DOI: 10.1111/spc3.12487

ARTICLE

WILEY

Addressing two blind spots of commonly used experimental designs: The Highly-Repeated Within-Person approach

Vivian Zayas[1] | Vasundhara Sridharan[2] | Randy T. Lee[1] | Yuichi Shoda[2]

[1]Cornell University, Ithaca, NY, USA

[2]University of Washington, Seattle, WA, USA

Correspondence
Vivian Zayas, Department of Psychology, Cornell University, Ithaca, NY, 14853-7601, USA.
Email: vz29@cornell.edu

Abstract

Two well documented but still neglected blind spots of often-used study designs limit a researcher's ability to make inferences about psychological phenomenon. First, typical designs focus on effects of conditions at the group level and are not able to assess the extent to which effects characterize each participant in the study. This blind spot can lead to erroneous (or incomplete) conclusions about the effects of manipulations both for a given participant and at the group level. Second, commonly used research designs often use a limited sample of stimuli, constraining conclusions to the particular stimuli. This blind spot can lead to non-replication when different stimuli are used. We propose that the Highly-Repeated Within-Person (HRWP) approach helps mitigate these limitations. Using a study on the effects of anti-smoking messages, we illustrate how the HRWP approach helps alert researchers when the conclusions at the group level may not apply to all (or any) participant, quantifies the heterogeneity of effects of manipulations across people, and increases confidence regarding the generalizability of the effects. We discuss how the HRWP approach may help conceptualize issues of replicability in a new light.

1 | INTRODUCTION

Commonly used research designs that focus on effects at the group level have two blind spots (e.g., Fisher, 2015; Gallistel, Fairhurst, & Balsam, 2004). First, the conclusions based on them may not apply to every, or possibly, any, individuals in the study. Second, studies conducted with only a limited sample of stimuli to represent the conditions are vulnerable to the concern that the conclusions reflect the idiosyncrasies of the stimuli used in the study. In the present paper, we describe the Highly-Repeated Within-Person (HRWP) approach (Lee, 2009; LeeTiernan, 2002; Shoda, 1999, 2004; Shoda & LeeTiernan, 2002; Shoda, Mischel, & Wright, 1994; Whitsett & Shoda, 2014; Wilson, 2008; Zayas & Shoda, 2007; Zayas, Whitsett, Lee, Wilson, & Shoda, 2008), which can serve to mitigate these limitations.

2 | STUDY DESIGNS AND THEIR LIMITATIONS

2.1 | Blind spot 1: Group-level effects do not necessarily correspond to the effect for any individual

Researchers are often interested in the effect of a manipulation, *x*, on an outcome of interest, *y*.[1] In the typical research design, groups of people are exposed to one or more conditions, and aggregate, group-level effects (means for each condition) are compared. A manipulation is considered effective if, on average, participants in the experimental condition fare better (at statistically significant levels) than participants in a control condition.

By not assessing the extent to which a manipulation is effective for a *given* participant, such approaches are blind to the potential individual-to-individual variability (or heterogeneity) in the effect of a treatment. In a between-subject design, groups of people are exposed to only one of a number of conditions. Thus, the design yields no information about how a person would respond to other conditions. In a within-subject design, people are exposed to all conditions. But here too, because each participant is typically observed only once in each condition, it is not possible to ascertain whether variations in a given participant's responses across conditions are random fluctuation or meaningful differences in how the person responds to different situations (e.g., "behavioral signatures," Shoda et al., 1994).

A practical and theoretical implication of this blind spot is that conclusions based on the group level may be erroneous for a group of individuals (e.g., Wood & Brumbaugh, 2009) or any individual (e.g., Gallistel et al., 2004). Even if a manipulation appears to have an effect at the group level, there could be many participants for whom it has no effect. Conclusions based on the group-level result would represent a Type I, or false positive, error (i.e., concluding that the manipulation had the intended effect when it did not) for these individuals. It is also possible that there may be individuals for whom the effect is reliably in the *opposite* direction from the one observed at the group level. In intervention research, exposing such individuals to an intervention that may in fact harm them is arguably unethical.

Conversely, a manipulation may appear ineffective because there is no statistically significant difference between group averages. But this may be because a reliable and positive effect that occurs for a subset of individuals is obscured by the presence of individuals for whom the intervention has no effect or possibly has an unintended, opposite effect. For the individuals for whom the true effect of the intervention is positive, conclusions based on the results at the group level would represent a Type II, or false negative, error (i.e., concluding that the intervention is not effective for them, while in actuality it is). False negatives can be costly, because a potentially useful intervention may be lost and individuals are denied the possibility of benefitting from it.

Moderation analyses offer only a limited solution to the problem of assessing individual-to-individual variability (or heterogeneity) in the effect of a treatment. This is because to examine moderation by an individual difference variable, one must identify *a priori* the moderating variable. But what if such variables are not even known? If no variable

that moderates the effect has been found, it may be tempting to conclude that the conclusions at the group level apply similarly to all individuals. However, such an inference is not warranted; one cannot rule out the possibility that there may exist yet-to-be-discovered moderator variables. The sobering reality is that one can never know if all, or even the most important, potential moderators have been identified *a priori*.

Thus, the presence of unexamined heterogeneity can lead to erroneous conclusions for given individuals. Unexamined heterogeneity can also lead to erroneous (or incomplete) conclusions about the effectiveness of manipulations.

2.2 | Blind spot 2: The treatment effects may be due to idiosyncrasies of the stimuli used

Another limitation of typical research designs is that the observed effect may be specific to the instantiation of the manipulation. For example, in a study on the effect of the gender of the confederate, the result may be specific to the idiosyncrasies of the particular individuals (stimuli) who serve as confederates (e.g., Fiedler, 2011; Judd, Westfall, & Kenny, 2012; Westfall, Kenny, & Judd, 2014; Westfall, Judd, & Kenny, 2015; Wells & Windschitl, 1999).[2] Although the importance of having an appropriate sample size of participants is well recognized, less attention is given to the sample size of stimuli. Just as researchers require reasonably sized samples of participants to draw conclusions from the sample to the population of interest, reasonably sized samples of stimuli are needed to draw conclusions involving the construct of interest. Failure to do so leaves the conclusions open to the possibility that they are limited to the particulars of the stimuli used and may not replicate when different stimuli are used. Yet most researchers still underestimate this limitation (Judd, Westfall, & Kenny, 2012; Wells & Windschitl, 1999) and often make unwarranted generalizations.

3 | THE HIGHLY-REPEATED WITHIN-PERSON (HRWP) APPROACH: ADDRESSING THE BLIND SPOTS OF TRADITIONAL EXPERIMENTAL DESIGNS

The HRWP design is well suited for addressing the blind spots of standard approaches. Here, we illustrate the steps of the HRWP approach (outlined in Figure 1) for conducting a study assessing the effectiveness of anti-smoking messages. Greater detail of the HRWP approach can be found elsewhere (Whitsett & Shoda, 2014; Zayas et al., 2008; see also Lee, 2009; LeeTiernan, 2002; Shoda, 1999, 2004; Shoda & LeeTiernan, 2002; Shoda et al., 1994; Wilson, 2008; Zayas & Shoda, 2007).

The present application of HRWP approach extends previous demonstrations of the HRWP approach, as well as other commonly used interactionist approaches (Fleeson, 2007a; Fleeson, 2007b; Fleeson, Malanos, & Achille, 2002; Wood & Brumbaugh, 2009). In particular, we focus on quantitatively and visually gauging the extent to which the effects of interest vary from participant to participant. Additionally, we show how the HRWP design, by encouraging the use of a greater number of representative stimuli, increases confidence in the construct validity of manipulations, and thus confidence in the generalizability of the conclusions beyond the specific stimuli used.[3] We discuss how features of the HRWP design may provide important insights into the generalizability of findings across different people as well as the replicability of findings across studies.

3.1 | Illustration of the HRWP design: An anti-smoking message study

Step 1. Identify behavior of interest and situations in which it occurs

In the illustrative study, we focused on attitudes towards smoking (our behavior of interest), and how it varies as a function of the anti-smoking message being viewed (our situations or stimuli of interest).

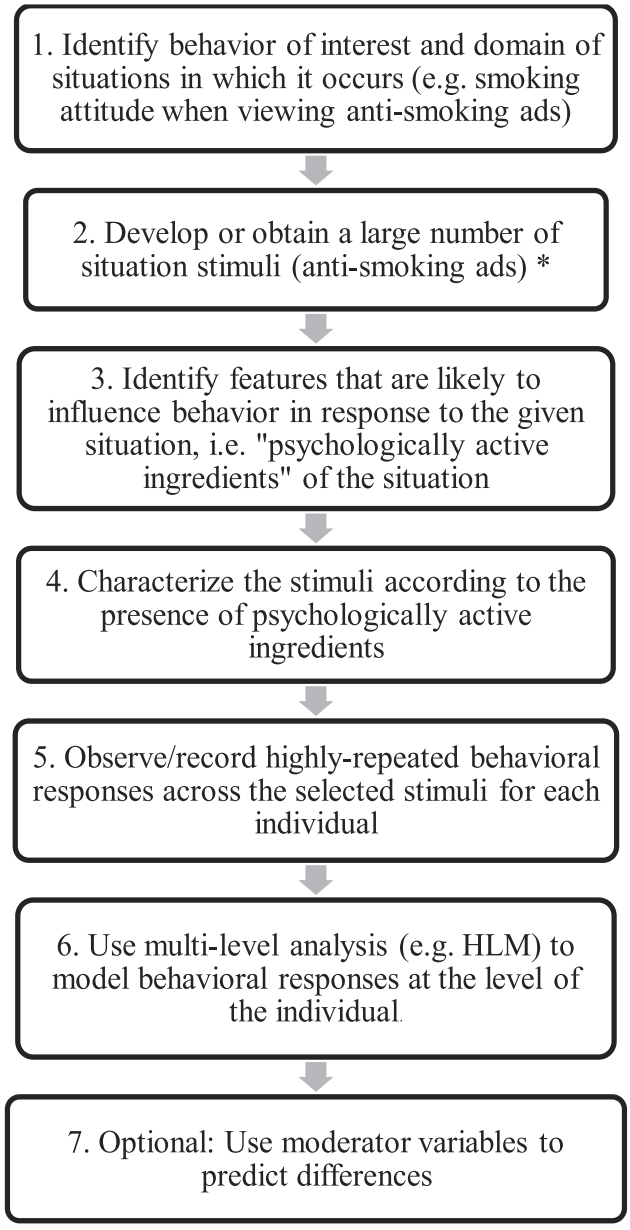

FIGURE 1 Steps involved in the Highly-Repeated Within-Person approach to assess the effects of psychological features of situations. *Note that the statistical power of this approach depends greatly on the number of stimuli. With a small number of stimuli (e.g., less than 30), confidence intervals for individual-level results are large, decreasing statistical power to detect an effect for a given individual, as well as making it more difficult to interpret null results as indicating no effect for the individual. Adapted from "An approach to test for individual differences in the effects of situations without using moderator variables," by Whitsett & Shoda, 2014. *Journal of Experimental Social Psychology, 50*, 94–104. Copyright 2014 by Elsevier. Adapted with permission

Step 2. Obtain or develop stimuli that represent the situations of interest

Sixty anti-smoking messages with still images were selected from extensive online searches of anti-smoking campaigns in the US (e.g., FDA warning labels, see Figure 2). For detailed information about methods, see Sridharan (2015).

Step 3. Identify the key features of situations

Researchers either manipulate or measure the construct of interest. Here, we focused on the following three constructs (also referred to as features) of anti-smoking messages, along which the 60 messages vary: (a) *Graphic*

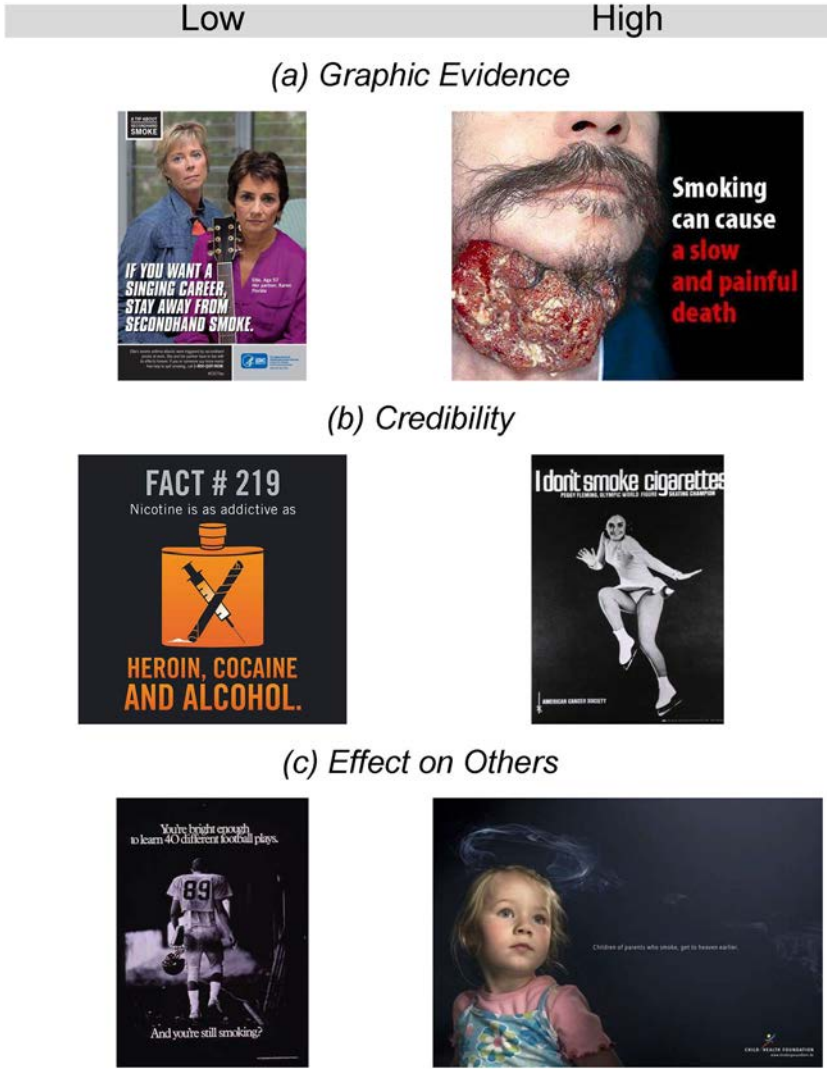

FIGURE 2 Example anti-smoking messages used as stimuli that are high versus low in the three message features of (a) graphic evidence, (b) credibility, and (c) effect on others

evidence—the degree to which messages contain shocking or disgusting imagery of the negative consequence of smoking. (b) *Credibility*—the degree to which messages sponsored by an official source or depicting real people. (c) *Effects on others*—the degree to which messages convey how smoking affects others (e. g., second hand smoke and effects on children of parents who smoke).[4]

Step 4. Characterize the stimuli regarding the feature(s) of interest

Next, to evaluate the anti-smoking messages on the extent to which they reflected the three features identified in Step 3, we recruited a sample of raters (N = 132). Each of the 22 message characteristics had six raters (see Table A1 in the Appendix). Ultimately, each anti-smoking message was characterized by three scores indicating the extent to which the message reflected *graphic evidence* (α = .94), *credibility* (α = .76), and *effects on others* (α = .92).

Step 5. Measure participants' response across the set of stimuli

Finally, in the main study, we recruited participants (N = 160), each of whom viewed all 60 anti-smoking messages (presented in randomized order) and indicated their attitude towards smoking in response to viewing each anti-smoking message. Specifically, after viewing each message, participants reported their attitude towards smoking using a feeling thermometer, which has been shown to correlate consistently with actual smoking behavior (Swanson, Rudman, & Greenwald, 2001) and is well suited to being completed repeatedly in a single session.

Step 6. Examine the effects of stimulus features for each individual

To illustrate the HRWP approach and heeding the recommendation by John Tukey (Tukey, 1977) and others, we begin by a descriptively examining the effects of the features of anti-smoking messages (e.g., credibility) within *each* individual. We then discuss results of a formal multilevel modeling (MLM) to test the statistical significance of the variability of the effects of these features across people.

3.1.1 | Descriptive illustration

To examine the effect of message feature on anti-smoking attitudes, we created within-person scatterplots, representing the effect of message feature on smoking attitude for *each* participant. For example, as shown in Figure 3, for participant 835, the degree to which messages focused on the *effects of smoking on others* was negatively correlated with attitudes towards smoking (r = −.30, p = .02). In contrast, for participant 535, the relation was also statistically significant, but in the opposite direction (r = .28, p = .03), and for participant 576, there was no clear relationship (r = .01, p = .94).

We repeated this process for each of the 160 participants separately and depicted the results as histograms of within-person correlations (Figure 4). The height of each bar corresponds to the number of individuals whose within-person correlation corresponds to the interval the bar represents. For example, the top, left panel of Figure 4 indicates that there was one person whose within-person correlation was between −0.7 and −0.6. This histogram shows that the median for the sample was −.09. However, the histogram also shows that within-person correlations varied across participants. Of course, such variability may be due to momentary fluctuations in behavior. Thus, even if the "true" within-person correlations are the same for all individuals, we would not expect the observed within-person correlations to be identical for all participants.

The question then becomes: Is the observed variability, depicted in the histograms in Figure 4, within the range expected by chance? Or is the observed variability greater than what is expected by chance? If the latter, this would

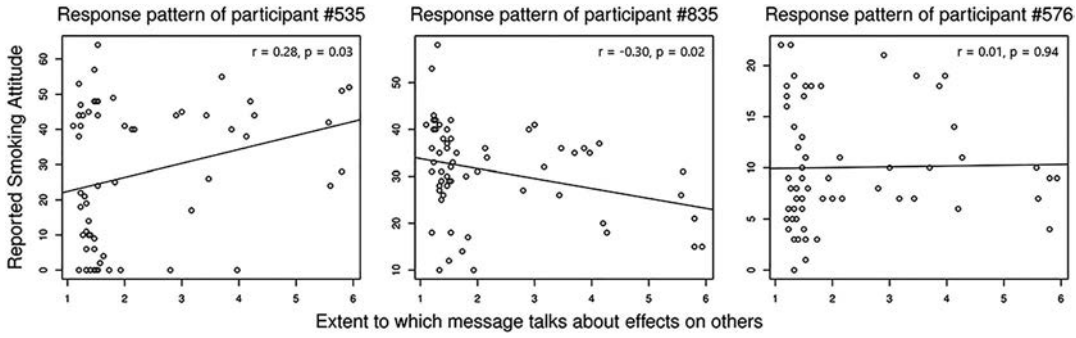

FIGURE 3 Data from three participants illustrating how messages that show the *effect of smoking on others* are associated with either higher (left panel) or lower (middle panel) positive attitudes towards smoking or where there is no association (right panel)

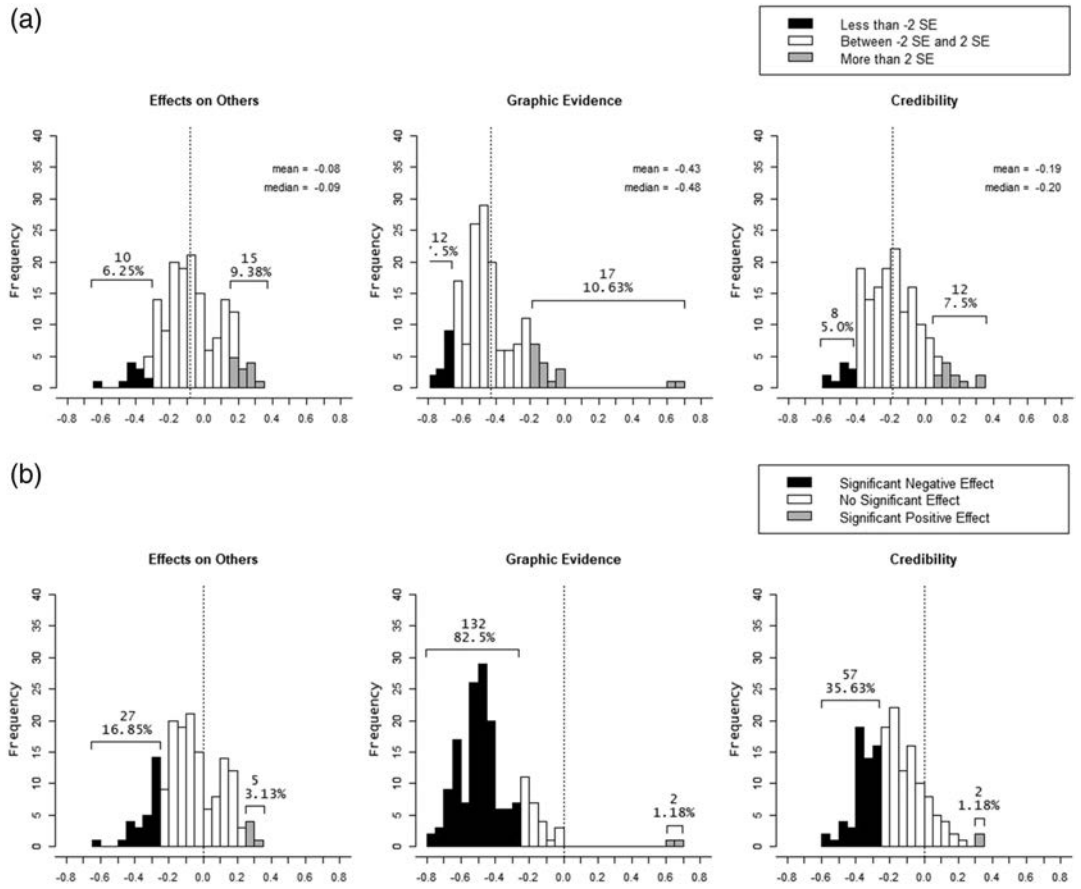

FIGURE 4 Histograms of within-person correlations between *effects on others* (left column), *graphic evidence* (middle column), and *credibility* (right column) and attitudes towards smoking. The vertical axis indicates the number of participants with within-person correlations corresponding to each bar's position on the horizontal axis. In panel A, areas of the histograms are shaded to indicate where less than 5% of the sample are expected to be, if person-to-person variability simply reflects chance. Specifically, the vertical dashed line indicates the average within-person correlation for the entire sample, areas shaded in black are less than 2 times the standard error of correlations for $N = 60$ (the number of observations used to compute each within-person correlation), and areas shaded in gray are more than 2 times the *SE*. The formula used to estimate the standard error assumes that the sampling variation reflects both random sampling of stimuli and momentary "noise." In the present study, however, all participants were exposed to the same set of 60 stimuli. Thus, this illustration is likely to be a conservative estimate of the proportion of participants whose effects deviate from the group average due to the genuine heterogeneity of effects across participants. In panel B, the vertical dashed line indicates a 0 within-person correlation, and areas of histograms are shaded to indicate participants whose within-person correlations are statistically significant at $p < .05$, two-tailed, and negative (shaded in black) and positive (shaded in white)

signal possible moderation—that the effect of the message feature on attitudes towards smoking systematically and reliably differed across participants and this variability is unlikely the result of chance.

Assuming a roughly normal distribution, under the null hypothesis that the within-person correlation in the population is the same for all participants, we would expect only about 5% of the individuals to have within-person correlations beyond 2 standard errors (SEs) from the mean within-person correlation.[5] In the top, left panel in Figure 4, the shaded area identifies those participants who had within-person correlations that were beyond 2 SEs in either direction from the mean. For predicting participants' attitude towards smoking from the message feature *effects on*

others, 16% of the participants had within-person correlations that were greater than 2 SEs from the mean. For *graphic evidence* and *credibility*, the shaded areas contain 18% and 13% of the participants, respectively. This is consistent with the existence of genuine individual-to-individual variations in within-person correlations, and the conclusion that participants in this sample differ from one another in how message characteristics (e.g., graphic evidence) affects their attitude towards smoking.

3.1.2 | Using MLM to examine individual-to-individual variations of treatment effects

To provide a statistical test of the extent of variability in these within-person correlations, we used mixed level modeling (MLM). Attitude towards smoking after viewing each message was modeled as a function of the extent to which the message is characterized by *graphic evidence*, *credibility*, and *effects on others*. Because all participants viewed the same 60 messages,[6] we explicitly modeled message as a nominal random factor. All the variables were entered simultaneously in the level-1 model, which is the within-person analysis. Critically, in our MLM model, we specified the effect of features as a random factor, varying across participants. As shown in Equation (1), the model predicts, for each participant *j*, his or her attitude towards smoking in response to each specific message *i* as a function of the three features:

Level-1 model

$$\text{ATTITUDE}_{ij} = b_{0ij} + b_{1j}*(\text{EonOTHERS}_{\text{i}}) + b_{2j}*(\text{CRED}_{\text{i}}) + b_{3j}*(\text{GRAPHIC}_{\text{i}}) + r_{ij}. \quad (1)$$

Of interest are the *b* coefficients representing the within-person regression coefficients predicting individuals' smoking attitudes from each message feature. For example, b_{3j} refers to participant *j*'s coefficient for predicting her smoking attitudes from the messages' level of *graphic evidence*, while statistically controlling for other message characteristics. The level-1 model also includes r_{ij}, which refers to the residual.

The level-2 model predicts the level-1 coefficients as a function of participants (i.e., subscript *j*) and stimuli (i.e., subscript *i*), as follows:

Level-2 model

$$b_{0ij} = \gamma_{00} + u_{0j} + u_{0i}, \quad (2)$$

$$b_{1j} = \gamma_{10} + u_{1j}, \quad (3)$$

$$b_{2j} = \gamma_{20} + u_{2j}, \quad (4)$$

$$b_{3j} = \gamma_{30} + u_{3j}. \quad (5)$$

Most relevant, as shown in Equations (3)–(5), the model predicts each participant *j*'s effect of the message feature (b_{1j}, b_{2j}, and b_{3j}) from the effect for the sample as a whole (γ_{10}, γ_{20}, and γ_{30}) and u_{1j}, u_{2j}, and u_{3j} representing residuals (i.e., how the effect of each feature for participant *j* deviates from γ_{10}, γ_{20}, and γ_{30}). For example, γ_{30} represents the average effect of *graphic evidence* in the sample on the whole and u_{3j} reflects the extent to which the effect of *graphic evidence* on smoking attitudes varies across participants. Equation (2) in the level-2 model predicts each participant *j*'s intercept, which is not of central interest here; it also includes u_{0j}, which represents the residual variability between participants, and u_{0i}, which represents the residual variability between messages after accounting for the three message characteristics.

Central to a main aim of the present work, namely, examining the heterogeneity of the effect of the feature, are the variance of the participant-to-participant variations in u (in particular, u_{1j}, $u_{2j,}$ and u_{3j}). Each variance component can be tested against the null (i.e., that the true effect is the same for all participants). Thus, when the variation is greater than what is expected by chance, this reflects that the effect of a psychological feature varies across participants.

3.1.3 | Interpreting MLM results

The results showed that, as a group on the whole, messages with greater *graphic evidence* and higher *credibility* significantly predicted anti-smoking attitudes (γ_{30} = −4.04, p < .001, and γ_{20} = −.93, p < .006, respectively). The extent to which the messages mentioned the effects of *smoking on others* did not (γ_{10} = −.08, p = .759).

But is there evidence for meaningful person-to-person variations in the effects of message characteristics? The answer was a resounding *Yes*. As shown in Table 1, according to the likelihood ratio test to assess the significance of a random effect (Hayes, 2006), the null hypothesis that the observed person-to-person variations in the effects was only due to measurement error, or momentary fluctuations in behavior, was clearly rejected for all three message features (*p*s < .001). Thus, the fact that different participants had different within-person correlations at least partly reflects genuine differences from person to person.

4 | DISCUSSION

4.1 | Addressing blind spot 1: When conclusions at the group level lead to erroneous inferences at the individual level

In the present data, the group-level conclusions did not always apply to an individual. To illustrate, both *graphic evidence* and *credibility* significantly predicted more negative attitudes towards smoking for the sample as whole. Yet as shown in Figure 4 (panel B), there were individuals, approximately 18% and 64%, respectively, for whom this was not the case. Indeed, as shown in Figure 4 (panel A), at least 11% and 8% of the sample had within-person correlations that were at least 2 *SE* away from the group average in the positive direction. For these individuals, the conclusion at the group level would lead to a false positive (Type I error). Moreover, if a recommendation is made for

TABLE 1 Estimates and *p* values of fixed and random effects of message feature on smoking attitudes

	Fixed effect		Random effect	
	γ	*p*	*SD* of *u*	*p*
Message feature				
Graphic evidence	−4.04	<.001	4.22	<.001
Credibility	−0.93	<.006	1.84	<.001
Effects on others	−0.08	0.759	1.35	<.001

Note. γ coefficients represent the effect of the message feature on attitudes towards smoking for the sample as a whole. The *p* value for the γ coefficients tests the null that the effect of the message feature on attitudes towards smoking is 0. The *SD* of *u* coefficients represents the variability across participants in the effect of the message feature on attitudes towards smoking (i.e., the level 1 slopes). The *p* value for *u* coefficients, based on the likelihood ratio test to assess the significance of a random effect (Hayes, 2006), tests the null hypothesis that the observed variability is due to chance or noise. A *p* value less than .05 indicates that there is greater variability in the slopes than is expected by chance and signals the presence of moderation.

widespread adoption of messages emphasizing these features, it would have overlooked the fact that these features reliably backfired for a subset of participants.

The heterogeneity of effects for the message feature *effect on others* also raised concerns about applying conclusions from group-level analyses to individuals. Averaged across all participants, the message feature *effects on others* had no appreciable effect on attitudes towards smoking. However, as shown in Figure 4 (panel B), when we examined the results for each person, *effects on others* had a significant negative effect for 27 (17%) participants and significant positive effect for 5 (3%) on attitude towards smoking. Thus, here, the conclusion at the group level would lead to a false negative (Type II) error for 17% of our participants, denying a possibly effective intervention.

4.1.1 | Testing for heterogeneity of effects in the absence of an explicit moderator

To be sure, past work has examined the possibility that effects of manipulations vary across participants. But virtually, all of them were limited to examining only variability in effects that can be reliably predicted by an individual difference moderator variable that had been identified *a priori*. A strength of the HRWP approach is that it identifies the presence of significant heterogeneity of effects even though no moderator variables were identified *a priori* (Whitsett & Shoda, 2014). In this way, the HRWP design may alert a researcher that there is significant variability to be investigated or, equally as important, that there is no appreciable variability to be investigated (i.e., the majority of the participants responded in a similar fashion to the construct of interest; e.g., Gaby & Zayas, 2017).

A natural next step in the HRWP approach is to identify the person characteristics that reliably predict individual differences in the effect of a situation feature or manipulation. This aim is practically important for work on psychological targeting to influence behavior (Matz, Kosinski, Nave, & Stillwell, 2017). Moreover, at a theoretical level, identifying the person characteristics serves to understand the nature of person-situation interactionism (Shoda, 1999; Wood & Brumbaugh, 2009). Still, the present approach is useful in providing a signal that the effects differ reliably across individuals (Whitsett & Shoda, 2014), that personal characteristics that moderate the effects are likely to exist, and that it may be worth the effort to search for them.

4.2 | Addressing blind spot 2: When the results reflect the idiosyncrasies of the stimuli

The HRWP design can also help address problems associated with insufficient stimulus sampling. To illustrate, what would have happened if we had conducted a study in which fewer stimuli were used to represent the construct of interest? We simulated the results of such a study using the present data. Specifically, because studies often employ stimuli that represent a dichotomous independent variable (e.g., high vs. low), for the purpose of demonstration, we classified stimuli in this study as either "high" or "low" on *effects on others* using a median split. Next, simulating a within-person design with two conditions, each of which used four stimuli, four anti-smoking messages were randomly sampled from the "low" category, and four messages, from the "high" category. Then, we calculated participants' smoking attitude for the two conditions, by averaging across responses to the four stimuli in each condition. We repeated this simulation, and illustrative examples are shown in Figure 5. When employing all 60 stimuli in the HRWP study, we found no evidence for the role of message characteristic *effect on others* (left panel). However, when sampling was limited to four stimuli per condition, the results sometimes led to the erroneous conclusion that *effects on others* worked, decreasing attitudes towards smoking (middle panel) and, at other times, that it backfired, increasing positive attitudes (right panel).

This simulation prompts the question: How many stimuli are needed to ensure sufficient sampling? The 60 messages in our study were certainly better than 2, 4, or even 8 messages, but to what extent does 60 messages lessen the concern about the results reflecting insufficient stimulus sampling. The answer to that question is beyond the scope of this paper. However, once an HRWP study is conducted, using many stimuli in a given domain, it becomes possible to perform a sensitivity analysis to empirically assess the extent to which the effect of interest depends on the number of stimuli (Whitsett & Shoda, 2014, Figure 6) provide an example

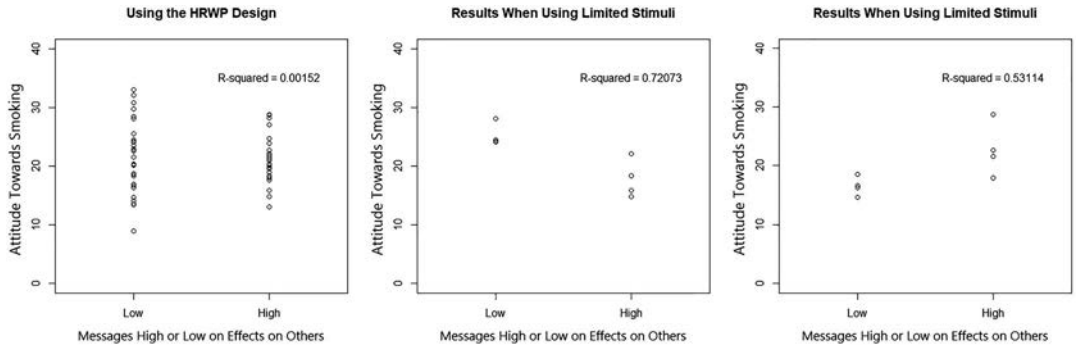

FIGURE 5 The left panel demonstrates that when using large enough and representative samples of stimuli in a study, there is little difference in smoking attitude after viewing anti-smoking messages low versus high in *effects on others*. The middle and right panels demonstrate how erroneous conclusions can be drawn when using smaller and nonrepresentative samples of stimuli in a study

of how to empirically estimate the number of stimuli needed to minimize Type I and Type II errors for a given study question, involving a given stimulus domain and outcome variable.

A second question worth asking is: Even if one ensures sufficient sampling, might a third variable that is correlated with the hypothesized effect be the driver of the effect? In the present study, we used stimuli that were already available, rather than experimentally creating the stimuli. Thus, the present work is a passive design that is susceptible to the limitations of such designs, particularly with regard to causal inference via the presence of a third variable. For example, perhaps, messages that were high in graphic evidence have catchier slogans, and the slogan lead to the message's effectiveness. One approach for investigating threats to causal inference posed by such third variable concerns is to identify stimuli coded as highest and lowest on a given dimension, say *graphic evidence*, and to assess whether these stimuli also vary on another dimension. If the stimuli do differ on other dimensions, these features can be added to the model. Moreover, if researchers make the stimulus set available, then factors can be coded retroactively. They can then be added to the MLM as additional predictors or covariates to evaluate hypotheses about mediating or confounding factors.

4.3 | Implications for examining replication of effects

There is an increasing awareness that it is difficult to replicate many of the published findings in psychological science. Often, a replication study is considered "successful" if the *mean* group-level effect of the replication study is similar, in the direction and magnitude, as the mean group-level effect of the original study. The HRWP approach suggests a possible future direction where researchers could take a more nuanced look at replication. First, the HRWP allows researchers to test the effect of an experimental condition for *each* participant in a study. Thus, the results of an HRWP study could be seen as akin to recent replication projects involving many studies (Klein et al., 2018; Klein et al., 2019; Open Science Collaboration, 2015). But instead of each replication being a study, in the HRWP study, each replication is an individual. In the present illustration, we see that the test of the effect of *graphic evidence* and *credibility* on smoking attitudes was in fact "replicated" for 83% and 36% of the sample, respectively.

Second, because the HRWP approach focuses not only on the mean effect of a manipulation at the group level but also on the distribution of the effect, a replication of a HRWP study could also focus on whether aspects of the *distribution* of effects across participants in the original study replicates in future studies. Suppose that in the original experiment, a positive and statistically significant within-person effect of the experimental manipulation occurred for 30% of the participants, and a negative and significant effect occurred for 10% of the participants. Suppose further

that when averaged across all participants, there is a significant group-level effect. Now, what if in a replication study a positive and significant effect was observed for 20% of the participants, and a negative and significant effect was observed for 20% of the participants? When averaged across all participants, it is very unlikely that there would be an appreciable group-level effect. Did the results of the original study replicate? With regard to the results averaged across participants, the original finding most likely did not replicate. However, in both studies, the manipulation did lead to a clearly positive effect for at least 20% of the sample. Moreover, in this hypothetical scenario, what might not have replicated is the relative proportions of the type of participants with regard to the effects of the experiment (e.g., % for whom there was a strong positive effect) that might reflect differences in sampling or recruitment strategies or populations being sampled. Much more work is certainly needed, but HRWP studies may provide a more nuanced view of the replicability of effects.

Finally, as highlighted by our simulation, a replication study may fail to reproduce the original effects due to differences in the stimuli used. Perhaps the findings of the original study, the replication study, or both reflect the idiosyncrasies of a small number of stimuli used in that study. In contrast, if studies were conducted with a representative and wide variety of stimuli to represent a construct (e.g., message credibility), idiosyncratic effects of each stimulus are less likely to affect the outcome of studies.

4.4 | Implications for increasing sensitivity to participants not well represented in the sample

Effects that are distinctive for a small minority of participants are often invisible to researchers because their responses are averaged out by those of the majority. For example, research that specifically recruited first-generation students found that for them, the effects of some interventions can be the opposite from continuing generation students (e.g., Stephens, Fryberg, Markus, Johnson, & Covarrubias, 2012). Yet because in many studies, first-generation students are a small minority in the sample, the focus on group-level average results would provide little indication that first-generation students respond differently from continuing generation students.

Of course, a fundamental solution to this problem must come from efforts to study populations that have not been well studied in psychology so far. But the HRWP design provides a small step in addressing this problem. First, the approach makes it possible for researchers to become aware of the existence of a minority of individuals for whom a manipulation has reliably different effects compared to the majority of participants. This in turn can provide an impetus for generating hypotheses about how to identify these individuals, and future studies can oversample people with those characteristics.

4.5 | Implications for statements about generalizability across people and stimuli

An implicit assumption among researchers is that findings from a given study generalize (or should generalize) to all people and all stimuli. But instead of assuming the universality of findings across people and settings, researchers should be more specific in stating the expected generality of findings (e.g., Simons, Shoda, & Lindsay, 2017). The HRWP approach allows for an empirical assessment of the generality of findings across people and stimuli to achieve a more nuanced, and deeper, understanding of the phenomena of interest. The knowledge from HRWP approach can help researchers specify for whom and when their manipulation works.

5 | CONCLUSION

The Highly-Repeated Within-Person approach helps mitigate the limitations of commonly used research designs, such as the fact that effects observed at the group average level do not necessarily correspond to the effects at the level of each individual and that conclusions based on a limited sample of stimuli may be specific to those stimuli.

The HRWP approach also provides an important step towards understanding why some studies fail to replicate, in turn facilitating the development of a cumulative science built on a deeper understanding of psychological phenomena.

DATA ACCESSIBILITY STATEMENT

The stimuli, message characteristics, presentation materials, Inquisit code, participant data, syntax for analysis, and analysis outputs can be found on Open Science Framework: https://osf.io/3x59z/.

ENDNOTES

[1] For simplicity, our description focuses on experiments wherein researchers *manipulate* the independent variable. However, the two blind spots also pose limitations for research designs wherein *x* is a naturally-occurring, measured variable.

[2] In some studies, such as studies that involve interacting with a confederate, the word "stimuli" may not be the best for describing the whole social situation involving the interaction with a confederate. That is why, throughout the paper, we also use the term "situations," in addition to stimuli. The HRWP approach can be applied to both types of studies.

[3] Some experiments in social cognition use a large number of stimuli. In these cases, data necessary for the HRWP approach are already collected. However, the HRWP approach calls for a different way to analyze the data such that heterogeneity of effects across individuals can be empirically assessed.

[4] The statistical power of this approach depends on the number of stimuli. Because there were 60 messages, the *N* for each within-person correlations was 60. As the number of stimuli decreases, confidence intervals for a given person's within-person correlation increase. This loss of statistical power makes it more difficult to find results that are statistically significant for a *given* individual. They also make it more difficult to interpret null results.

[5] The *SE* is provided by the formula, $\sqrt{(1-r^2)/(n-2)}$, given that the sampling distributions of correlation coefficients are approximated by a *t* distribution when the sample sizes are not very small and the correlations are not extreme (Kendall & Stuart, 1973, section 31.19; also, see Rahman, 1968).

[6] When this is not the case, for example, in experience sampling studies in which participants encounter different random samples of stimuli, researchers may use a different variant of MLM models such as hierarchical linear models (Raudenbush & Bryk, 2002).

ORCID

Vivian Zayas https://orcid.org/0000-0002-9534-3721

REFERENCES

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171. https://doi.org/10.1177/1745691611400237

Fisher, A. J. (2015). Toward a dynamic model of psychological assessment: Implications for personalized care. *Journal of Consulting and Clinical Psychology*, *83*, 825–836. https://doi.org/10.1037/ccp0000026

Fleeson, W. (2007a). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, *75*, 825–862. https://doi.org/10.1111/j.1467-6494.2007.00458.x

Fleeson, W. (2007b). Using experience sampling and multilevel modeling to study person-situation interactionist approaches to positive psychology. In *Oxford handbook of methods in positive psychology* (pp. 501–514). New York, NY, US: Oxford University Press.

Fleeson, W., Malanos, A. B., & Achille, N. M. (2002). An intraindividual process approach to the relationship between extraversion and positive affect: Is acting extraverted as "good" as being extraverted? *Journal of Personality and Social Psychology*, *83*, 1409–1422. https://doi.org/10.1037/0022-3514.83.6.1409

Gaby, J., & Zayas, V. (2017). Smelling is telling: Human olfactory cues influence social judgments in semi-realistic interactions. *Chemical Senses*, *42*, 405–418. https://doi.org/10.1093/chemse/bjx012

Gallistel, C., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 13124–13131. https://doi.org/10.1073/pnas.0404965101

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32*, 385–410. https://doi.org/10.1111/j.1468-2958.2006.00281.x

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics, volume 2: Inference and relationship*. Griffin. ISBN 0–85264–215-6 (Section 31.19)

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2018, October 2). Investigating variation in replicability: A "many labs" replication project. Retrieved from osf.io/wx7ck

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B. Jr., Alper, S., … Nosek, B. A. (2019, February 11). Many labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.17605/OSF.IO/8CD4R

Lee, J. (2009). The situation and the person: A social-cognitive approach to modeling and predicting people's unique patterns of emotional and behavioral responses to complex social situations, ProQuest Dissertations and Theses.

LeeTiernan, S. (2002). Modeling and predicting stable response variation across situations, ProQuest Dissertations and Theses.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *114*, 12714–12719. https://doi.org/10.1073/pnas.1710966114

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943–950.

Rahman, N. A. (1968). *A course in theoretical statistics* (p. 1968). Charles Griffin and Company.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods. Advanced quantitative techniques in the social sciences 1 (Vol. 2nd).

Shoda, Y. (1999). A unified framework for the study of behavioral consistency: Bridging person-situation interaction and the consistency paradox. *European Journal of Personality*, *13*, 361–387. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<361::AID-PER362>3.0.CO;2-X

Shoda, Y. (2004). Individual differences in social psychology: Understanding situations to understand people, understanding people to understand situations. In C. Sansone, C. Morf, & A. Panter (Eds.), *Handbook of methods in social psychology* (pp. 117–141). Sage.

Shoda, Y., & LeeTiernan, S. (2002). What remains invariant?: Finding order within a person's thoughts, feelings, and behaviors across situations. In D. Cervone, & W. Mischel (Eds.), *Advances in personality science*, 1 (pp. 241–270). New York, NY: Guilford Press.

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, *67*, 674–687. https://doi.org/10.1037/0022-3514.67.4.674

Simons, D., Shoda, Y., & Lindsay, D. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives On Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Sridharan, V. (2015). *Evaluating smoking attitudes in response to different types of anti-smoking messages using a Highly-Repeated Within-Person design (master of science)*. University of Washington.

Stephens, N., Fryberg, S., Markus, H., Johnson, C., & Covarrubias, R. (2012). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology*, *102*(6), 1178–1197. https://doi.org/10.1037/a0027143

Swanson, J. E., Rudman, L. A., & Greenwald, A. G. (2001). Using the implicit association test to investigate attitude-behaviour consistency for stigmatised behaviour. *Cognition and Emotion*, *15*, 207–230. https://doi.org/10.1080/0269993004200060

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass: Addison-Wesley.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science*, *10*, 390–399. https://doi.org/10.1177/1745691614564879

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020–2045. https://doi.org/10.1037/xge0000014

Whitsett, D. D., & Shoda, Y. (2014). An approach to test for individual differences in the effects of situations without using moderator variables. *Journal of Experimental Social Psychology*, *50*, 94–104. https://doi.org/10.1016/j.jesp.2013.08.008

Wilson, N. (2008). Identifying the features of stressful situations, ProQuest Dissertations and Theses.

Wood, D., & Brumbaugh, C. C. (2009). Using revealed mate preferences to evaluate market force and differential preference explanations for mate selection. *Journal of Personality and Social Psychology*, *96*, 1226–1244. https://doi.org/10.1037/a0015300

Zayas, V., & Shoda, Y. (2007). Predicting preferences for dating partners from past experiences of psychological abuse: Identifying the psychological ingredients of situations. *Personality and Social Psychology Bulletin*, *33*, 123–138. https://doi.org/10.1177/0146167206293493

Zayas, V., Whitsett, D. D., Lee, J. J. Y., Wilson, N., & Shoda, Y. (2008). From situation assessment to personality: Building a social cognitive model of a person. In *The sage handbook of personality theory and assessment* (pp. 375–401). Thousand Oaks, CA: Sage Publishing.

AUTHOR BIOGRAPHIES

Vivian Zayas received her PhD from the University of Washington, and is an Associate Professor of Psychology at Cornell University. Her research focuses on the 'relational mind' examining questions such as: how do we mentally represent the emotional complexity of our closest relationships? Why do we "click" with some people and not with others? Much of her work focuses on processes that operate nonconsciously (i.e., outside of conscious awareness). Her research appears in journals such as *Psychological Science*, *Proceedings of the National Academy of Sciences*, *Nature Communications*, *Journal of Personality and Social Psychology*, *Personality and Social Psychology Bulletin*, *Social Psychological and Personality Science*, *Child Development*, and *Journal of Personality*, *as well as in the popular press*, *such as the New York Times*, *Quartz*, *Newsweek*, *Discover Magazine*, *and Psychology Today*. Her research has received funding from National Science Foundation and National Institutes of Health.

Vasundhara Sridharan received her PhD in Social and Personality Psychology from the University of Washington, and is a Senior Researcher working in the digital health industry. Her research focuses on designing digital evidence-based psychological interventions for a variety of health behaviors such as smoking and self-management of chronic diseases.

Randy T. Lee is a PhD student in the Social and Personality area in the Department of Psychology at Cornell University. He is interested in three distinct, but interrelated topics: (a) emotion and emotion theory; (b) the deliberate and automatic ways we intrapsychically and interpersonally regulate emotion; and (c) the causes, responses, and outcomes of and to social exclusion at the intrapsychic, interpersonal, communal, and institutional levels.

Yuichi Shoda received his PhD from Columbia University, and is a Professor of Psychology at the University of Washington. His current work includes "quantitative idiography" and further development of the Highly-Repeated Within-Person approach for detecting individual differences in the effects of specific situations without relying on moderator variables.

How to cite this article: Zayas V, Sridharan V, Lee RT, Shoda Y. Addressing two blind spots of commonly used experimental designs: The Highly-Repeated Within-Person approach. *Soc Personal Psychol Compass*. 2019;13:e12487. https://doi.org/10.1111/spc3.12487

APPENDIX

TABLE A1 Loadings of different message characteristics on three factors or *psychologically active ingredients* of anti-smoking messages

Item	Graphic evidence	Credibility	Effects on others	Mean	*SD*	Alpha
The message shows the future consequences of smoking	.929	.171	.099	4.80	1.43	.87
This message talks about the negative consequences of smoking	.918	.019	.052	5.46	1.55	.87
The message shows that smoking is as bad or worse than using other drugs	.879	.025	−.085	4.22	1.04	.74
The message provides evidence for the negative consequences of smoking	.871	.139	.149	5.21	1.07	.81
The message shows that smokers will regret taking up smoking	.858	.167	.183	4.63	1.39	.85
The message shows that the damage from smoking is permanent	.854	.153	−.047	4.70	1.26	.84
The message is shocking	.809	.221	.002	4.37	1.10	.79
It presents facts related to smoking	.736	−.125	.161	4.73	0.80	.76
The message emphasizes the need to quit smoking immediately	.733	.265	.174	4.96	1.07	.85
There is an emphasis on how certain part(s) of the body is affected by smoking	.667	.510	−.242	4.08	1.82	.92
The message depicts a medical setting (e.g., a hospital)	.637	.590	.053	2.41	1.57	.91
Statistics related to smoking behavior are shown	.583	−.235	.198	3.55	1.20	.69
The message is from an official source	−.004	.914	.112	4.23	1.31	.64
The message shows real people who are affected by smoking (as opposed to actors or cartoons)	.172	.909	.005	4.09	2.11	.93
The message has a cartoon-like drawing	−.162	−.908	−.013	3.25	1.63	.89
It contains information about specific actions that people can take to quit smoking	.181	.840	.019	2.45	1.24	.86
The message looks fake (e.g., it looks photoshopped)	−.206	−.691	.033	3.7	1.48	.79
The message shows that smoking can lead to the death of children	.166	−.035	.840	1.98	1.38	.90
The message shows that cigarette smoke can be harmful to nonsmokers who are exposed to it	.139	.254	.838	2.93	1.86	.88
The message shows that smoking by pregnant women causes prenatal harm	.135	−.071	.809	1.90	1.09	.85
The message shows the harmful effects of secondhand smoke on others	−.020	.348	.791	2.33	1.94	.95
It is only relevant to people with children	−.042	−.295	.630	2.10	1.59	.92

Note. The number of raters for each item is six.