, other small-*n* techniques may be most appropriate (*see* Rapoff & Stark, 2008 *for review*). In particular, a multiple-baseline design may be of interest when a treatment effect is not expected to be reversible within a given participant. While it is beyond the scope of this article, it is possible to apply multilevel modeling to this design with each day representing within-person variability and experimental condition as a covariate. However, it is important to note that in multiple-baseline designs observations will not be balanced within experimental condition. That is, some participants will spend much more time in intervention than control and vice versa. As such, it is critically important to define an appropriate time metric and may limit the ability to answer the *for whom does the treatment work* question, which is better addressed by N-of-1 RCTs.

Conclusions

The combination of N-of-1 RCTs and multilevel modeling may represent a methodological advancement in pediatric psychology, leading the field to new scientific discovery and providing an accessible tool for clinicians to conduct scientifically informative clinical work. We have attempted to provide an overall narrative and a data analysis example that is rigorous yet accessible to a doctoral-level pediatric psychologist. At a minimum, most practicing pediatric psychologists can likely use the fixed-effect models presented in Study I within commonly available statistical packages (e.g., SPSS, SAS) to answer the question of *what works for whom*. Admittedly, the question of *why* an intervention works is more complex, but we have attempted to provide a springboard to allow researchers to answer these questions with random-effects models.

Supplementary Data

Supplementary data can be found at: http://www.jpepsy.oxfordjournals.org/

Acknowledgments

The authors gratefully acknowledge Alayna Tackett for her comments on a previous version of this manuscript.

Conflicts of interest: None declared.

References

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology, 27*, 379–387. doi:10.1037/0278-6133.27.3.379

Altman, D. G., & Bland, J. M. (1999). Statistics notes: How to randomise. *British Medical Journal, 319*, 703–704.

Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition order. *Psychological Methods, 17*, 600–614. doi:10.1037/a0029310

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretaions. *Psychological Bulletin, 104*, 396–404.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*, 467–478. doi:10.3758/BRM.40.2.467

Cushing, C. C., Jensen, C. D., & Steele, R. G. (2011). An evaluation of a personal electronic device to enhance self-monitoring adherence in a pediatric weight management program using a multiple baseline design. *Journal of Pediatric Psychology, 36*, 301.

Cushing, C. C., & Steele, R. G. (2010). A meta-analytic review of eHealth interventions for pediatric health promoting and maintaining behaviors. *Journal of Pediatric Psychology, 35*, 937–949. doi:10.1093/jpepsy/jsq023

Guyatt, G., Sackett, D., Adachi, J., Roberts, R., Chong, J., Rosenbloom, D., & Keller, J. (1988). A clinician's guide for conducting randomized trials in individual patients. *CMAJ: Canadian Medical Association Journal, 139*, 497–503.

Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research, 42*, 609–629.

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development, 6*, 97–120.

Ingerski, L. M., Hente, E. A., Modi, A. C., & Hommel, K. A. (2011). Electronic measurement of medication adherence in pediatric chronic illness: A review of measures. *The Journal of Pediatrics, 159*, 528. doi:10.1016/j.jpeds.2011.05.018

Kazdin, A. E. (1997). A model for developing effective treatments: Progression and interplay of theory, research, and practice. *Journal of Clinical Child Psychology, 26*, 114–129.

Kahana, S., Drotar, D., & Frazier, T. (2008). Meta-analysis of psychological interventions to promote adherence to treatment in pediatric chronic health conditions. *Journal of Pediatric Psychology, 33*, 590–611. doi:10.1093/jpepsy/jsm128

Keller, J. L., Guyatt, G. H., Roberts, R. S., Adachi, J. D., & Rosenbloom, D. (1988). An N of 1 service: Applying the scientific method in clinical practice. *Scandinavian Journal of Gastroenterology, 23*(S147), 22–29.

Meltzer, L. J., Montgomery-Downs, H. E., Insana, S. P., & Walsh, C. M. (2012). Use of actigraphy for assessment in pediatric sleep research. *Sleep Medicine Reviews, 16*, 463–475. doi:10.1016/j.smrv.2011.10.002

Motl, R. W., Dishman, R. K., Saunders, R. P., Dowda, M., & Pate, R. R. (2007). Perceptions of physical and social environment variables and self-efficacy as correlates of self-reported physical activity among adolescent girls. *Journal of Pediatric Psychology, 32*, 6–12. doi:10.1093/jpepsy/jsl001

Nelson, T. D., Aylward, B. S., & Rausch, J. R. (2011). Dynamic p-technique for modeling patterns of data: Applications to pediatric psychology research. *Journal of Pediatric Psychology, 36*, 959–968. doi:10.1093/jpepsy/jsr023

Rapoff, M., & Stark, L. (2008). Editorial: Journal of Pediatric Psychology statement of purpose: Section on single-subject studies. *Journal of Pediatric Psychology, 33*, 16–21. doi:10.1093/jpepsy/jsm101

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

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., & Martinez, A. (2011). Optimal design. software for multi-level and longitudinal research (Version 3.01) [Software]. Available from http://hlmsoft.net/od/.

Schluter, P., & Ware, R. (2005). Single patient (N-of-1) trials with binary treatment preference. *Statistics Medicine, 24*, 2625–2636. doi:10.1002/sim.2132

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). New York: Houghton Mifflin Company.

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage Publications Inc.

Sniehotta, F. F., Presseau, J., Hobbs, N., & Araújo-Soares, V. (2012). Testing self-regulation interventions to increase walking using factorial randomized N-of-1 trials. *Health Psychology, 31*, 733–737.

Steele, R. G., Aylward, B. S., Jensen, C. D., Cushing, C. C., Davis, A. M., & Bovaird, J. A. (2012). Comparison of a family-based group intervention for youths with obesity to a brief individual family intervention: A practical clinical trial of positively fit. *Journal of Pediatric Psychology, 37*, 53–63.

Tsapas, A., & Matthews, D. R. (2008). N of 1 trials in diabetes: Making individual therapeutic decisions. *Diabetologia, 51*, 921–925. doi:10.1007/s00125-008-0983-2

Wysocki, T., Harris, M. A., Buckloh, L. M., Mertlich, D., Lochrie, A. S., Taylor, A., . . . White, N. H. (2006). Effects of behavioral family systems therapy for diabetes on adolescents' family relationships, treatment adherence, and metabolic control. *Journal of Pediatric Psychology, 31*, 928–938. doi:10.1093/jpepsy/jsj098