Moving From the Lab to the Field: The Role of Fidelity and Achieved Relative Intervention Strength

Chris S. Hulleman & David S. Cordray

Vanderbilt University, Nashville, Tennessee, USA

Published online: 14 Jan 2009.

To cite this article: Chris S. Hulleman & David S. Cordray (2009) Moving From the Lab to the Field: The Role of Fidelity and Achieved Relative Intervention Strength, Journal of Research on Educational Effectiveness, 2:1, 88-110, DOI: 10.1080/19345740802539325

To link to this article: http://dx.doi.org/10.1080/19345740802539325

Please scroll down for article

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.
METHODOLOGICAL STUDIES

Moving From the Lab to the Field: The Role of Fidelity and Achieved Relative Intervention Strength

Chris S. Hulleman and David S. Cordray
Vanderbilt University, Nashville, Tennessee, USA

Abstract: As an example of how the effectiveness of well-designed laboratory interventions is often diffused in the field, we examined the effects of a motivation intervention on students’ perceptions and learning. The intervention proved to be more effective in the laboratory (g = 0.45) than the field (g = 0.05) in enhancing subsequent motivation. We explored this reduction in treatment effectiveness through a fidelity analysis that examined the extent to which participants responded to the treatment. We calculated fidelity as three indices of achieved relative treatment strength (Cordray & Pion, 2006), and found that, regardless of how fidelity was calculated, achieved relative strength was about 1 standard deviation less in the classroom than the laboratory. In addition, greater levels of achieved relative strength were associated with greater differences in the motivational outcome—indicating that intervention was more effective for participants who actually received the treatment than those who did not. Multilevel analyses indicated that the drop in classroom treatment fidelity was partially because of teacher, rather than student, factors. Implications for theoretical models of change and research are discussed.

Keywords: Treatment fidelity, intervention strength, treatment effectiveness, multilevel modeling, student motivation

Ideally, educational interventions should be grounded in theory and research about basic processes on learning, motivation, and instruction. Most often, studies on these basic processes are undertaken in controlled settings (i.e., the research laboratory), which facilitates some aspects of causal inference (i.e., internal validity) by ruling out extraneous and confounding influences on the outcome. However, a major concern is how well the laboratory-based
results will replicate in actual educational settings. When the laboratory environment is too scripted and controlled, it is reasonable to question how effective the intervention will be when confronted with the relatively uncontrolled world of classrooms and schools. Invariably, the classroom intervention is not an exact replica of the one designed in the laboratory because the treatment is often not implemented with complete fidelity. For example, not every teacher may attend all of the professional development training sessions for a new curriculum, and thus not be able to implement the curriculum as designed. In addition, there is often a tension between maintaining perfect intervention fidelity and adapting components of the intervention to match the educational context. Teachers may alter portions of the curriculum to better match their students’ needs and therefore enhance its effectiveness, or they may change portions that require too much advanced preparation and therefore undermine its effectiveness (Gottfredsen, 1984).

Regardless of how these fidelity and adaptation processes alter the implementation of the intervention in the field, the overall effectiveness of the intervention is primarily assessed by asking the question, Was the intervention effective as implemented? This intent-to-treat analysis assesses whether the intervention—as it was implemented—was effective in promoting the desired outcomes. Here, the causal question is whether the resources invested in the intervention—money, time, effort—made a difference, on average. However, answering secondary questions (e.g., how were effects produced or not?) is also important for policymakers and practitioners who intend to improve the intervention based on experimental tests of effectiveness. How can researchers and practitioners ensure that an effective intervention, designed and tested in a controlled environment, retains its effectiveness in the field? What can be learned from analyzing why an intervention failed to be as effective in the field as it was in the laboratory? This article focuses on issues of fidelity in the process of transferring laboratory interventions to the field.

INTERVENTION FIDELITY AND RELATIVE STRENGTH

The notion of intervention fidelity has been captured under a broad array of labels, such as treatment integrity, adherence, compliance, dose, exposure, quality of delivery, and treatment differentiation (for reviews, see Dane & Schneider, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003; Mowbray, Holter, Teague, & Bybee, 2003; O’Donnell, 2008). Treatment integrity, compliance, and adherence refer to the extent to which participants (e.g., teachers) deliver the intended innovation and whether other participants (e.g., students) accept or receive or are responsive to the intended services, at the intended level of treatment strength (see Boruch & Gomez, 1977; Cordray & Pion, 2006; Yeaton & Sechrest, 1981). In practice, these constructs are often operationalized by indices of dose, exposure, and quality.
Assessments of intervention fidelity involve the specification of a “gold standard” or basis for comparison—a theory, model, or conception of the educational intervention—to which something is faithful. As such, fidelity assessments must begin with a full characterization of the intervention “in theory.” These theories can be grand or small. They have in common a well-stated set of expectations about how the intervention is supposed to work, its underlying logic, and rationales for how and why these actions will produce the desired enhancements in student learning, motivation, and achievement. Fidelity assessments tell us how closely the intervention, in practice, met these specifications. Reliable and valid measures of achieved intervention fidelity index the degree of discrepancy between what should have been implemented and what was actually implemented.

Figure 1 defines intervention fidelity within the context of a randomized control trial. Cordray and Pion (2006) stated that the outcome \( Y_i \) for a participant is determined by the achieved fidelity of the treatment, as implemented and received by that individual \( t_{TX} \). When it is possible to stipulate the intended or theoretical strength of an intervention \( T^{TX} \), true indices of compliance, adherence, or treatment integrity can be derived. As such, the achieved intervention fidelity

\[
\begin{align*}
T^{TX} & \quad \text{Integrity} \\
& \quad \text{Achieved} \\
& \quad \text{Outcome} \\
0.45 & \quad 100 \\
0.40 & \quad 90 \\
0.35 & \quad 85 \\
0.30 & \quad 80 \\
0.25 & \quad 75 \\
0.20 & \quad 70 \\
0.15 & \quad 65 \\
0.10 & \quad 60 \\
0.05 & \quad 55 \\
0.00 & \quad 50 \\
\end{align*}
\]

\[
\begin{align*}
\frac{Y_i - \bar{Y}}{sd_{pooled}} & = d \\
\text{Achieved Relative Strength} = 15 \\
d & = \frac{85 - 70}{30} = 0.50 \\
\text{Expected Relative Strength} = 25 \\
d & = \frac{50}{30} = 0.83 \\
\end{align*}
\]

**Figure 1.** Representing fidelity and relative strength in experiments. Adapted from Cordray and Pion (2006, p. 116).
fidelity (or treatment integrity) can be represented as the difference between treatment as theorized ($T^{Tx}$) and the treatment as realized for individuals or groups of individuals ($t^{Tx}$). In Figure 1, the degree of treatment infidelity, $T^{Tx} - t^{Tx}$, across all participants, is .05 “strength” units ($0.40 - 0.35 = 0.05$).

Cordray and Pion (2006) also incorporated treatment differentiation (see Waltz, Addis, Koerner, & Jacobson, 1993) into their definition of intervention fidelity. Treatment differentiation suggests that $T^{Tx}$ has to be stronger than or different from the counterfactual condition. Counterfactual conditions rarely are unorganized collections of activities. Rather, as is often the case in education research, control conditions frequently consist of “business as usual” in terms of curriculum activities. Holland (1986) stipulated that theories, or models of causality, are embedded in control conditions. These theories/models can be designated as $T^C$. As such, the causal effect—on the outcome $E(Y^{Tx})$—of the target treatment ($T^{Tx}$) has to be considered relative to the causal components in $T^C$ associated with the production of the outcome—$E(Y^C)$—in the counterfactual condition. As previously stated, infidelity can occur when the actual comparison condition ($t^C$, as opposed to its theoretical counterpart, $T^C$) becomes more like the $T^{Tx}$. This migration can be due to contamination or leakage of the $t^C$ with elements of the $T^{Tx}$ (Orwin, Sonnefeld, Cordray, Pion, & Perl, 1998; Shadish, Cook, & Campbell, 2002), such as when core components of the treatment are provided to participants in the control condition. To determine if this happens, a parallel “fidelity” assessment of programmatic components in counterfactual conditions also is required. Cordray and Jacobs (2005) linked fidelity assessment to contemporary statistical models of causal inference and refer to the difference between intervention and control conditions as the achieved relative strength of the contrast. More specifically, the achieved relative strength is the difference between the treatment, as implemented, and the control, as implemented (i.e., $t^{Tx} - t^C$). In turn, the estimates of effects on the outcome are the result of the achieved relative strength of the contrast ($t^{Tx} - t^C$), not the theoretically expected difference ($T^{Tx} - T^C$). Because of these sources of infidelity, the observed effects can be less than originally expected. For the hypothetical example in Figure 1, the expected effect of a perfectly implemented intervention, based on $T^{Tx} - T^C$, is $d = 0.83$ standard deviation units. On the other hand, because of infidelity in both conditions, the achieved effect ($t^{Tx} - t^C$) on the outcome is only $d = 0.50$.

In this article, we operationalized the Cordray and Pion (2006) conception of intervention fidelity and achieved relative strength in a pair of experimental studies, one laboratory based and the other field based. These studies were designed to examine the effects of a motivation intervention on students’ perceptions and learning. In addition to documenting the achieved relative treatment strength for each intervention site (lab and classrooms), we investigated the effects of differences in achieved relative strength between sites on outcomes. In the final section of the article we examined sources of infidelity in classroom settings to propose remedial actions in future field studies.
AN EXAMPLE: A MOTIVATIONAL INTERVENTION

Hulleman and his colleagues (Hulleman, 2008; Hulleman, Godes, Hendricks, & Harackiewicz, 2008; Hulleman & Harackiewicz, 2008; Hulleman, Hendricks, & Harackiewicz, 2007) have developed and tested an instructional intervention designed to increase student motivation and performance. The intervention manipulates the relevance and usefulness of the course material (i.e., utility value) by asking students to make connections between the topic they are studying and their lives. Expectancy-value models of motivation (Eccles et al., 1983) suggest that the desire to engage in an activity is dependent on both the individual’s expectancies for success in the activity and their perceived importance and value. In particular, prior research demonstrates that utility value is correlated with both classroom interest and achievement (Hulleman, Durik, Schweigert, & Harackiewicz, 2008; Malka & Covington, 2005; Simons, Dewitte, & Lens, 2003). In the Hulleman and colleagues intervention, participants in the relevance condition were asked to select a topic that they had been studying and write about how it was related to their life in some way. Participants in the control condition were asked to select a topic they had been studying and to write a summary of what they had learned. By manipulating relevance, Hulleman and colleagues hypothesized that the subsequent increase in utility value would lead to more interest in the material and higher grades.

In this article we utilized data collected in two randomized experiments that have been more fully reported elsewhere (Hulleman, 2008; Hulleman, Godes, et al., 2008; Hulleman & Harackiewicz, 2008; Hulleman et al., 2007). First, a laboratory study was conducted to develop and test the effectiveness of the intervention. Second, a field experiment in high school science classrooms was conducted to assess the effectiveness of the intervention in a real-world context. This article focuses on the relationship between implementation fidelity and treatment effectiveness, which was not considered in the earlier publications.

In the first randomized experiment, laboratory participants were 107 undergraduate students who learned a new method of solving multiplication problems, and randomly assigned to either the relevance or control conditions. Participants in the relevance condition wrote about how useful the new technique could be to their lives, and those in the control condition wrote a summary of the math technique. The second study was a randomized field experiment in which 182 high school students from 13 science classrooms taught by eight teachers across three different high schools who were randomly assigned to either write about the usefulness and utility value of the course material in their lives (relevance condition) or write a summary of the material they were studying (control condition). Classroom teachers were asked to assign the writing intervention to students as part of their exam review activities. Because of variability in syllabi and course planning, the number of writing assignments offered to students during the semester ranged from two to eight (M = 5.32, SD
In both studies, following the intervention participants reported how useful and relevant they found the math activity or the science course material to their lives.

In the laboratory, the intervention produced a positive effect on perceptions of utility value, $t(108) = 2.36, p = .02, g = 0.45$, such that participants in the treatment group perceived more utility value in the math task ($M = 5.28, SD = 0.94$) compared to those in the control group ($M = 4.78, SD = 1.28$). In the classroom, the effect was nearly zero, $t(180) = 0.43, p = .67, g = 0.05$, such that participants in the treatment group perceived only slightly more utility value in the course material ($M = 3.62, SD = 0.95$) compared to those in the control group ($M = 3.56, SD = 0.92$).

**FIDELITY MEASUREMENT AND ACHIEVED RELATIVE STRENGTH**

To explore why the treatment was less effective in the classroom, we examined treatment fidelity in both experiments by assessing the extent to which the intervention was implemented in treatment and control conditions. As presented in Figure 1, the actual difference in treatment fidelity between treatment and control conditions ($t^{Tx} - t^{C}$) allows us to operationalize the achieved relative strength of the intervention. However, before we could calculate achieved relative strength, we first needed to determine how to assess fidelity. This required specifying which intervention components were essential to the program’s effectiveness (i.e., core components) and which were less essential. Theoretically, the degree of relevance that participants perceived in the math activity or science course material was identified as the single core component, or “active ingredient,” of the intervention. That is, if participants did not comply with the instructions to make connections between themselves and the material, then the psychological mechanism behind the construct of perceived utility value would not be triggered. Indices of the strength of the intervention were dose and participant responsiveness.$^2$

---

1This effect size is calculated using Hedges’s $g$ and corrected for clustering in the classroom (Hedges, 2007). See the appendix for formulas.

2Implementation fidelity, as we have defined it, is similar to a manipulation check in experimental psychology. The difference between a manipulation check and a mediational mechanism is not always large and can often be a matter of timing as well as definition. In the case of the motivational intervention, participant responsiveness refers to whether participants responded to the intervention instructions. Thus, the essays are the check that participants followed the instructions in each experimental condition. The actual mediating process of the intervention is subsequent perceptions of utility value and is theorized to be activated by the manipulation of relevance, which is theorized to increase motivation to engage in the activity.
Dose was operationalized as the number of opportunities each participant had to write about the relevance of the math activity to their lives (treatment) or a topic summary (control). Because the study design called for a single dose to be delivered in the laboratory and multiple doses to be delivered in the classroom, the dose index is relevant only to the classroom condition. Participant responsiveness was operationalized in two ways: (a) the frequency of responding to the instructions to write an essay, and (b) the quality of response in the essay. Frequency was assessed by summing the number of essays that students wrote—-independent of the quality of the essay. That is, as long as participants wrote something in their essays, this was considered an observable instance for frequency of responsiveness. In contrast, quality of responsiveness was assessed as the extent to which participants connected the math activity or science topic to their lives in their essays. Two independent raters, blind to condition, read and coded each essay for quantity and quality of connections on a 4-point scale: 0 (no connections), 1 (weak connections), 2 (moderate connections), and 3 (strong connections). Raters demonstrated 81% agreement on the laboratory essays and 88% agreement on the classroom essays, and differences were resolved through discussion.

As reported in Table 1, all participants in the laboratory study received one opportunity to write about the relevance of the math activity to their lives, whereas there was considerable variability in the number of opportunities (dosage) across classrooms. Because of the designed invariance in laboratory dosage, this measure of fidelity could not be utilized in explaining the differences in fidelity between the lab and classroom. In contrast, participant responsiveness was variable in both the laboratory and classroom and thus it was the focus of the initial fidelity analyses. As presented in Table 2,

### Table 1. Levels of treatment dosage by study and condition

<table>
<thead>
<tr>
<th>Dosage</th>
<th>Laboratory</th>
<th></th>
<th></th>
<th>Classroom</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>Tx</td>
<td>C</td>
<td>Tx</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1</td>
<td>47</td>
<td>100</td>
<td>63</td>
<td>100</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>43</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>100</td>
<td>63</td>
<td>100</td>
<td>90</td>
<td>100</td>
</tr>
</tbody>
</table>

| M      | 0.00 | 1.00 | 3.17 | 3.14 |
| SD     | 0.00 | 0.00 | 1.05 | 1.01 |

*Note. C = control; Tx = treatment.*
Table 2. Frequency of participant responsiveness by study and condition

<table>
<thead>
<tr>
<th>Response Frequency</th>
<th>Laboratory</th>
<th></th>
<th>Classroom</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>Tx</td>
<td>C</td>
<td>Tx</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>47</td>
<td>100</td>
<td>63</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>100</td>
<td>63</td>
<td>100</td>
</tr>
</tbody>
</table>

*M* 0.00 1.00 2.51 2.49

*SD* 0.00 0.00 1.12 1.22

*Note.* Response frequency refers to the number of essays that participants wrote during the experiment (during either one experimental session or one semester). C = control; Tx = treatment.

the controlled conditions of the laboratory prevented frequency of participant responsiveness to vary across treatment and control conditions. In contrast, Table 3 demonstrates there were differences in quality of responsiveness across the lab and classroom studies. Whereas 89% of participants in the laboratory treatment condition made at least a weak connection, only 59% of those in the

Table 3. Quality of participant responsiveness by study and condition

<table>
<thead>
<tr>
<th>Quality of Responsiveness</th>
<th>Laboratory</th>
<th></th>
<th>Classroom</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>Tx</td>
<td>C</td>
<td>Tx</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>47</td>
<td>100</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>46</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>100</td>
<td>63</td>
<td>100</td>
</tr>
</tbody>
</table>

*M* 0.00 1.73 0.04 0.74

*SD* 0.00 0.90 0.21 0.71

*Note.* Quality of responsiveness refers to the extent to which participants made connections between the topic and their lives, and could range from 0 (*no connections*), 1 (*weak connections*), 2 (*moderate connections*), to 3 (*strong connections*). C = control; Tx = treatment.
classroom did so. In other words, nearly half (41%) of the participants in the classroom treatment conditions failed to make connections with the science topic, whereas only 11% of treatment participants in the laboratory failed to make a connection with the new math technique.

INDEXING FIDELITY

The data on dosage, frequency of responsiveness, and the quality of participants’ responsiveness can be reformulated as indices of intervention fidelity and relative strength. For the purposes of this discussion we focused on one source of data, namely, the quality of the participants’ responsiveness. We compared three different methods in calculating a treatment fidelity index: an absolute fidelity index, an average index, and a binary complier index.

Absolute Fidelity Index

The first method involved determining whether there was an absolute, or maximum, level of fidelity from which to compare participants’ responses. This method most closely matches the conceptualization of fidelity and achieved relative strength presented in Figure 1: The theoretical maximum value is $T_{T_x}$, and the observed participant response is $t_{T_x}$. However, this method is difficult to operationalize because we rarely know the maximum value. Either prior theory or empirical results may be used in determining the absolute level of participant responsiveness from which to compare current levels. For the motivational intervention, we used the maximum possible response quality score from the essay coding system, which was three ($T_{T_x}$). The absolute fidelity index was calculated by dividing the mean value in response quality for each condition by this hypothetical maximum value (treatment: $t_{T_x}$, control: $t_c$). This proportion was multiplied by 100 to produce the percentage of absolute fidelity (i.e., the absolute fidelity index). The percentage of absolute fidelity attained in treatment conditions was greater in the laboratory ($t_{T_x} = 1.73$ out of 3; 58%) than in the classroom ($t_{T_x} = 0.74$ out of 3; 25%), whereas there were equally low levels of absolute fidelity in control conditions in both the laboratory ($t_C = 0$%) and classroom ($t_C = 4$%).

Average Fidelity Index

A second method considered mean levels of response quality in each condition as an indicator of treatment receipt (i.e., the average fidelity index). This method is perhaps more intuitive than the absolute fidelity index and matches the more common approach of using means and standard deviations in statistical analyses. Of importance, this method does not require specifying an
absolute threshold of fidelity, as does the absolute fidelity index, and can thus be generated from the observed fidelity ratings/scores. As such, it indexes achieved fidelity for each condition and setting. For each condition, the index corresponds to $t^T_x$ and $t^C$ in Figure 1. As presented in Table 3, average levels of fidelity in the treatment conditions were higher in the laboratory ($t^T_x = 1.73$) than the classroom ($t^C = 0.74$), whereas average levels of fidelity in the control conditions were equally low in both the lab ($t^T_x = 0.00$) and classroom ($t^C = 0.04$).

**Binary Complier Index**

A third method created a binary variable to indicate whether participants had received the treatment (or not). This is referred to as the binary complier index. This approach is often utilized to estimate the effects of an intervention for only those treatment participants who actually received the intervention, that is, complier average causal effect estimation (e.g., Angrist, Imbens, & Rubin, 1996; Bloom, 1984; Jo, Asparouhov, Muthen, Ialongo, & Brown, 2008). The binary index requires the specification of a minimum cut-point designating a level of responsiveness that is judged to be sufficient to warrant a claim that a participant received full exposure. This cut-point can be determined a priori, either by using a theoretically or empirically derived value or post hoc by examining the distribution of participant responsiveness. In our pair of studies, we used the laboratory distribution of quality of responsiveness to determine the cutoff score for the classroom. An inspection of Table 3 revealed that 65% of treatment participants in the laboratory study made moderate or strong connections in their essays. These participants were coded as having received the full exposure of treatment. The remaining participants made weak (24%) or no connections (11%) and were coded as receiving less than a full dose of treatment. In the classroom, only 15% of treatment participants made moderate connections and were coded as receiving a full dose (no participants made strong connections), whereas the majority of participants made either weak (43%) or no connections (41%) and were coded as receiving less than a full dose of treatment.

Relating this index back to Figure 1, the theoretical model specified that all treatment participants would receive the treatment ($T^T_x = 1.0$), whereas no control participants would ($T^C = 0.0$). In actuality, only 65% of laboratory treatment participants ($t^T_x = 0.65$) and 15% of classroom treatment participants received the treatment ($t^T_x = 0.15$), whereas 0% of control participants in both studies received the treatment ($t^C = 0$). As calculated in Figure 1, the degree of laboratory infidelity was 0.35 for the treatment ($1 - t^T_x = 1 - 0.65$) and zero for the control ($0 + t^C = 0 + 0$), whereas the degree of classroom infidelity was 0.85 for the treatment ($1 - t^T_x = 1 - 0.15$) and zero for the control ($0 + t^C = 0 + 0$).
INDEXING ACHIEVED RELATIVE STRENGTH

Within Study Contrasts

Now that we have operationalized treatment fidelity, we can utilize these measures to create indices of achieved relative strength. By itself, treatment fidelity does not help us distinguish the strength of the intervention—it is only in comparison to the control condition that we can determine the achieved relative strength of the intervention. Therefore, we quantified achieved relative strength as a standardized difference in the fidelity index across treatment and control conditions. As with conventional effect sizes, the effect size measure of achieved relative strength is expressed in standard deviation units. The Achieved Relative Strength (ARS) Index for each of these three conceptualizations of participant responsiveness is based on Hedges’s $g$ and corrected for clustering in the classroom (Hedges, 2007).

Conceptually, the ARS Index is calculated as the difference between achieved fidelity in the treatment and control groups divided by their pooled standard deviation (i.e., $\frac{t_{Tx} - t_{C}}{\sqrt{\frac{1}{n_{Tx}} + \frac{1}{n_{C}}}}$). Because both the absolute fidelity index and the binary complier index are based on proportions, their equations differ from the average fidelity index which is based on means. The appendix presents the equations for each type of ARS Index. For example, we calculated the ARS Index for absolute fidelity in the lab to be $g = 1.71$ by inserting the following values into Equation A3: $t_{Tx} = 0.58$, $t_{C} = 0.00$, $n_{Tx} = 63$, and $n_{C} = 47$. In the classroom, the ARS Index for absolute fidelity was calculated as $g = 0.81$ using Equation A4 to correct for clustering (intraclass correlation [ICC] = 0.08, $t_{Tx} = 0.25$, $t_{C} = 0.04$, $n_{Tx} = 92$, and $n_{C} = 90$). Table 4 presents the ARS Indices for each type of fidelity index in both the lab and the field.

Lab Versus Field Contrasts

Once we have created the ARS Indices for each study, we can then compare the achieved relative strength of the intervention for each of the three types of
fidelity (absolute, average, binary). As presented in last column of Table 4, all three measures of achieved relative strength indicated that participants were more responsive to the treatment in the lab than the classroom. In particular, the difference in the ARS Indices between the laboratory and classroom were all near 1.0 (Absolute = 0.91, Average = 1.20, Binary = 1.08), indicating that the three indicators of achieved relative strength all revealed approximately a 1 standard deviation decrement in treatment strength when moving from the laboratory to the field.

**LINKING ACHIEVED RELATIVE STRENGTH TO OUTCOMES**

Not only do these three methods of conceptualizing quality of participant responsiveness portray the achieved relative strength of the intervention across studies, but differences in achieved relative strength were concomitant with differences in treatment effectiveness. Figure 2 presents the results for the binary ARS Index where treatment participants were classified as either receiving a full dose of the treatment (or not). In comparison to the laboratory control group ($M = 4.78, SD = 1.28$), participants in the treatment condition who received a full dose of the treatment perceived more utility value in the math activity ($M = 5.64, SD = 0.73, g = 0.81$), whereas those who did not receive a full dose actually perceived slightly less utility value in the math activity ($M = 4.63, SD = 0.96, g = −0.12$). In comparison to the classroom control

![Figure 2](image.png)

**Figure 2.** Motivational outcome effect sizes at different levels of treatment receipt. *Note.* The error bars represent the 95% confidence interval for each effect size (see the appendix for equations).
group \((M = 3.59, SD = 0.89)\), participants in the treatment condition who received a full dose of the treatment perceived more utility value in the math activity \((M = 3.87, SD = 0.86, g = 0.32)\), whereas those who did not receive a full dose perceived only slightly more utility value in the math activity \((M = 3.60, SD = 0.93, g = 0.01)\). Not surprisingly, the tightly controlled laboratory study produced a statistically reliable treatment effect. In contrast, although the classroom treatment effect was greater than zero, this difference was not statistically reliable (i.e., the 95% confidence interval included zero). Therefore, it appears that fidelity does matter but only reliably so in the laboratory.

**SOURCES OF INFIDELITY IN THE CLASSROOM**

Given the reduction in fidelity and achieved relative strength in the transition from our lab to our classroom studies, we next consider how fidelity and relative strength can be improved in field settings. In particular, we focus on the processes and structures associated with the classroom interventions that may be responsible for the variability in treatment fidelity. In other words, we used student and teacher characteristics to predict our index of treatment fidelity—namely, the quality of responsiveness. Given the degree of variability in treatment infidelity, and the nested structure of the data, the classroom experiment provided the opportunity to investigate potential sources of treatment infidelity. First, we examined the influence of teachers/class and students on the quality of responsiveness. We then examined how dose and frequency of responsiveness affected the quality of responsiveness. The results of these analyses should provide suggestions on how to control sources of infidelity and, in turn, enhance student outcomes.

In identifying the sources of implementation infidelity for the Hulleman et al. motivational intervention, there are several possible places for the treatment implementation to breakdown in the classroom. First, teachers were responsible for providing the opportunity for students to complete the essays. If teachers failed to provide in-class time for the essays to be written, then students would not have had the opportunity to receive the treatment. Thus, teachers were responsible for determining the dosage of the intervention (see Table 1). Second, students were responsible for completing the essays when offered by the teachers. Although students would fail to receive points toward their final course grade if they did not complete the essay, the choice to complete (or not) the essay was theirs to make. Thus, students were responsible for the frequency of participant responsiveness to the treatment (see Table 2). It is important to remember that student responsiveness is contingent on teacher dosage; that is, students were unable to respond to the treatment unless teachers provided the opportunity. Thus, participant responsiveness was nested within teacher dosage. We can use these implementation factors (teacher dosage, student response frequency) to predict quality of participant responsiveness (i.e.,
the degree to which participants made connections with the material in their essays).

Hierarchical Linear Modeling

Given the nested structure of the data, hierarchical linear modeling (HLM) is an appropriate method to analyze the data (Raudenbush & Bryk, 2002). HLM estimates a statistical model that takes into account the interdependencies in the data—between students in the same classroom, between classrooms taught by the same teacher, and between teachers within the same school—by estimating within and between component variance. By allowing the parameters of the model to vary randomly across levels (i.e., the treatment effect can vary randomly across students and teachers), the HLM model estimates the amount of variability in treatment fidelity that resides between students, classrooms, teachers, and schools. In other words, HLM estimates the amount of “leftover” variability in quality of participant responsiveness after accounting for the treatment effect. We can then use response frequency and teacher dosage to predict the residual variation in quality of responsiveness across students and classrooms. If either factor reduces the amount of residual variance in response quality, then we can attribute the reduction to either students or teachers. This is one method of understanding which sources are responsible for implementation failures.

Baseline Analyses

We used the program HLM 6.04 (Raudenbush, Bryk, & Congdon, 2007) to analyze a two-level model. Because the HLM program does not allow for missing data, we used only those cases that had complete data on both response quality and the motivational outcome (perceived utility value). This produced a sample size of 153 students, 9 classrooms, six teachers, and two schools (compared with our prior analyses which contained 183 students, 13 classrooms, eight teachers, and three schools). A dummy code representing the treatment effect (0 = control, 1 = treatment) was entered at Level 1. We examined a random effects model that allowed all residual variances to vary randomly across levels. Preliminary testing using a three-level model (Level 1: students; Level 2: classrooms; Level 3: teachers and school dummy code) indicated that the two high schools and nine classrooms did not account for a significant amount of

3Although students were nested within classrooms, teachers, and schools, the intraclass correlation (ICC) for response quality was quite low (0.077), thus indicating that the clustering of students would have little impact on the results. The ICC for the motivational outcome, perceived utility value, was even smaller (0.01).
variance. Therefore, we analyzed a two-level model predicting response quality with students at Level 1 (Equation 1) and teachers at Level 2 (Equations 2 and 3). Residual variance in response quality was represented by $r_{ij}$ (residual student-level variance), $u_{0j}$ (residual variance in the mean level of response quality across teachers), and $u_{10j}$ (residual variance in the treatment slope across teachers). Although not presented in Equations 2 and 3, dummy codes for classrooms and schools were included in Level 2 to ensure that the standard errors were not biased (Raudenbush & Bryk, 2002).

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{TREATMENT})_{ij} + r_{ij},$$

$$\beta_{0j} = \gamma_{00} + u_{0j},$$

$$\beta_{1j} = \gamma_{10} + u_{10j}$$

The baseline analyses revealed a significant treatment effect on response quality ($\beta_1 = 0.59$, $p < .01$, $g = 1.08$), indicating that students in the treatment condition made more connections between the science topic they chose and their lives ($M = 0.69$, $SD = 0.67$) than those in the control condition ($M = 0.05$, $SD = 0.20$). As indicated in Table 5 (“Baseline Model”), the residual variance components analysis revealed that 52% of the unexplained variability in response quality was because of student factors ($r_{ij}$), and 48% was because of teacher factors ($u_{0j}$, and $u_{10j}$).

**Explanatory Analyses**

Next, we analyzed the effects of student response frequency and teacher dosage on quality of responsiveness. The number of essays that students completed (response frequency) was entered as a Level 1 predictor of response quality (Equation 4). Because response frequency was nested within teacher dosage, teacher dosage was entered as a Level-2 predictor of both the treatment slope and response frequency (see Equations 6 and 7). The mean level of

<table>
<thead>
<tr>
<th>Residual Variance Component</th>
<th>Baseline Model</th>
<th>Explanatory Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Variance</td>
<td>% of Total</td>
</tr>
<tr>
<td>Level 1 (Student)</td>
<td>0.15437*</td>
<td>52</td>
</tr>
<tr>
<td>Level 2 (Teacher)</td>
<td>0.13971*</td>
<td>48</td>
</tr>
<tr>
<td>Total</td>
<td>0.29408</td>
<td>0.20270</td>
</tr>
</tbody>
</table>

*p < .001.
response quality was allowed to vary randomly across teachers as represented by Equation 5.

$$Y_{ij} = \beta_0 + \beta_1(TREATMENT)_{ij} + \beta_2(RESPONSE\ FREQUENCY)_{ij} + r_{ij}.$$  \hspace{1cm} (4)

$$\beta_0 = \gamma_0 + u_{0j},$$  \hspace{1cm} (5)

$$\beta_1 = \gamma_1 + \beta_{10}(TEACHER\ DOSAGE)_{j} + u_{10j}$$  \hspace{1cm} (6)

$$\beta_2 = \gamma_2 + \beta_{20}(TEACHER\ DOSAGE)_{j} + u_{20j}$$  \hspace{1cm} (7)

As presented in Table 5 (“Explanatory Model”), teacher dosage accounted for a significant (65%) reduction in the unexplained Level 2 variance ($u_{10j}$) in the treatment effect, $\chi^2(1, N = 153) = 17.80, p < .001$. In contrast, student response frequency did not significantly (< 1%) reduce the unexplained student-level variance ($r_{ij}$) in the treatment effect, $\chi^2(1, N = 153) = 0.35, p > .50$. In addition, teacher dosage was a marginally significant predictor of participant responsiveness, $\beta_{10} = .26$, $t(4) = 2.47, p = .065$, indicating that the more essays offered by the teacher then the more connections were made by students. These results indicated that the number of essays offered by teachers accounted for variability in participant responsiveness, whereas student compliance did not.

Results Summary and Discussion

As an example of how the effectiveness of well-designed laboratory interventions is often diffused in the field, in this article we examined a motivational intervention designed by Hulleman and colleagues (Hulleman, 2008; Hulleman, Godes, et al., 2008; Hulleman & Harackiewicz, 2008; Hulleman et al., 2007). The intervention proved to be more effective in the laboratory ($g = 0.45$) than the field ($g = 0.05$) in predicting subsequent motivation. We explored this diffusion in treatment effectiveness through a fidelity analysis that examined the extent to which participants responded to the treatment. We calculated fidelity as three indices of achieved relative treatment strength—binary treatment receipt, average levels of responsiveness, and percentage of absolute amount of responsiveness. Regardless of how fidelity was calculated, achieved relative strength was about 1 standard deviation less in the classroom than the laboratory. In addition, greater levels of achieved relative strength were associated with greater differences in the motivational outcome—indicating that the intervention was more effective for participants who actually received the treatment than those who did not. Finally, our multilevel analyses of the sources of treatment infidelity indicated that the drop in participant responsiveness in the classroom was partially because of teacher (i.e., dosage), rather than student (i.e., frequency) factors.
IMPLEMENTATION SYSTEMS AND PROCESSES

The multilevel fidelity analysis presented in this article is unique and represents an approach to understanding treatment effectiveness and implementation fidelity recommended by both researchers and evaluation experts (e.g., Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Jo et al., 2008; Sloane, 2005). Just as intervention outcomes are subject to influences from multiple levels (e.g., students nested within classrooms and schools), so too are measures of implementation fidelity. Indeed, our analyses revealed that teacher behaviors (dosage) influenced quality of participant responsiveness to the intervention, rather than student factors, and suggest that remedial actions undertaken to ameliorate treatment infidelity would be most effective at the teacher level. That is, we would be best served by adding protocols or components to the intervention program which would increase the likelihood that teachers were providing an equal and sufficient number of treatment exposures.

Fixsen et al. (2005) referred to core implementation processes that are essential for intervention effectiveness as “implementation drivers.” Potential implementation drivers for our motivational intervention could be additional teacher training, ongoing coaching, and monitoring and feedback regarding how well the intervention is being implemented. Specific guidelines in terms of when the treatment should be administered, logbooks to record treatment delivery, and feedback (from the researcher) in terms of percentage of exposures provided relative to the theoretical program model could help keep teachers at a high level of implementation fidelity. Although not analyzed in our example, Fixsen et al. described two additional levels of implementation drivers. Organizational implementation drivers are at a slightly more distal level than the core implementation drivers and include staff selection, program evaluation, administrative support, and systems interventions. A third level of implementation drivers, called influence factors, includes social, economic, and political influence factors that may operate in a more indirect fashion. The more distal levels of the model—organizational and influence factors—can influence program implementation directly, or indirectly by impacting lower level factors.

REDEFINING THE CAUSAL AGENT

The analysis of fidelity presented in this article requires that we amend our conceptualization of the causal agent in interventions studies in two ways. First, if we accept that the intervention is not perfectly implemented—and most are not as exemplified by our laboratory intervention attaining only 58% of absolute fidelity—then the effects of the intervention rely on the degree to which intervention components were actually present ($T_{x} - T_{c}$) rather than on what was theoretically proscribed ($T_{x}^T - T_{c}^C$). In general, researchers are often in the position where they cannot explicate an a priori standard of treatment
fidelity that is necessary and sufficient for the treatment to be effective. This challenge is overcome through the conceptualization of treatment fidelity as achieved relative strength, thus allowing researchers to at least understand the relative differences between treatment and control conditions. However, this conceptualization of the intervention transforms the causal element into the relative difference between treatment and control conditions, or achieved relative strength (Cordray, 2006).

Second, regardless of the type or level of changes that are made, even if they are not intended to alter the theoretical model of change (i.e., implementation drivers), any modifications made to the intervention in the field, therefore, change the intervention. For example, by adding teacher supports for implementation, the intervention model is altered because the core components of the intervention have changed. In addition to quality of responsiveness, the model of change now requires a specific level of exposure supported by specific types of monitoring and feedback to teachers. In addition, implementation drivers that are added to an intervention program can add new, theoretically viable elements into the causal model. Perhaps providing additional monitoring and feedback to teachers produces a feeling of evaluation among teachers that is not present in the counterfactual group, and which impacts other teacher behaviors and student learning independent from the intervention. It is in this way that the model of change is altered, knowingly or not, to include this evaluation effect. From this perspective, it is not possible to exactly replicate a lab intervention in the field because the field setting requires additional implementation drivers that change the intervention and alter the theory of change. This is an often overlooked aspect of causality and inference in applied social science.

Now that the field of applied social science is beginning to understand how implementation fidelity (and a lack thereof) impacts our conclusions and inferences regarding program effectiveness, we encourage applied social scientists and program evaluators to consider how our models of causality are impacted by the heightened prominence given to implementation fidelity in randomized experiments. It is by applying the same sophisticated scientific analysis to implementation fidelity that we currently do to intervention outcomes that our understanding of intervention effectiveness will ultimately be advanced.

In this article we used a straightforward motivational intervention to demonstrate the methodological processes involved in analyzing implementation fidelity. The next step is to apply the indices of fidelity and achieved relative strength to more complex, programmatic interventions that contain multiple components. Doing so will raise important questions: Which measures of fidelity should be included in computing the achieved relative strength indices? Should all measured components be utilized, should some be dropped, or should a weighting scheme be employed? The real world of educational interventions contains multifaceted intervention programs with multiple layers of components. Integrating the multitude of factors into indices of fidelity and
achieved relative strength will be essential for properly evaluating intervention effectiveness.

ACKNOWLEDGMENTS


 The research was in part based on a doctoral dissertation submitted by Chris Hulleman to the University of Wisconsin-Madison under the supervision of Judith M. Harackiewicz. Chris Hulleman was supported in part by the Department of Psychology Research Award at the University of Wisconsin-Madison, and Institute for Education Sciences grants R305B050029 and 144-NL14. David Cordray was supported in part by grant R305U06002 from the Institute for Education Sciences. The authors wish to thank Joy Lesnick for her helpful comments and suggestions on an earlier draft of this manuscript.

REFERENCES


Achievement and achievement motives: Psychological and sociological approaches (pp. 75–146). San Francisco: W. H. Freeman.


**APPENDIX**

**Hedges’s g for means (Average Fidelity Index)**

\[
\hat{g} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{N_{\text{total}} - 2}}} \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right). \quad (A1)
\]

**Hedges’s g for means corrected for clustering**

\[
\hat{g} = \left(\frac{\bar{X}_1 - \bar{X}_2}{S_T}\right) \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right) \times \sqrt{1 - \frac{2(n - 1)p}{N - 2}} \quad (A2)
\]

Where,

\(\bar{X}_1\) = mean for group 1 (t\text{Tx})

\(\bar{X}_2\) = mean for group 2 (t\text{C})

\(S_T\) = pooled within groups standard deviation

\(n\) = average cluster size

\(p\) = Intra-class correlation (ICC)

\(N\) = total sample size

\(N_{\text{tx}}\) = treatment sample size

\(N_{\text{c}}\) = control sample size

With variance,

\[
V_g = \left(\frac{N_{\text{tx}} + N_c}{N_{\text{tx}} \times N_c}\right)(1 + (n - 1)p)
+ g\left(\frac{(N - 2)(1 - p)^2 + n(N - 2n)p^2 + 2(N - 2n)p(1 - p)}{2(N - 2)[(N - 2) - 2(n - 1)p]}\right)
\]
Hedges’s $g$ for proportions (Absolute Fidelity Index and Binary Complier Index)

\[
\hat{g} = 2^* \arcsin \theta(\sqrt{p_{Tx}}) - 2^* \arcsin \theta(\sqrt{p_{C}}) \times (1 - X \left(1 - \frac{3}{4(n_1 + n_2)} - 9\right)) \tag{A3}
\]

Where,

$p_{Tx}$ = proportion for the treatment group ($t_{Tx}$)

$p_{C}$ = proportion for the control group ($t_{C}$)

With variance,

\[
V_g = \frac{n_C + n_{Tx}}{n_C \times n_{Tx}} + \frac{g^2}{2 \times (n_C + n_{Tx})}
\]

Where,

$n_C$ = sample size in Control group

$n_{Tx}$ = sample size in Tx group

$\hat{g}$ = Hedges’s $g$

Hedges’s $g$ for proportions corrected for clustering

\[
\hat{g} = 2^* \arcsin \theta(\sqrt{p_{Tx}}) - 2^* \arcsin \theta(\sqrt{p_{C}}) \times \left(1 - \frac{3}{4(n_1 + n_2)} - 9\right) \times \sqrt{1 - \frac{2(n-1)p}{N-2}} \tag{A4}
\]

Where,

$p_{Tx}$ = proportion for the treatment group ($t_{Tx}$)

$p_{C}$ = proportion for the control group ($t_{C}$)

$p$ = Intra-class correlation (ICC)

With variance,

\[
V_g = \left(\frac{N_{Tx} + N_c}{N_{Tx} \times N_c}\right)(1 + (n - 1)p) + g \left(\frac{(N-2)(1-p)^2 + n(N-2n)p^2 + 2(N-2n)p(1-p)}{2(N-2)[(N-2) - 2(n-1)p]}\right)
\]

Where,

$g$ = Hedges’s $g$
\[ P = \text{Intra-class correlation (ICC)} \]

*Note.* The lower and upper bound for the Hedges’s \( g \) effect sizes are calculated using the formula for the 95\% confidence interval \( (g + 1.96 \times \text{Standard Error of } g) \). See Hedges (2007) for further details.