Published in final edited form as:
J Am Stat Assoc. 2013 June 1; 108(502): 469–482. doi:10.1080/01621459.2013.779832.

Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention

Tyler J. VanderWeele,
Harvard School of Public Health

Guanglei Hong,
University of Chicago

Stephanie M. Jones, and
Harvard Graduate School of Education

Joshua L. Brown
Fordham University

Abstract

Peer influence and social interactions can give rise to spillover effects in which the exposure of one individual may affect outcomes of other individuals. Even if the intervention under study occurs at the group or cluster level as in group-randomized trials, spillover effects can occur when the mediator of interest is measured at a lower level than the treatment. Evaluators who choose groups rather than individuals as experimental units in a randomized trial often anticipate that the desirable changes in targeted social behaviors will be reinforced through interference among individuals in a group exposed to the same treatment. In an empirical evaluation of the effect of a school-wide intervention on reducing individual students' depressive symptoms, schools in matched pairs were randomly assigned to the 4Rs intervention or the control condition. Class quality was hypothesized as an important mediator assessed at the classroom level. We reason that the quality of one classroom may affect outcomes of children in another classroom because children interact not simply with their classmates but also with those from other classes in the hallways or on the playground. In investigating the role of class quality as a mediator, failure to account for such spillover effects of one classroom on the outcomes of children in other classrooms can potentially result in bias and problems with interpretation. Using a counterfactual conceptualization of direct, indirect and spillover effects, we provide a framework that can accommodate issues of mediation and spillover effects in group randomized trials. We show that the total effect can be decomposed into a natural direct effect, a within-classroom mediated effect and a spillover mediated effect. We give identification conditions for each of the causal effects of interest and provide results on the consequences of ignoring "interference" or "spillover effects" when they are in fact present. Our modeling approach disentangles these effects. The analysis examines whether the 4Rs intervention has an effect on children's depressive symptoms through changing the quality of other classes as well as through changing the quality of a child's own class.

Keywords

Direct/indirect effects; interference; multilevel models; social interactions

Correspondence: Tyler J. VanderWeele Harvard School of Public Health, Departments of Epidemiology and Biostatistics 677 Huntington Avenue, Boston, MA 02115 617-432-7855 (phone); 617-432-1884 (fax) tvanderw@hsph.harvard.edu.

Introduction

Can schools do a better job improving children's social-emotional well-being? The Reading, Writing, Respect and Resolution (4Rs) program is aimed at promoting not only literacy development but also intergroup understanding and conflict resolution. This school-wide intervention program has three components: (i) a literacy-based curriculum in conflict resolution and social-emotional learning, (ii) training and ongoing coaching of teachers in the delivery of the 4Rs curriculum, and (iii) a family-based parent-child homework arrangement. The intervention was designed for an entire school on the basis of the theory that there would be reinforcement amongst teachers and students across different classrooms in the same building. In particular, students' social-emotional development was expected to improve by benefitting not only from what would potentially be an enhanced environment within their own classroom but also from an enhanced environment in the school as a whole. This is because students interact not only within a classroom but also in the hallways, in the cafeteria and on the playground with students from other classrooms. Therefore, the improvement of the quality of one classroom might affect the social-emotional outcomes of children enrolled in other classrooms in the same school.

This study investigates the mediating role of changes in class quality induced by the 4Rs intervention in influencing child depressive symptoms. Class quality encompasses instructional support, emotional support, and organizational climate. Specifi-cally, we evaluate the extent to which the effect of the 4Rs intervention on children is (i) mediated by changes in a child's own classroom quality, (ii) mediated by changes in the quality of classes other than the child's own, and (iii) potentially through other pathways. We analyze data from a group randomized experiment conducted in New York City in which elementary schools in matched pairs were assigned at random to either the 4Rs program or the control condition (Jones et al., 2010; Brown et al., 2010).

A conventional approach to mediation analysis consists of regressing the outcome on the treatment with and without the mediator variable. This approach takes the coe¢ cient for the treatment in the model without the mediator as a "total effect;" it takes the coefficient for the treatment in the model with the mediator as a "direct effect." The difference between the "total effect" and the "direct effect" is a measure of the "indirect effect." This approach to mediation is sometimes referred to as the "difference method." When mediator and outcome are both continuous, this approach gives the same estimate as the "product-of-coefficients method," which itself takes as the "indirect effect" the product of the coefficient of the treatment in a model for the mediator and the coefficient of the mediator in a model for the outcome. This product-of-coefficients method is what is typically employed in mediation analyses with structural equation models. Although the product and difference methods give the same estimates with linear models, similar results do not hold with dichotomous outcomes or in non-linear model (MacKinnon, 2008).

The conventional approach is subject to several limitations when applied to the current study. First, the standard regression approach typically ignores selection into mediator levels. Although the treatment was randomized, teachers and students within a school were not randomly assigned to alternative levels of class quality. Analyses ignoring this selection issue are subject to potentially severe biases due to confounding (Judd and Kenny, 1981; Robins and Greenland, 1992; Pearl, 2001). While the problem of selection into mediator levels is endemic to all meditational studies, the current study conditions on covariates to make the "no-selection" assumption more plausible. The second issue with the standard regression approach is that potential interactions between the effects of treatment and mediator on the outcome are typically ignored. The assumption of no treatment-mediator interaction would be violated if, for example, improvement in class quality in a 4Rs school

produced a larger impact on child outcomes than it would had the school been assigned to the control condition instead. Recent literature on causal inference has made clear that mediation analysis becomes considerably more complex when such interactions are present (Pearl, 2001).

Perhaps even more importantly, the literature on mediation has ignored spill-over effects in organizational settings. The issue is referred to as one of “interference between units”in the statistics literature (Cox, 1958). No interference between units is a component of Rubin's Stable Unit Treatment Value Assumption or SUTVA (Rubin, 1980, 1986). The assumption will be violated in settings in which social interactions allow one individual's exposure or treatment assignment to affect the outcomes of other individuals. The assumption will also be violated in mediation studies if the mediator value displayed by one unit affects the outcomes of other units even if these units have been assigned to the same treatment. Such interference is part of the rationale of the 4Rs program. In theory, a child's social-emotional outcomes may depend not only on the class quality experienced in the child's own classroom but also on the quality of other classrooms experienced by other children attending the same school due to social interactions among children from different classrooms. The school-wide intervention was designed to bring together educators’ collective efforts within a school so as to maximize positive social interactions and minimize negative interactions among children. Conceptualizing and analyzing the mediation effects are considerably more complex in the face of such interference.

The literature on mediation in a multilevel setting (VanderWeele, 2010) and on causal inference under interference (Hong and Raudenbush, 2006; Sobel, 2006; Rosenbaum, 2007; Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2011) has not explicitly considered how to deal with the assessment of spillover effects in mediation analysis. To address the question of spillover effects in the 4Rs intervention, we will extend earlier work. Specifically, our work extends beyond the results of VanderWeele (2010) by (i) giving counterfactual definitions for within-classroom and spillover mediated effects and showing that not only does the total effect of the intervention decompose into a direct effect and an indirect effect but also that the indirect effect itself decomposes into an effect mediated through the quality of a child's own classroom and a spillover effect mediated by the quality of the other classrooms at a school; (ii) giving identification results for within-classroom and spillover mediated effect; (iii) giving two results on the consequences of ignoring interference when it is present; (iv) mapping the conceptual framework to a class of models that disentangle these effects; and (v) applying the methodology to the 4R's intervention study.

This case study itself will provide a basic template for the numerous group-randomized trials and quasi-experimental studies in which questions of mediation are of scientific interest and interference is likely present. Interference among units is common in social settings. For example, banning smoking in office buildings could affect individual health not only through changing one's own smoking behavior but also through enforcing the new rules upon one's colleagues. We hope that the approach described in the present paper will offer a new framework and concrete guidance regarding how to investigate causal mechanisms that involve interference among units. This template will be applicable to settings in which the treatment is assigned at the organization or community level while the mediator and the outcome are at lower levels than the treatment. The paper not only provides concepts and results for new analyses but also highlights the strong assumptions required to identify causal quantities that may be of interest. Such strong assumptions are often implicitly made but not acknowledged in the conventional analysis. By drawing attention to these strong assumptions and evaluating them in the context of the 4Rs application study, we hope that subsequent studies can be designed and data can be collected that may render these

assumptions more plausible so that investigators can estimate the mediated effects of substantive interest.

The paper is organized as follows: Section 2 introduces notation and definitions of the causal effects. Section 3 presents the assumptions required for identifying direct, indirect, and spillover effects. Section 4 describes the analytic procedure and gives the estimation results. Section 5 concludes by discussing the strengths and limitations of the study.

2. Notation, Definitions, and Framework

In the 4Rs study, the randomized treatment occurred at the school level; the mediator was measured at the classroom level and the outcome at the child level. We adapt the notation and multilevel mediation framework of VanderWeele (2010) to accommodate this setting; and we apply and extend the work on interference of Hong and Raudenbush (2006) and Hudgens and Halloran (2008) to consider the spillover effects of interest. Let T_k denote the school-wide randomized treatment (1 for the 4Rs intervention; 0 for control) for school k. Let M_{jk} denote the classroom level mediator for classroom j in school k indicating classroom quality. Let J_k denote the number of classrooms in school k. Let Y_{ijk} denote the child level outcome for child i in classroom j and school k indicating child depressive symptoms.

2.1 Total Effects

In order to define the causal effects of interest we proceed by introducing potential outcome variables corresponding to values of the outcome under possibly contrary to fact treatment and mediator scenarios and also values of the mediator under possibly contrary to fact treatment scenarios. In particular, let $Y_{ijk}(t_k)$ denote the potential or counterfactual outcome that child i in classroom j and school k would have obtained if the school-level treatment, T_k, were set to t_k. Similarly, let $M_{jk}(t_k)$ denote the potential or counterfactual mediator that classroom j in school k would have obtained if the school-level treatment, T_k, were set to t_k. We assume that children do not change schools as the result of assigning a particular school to a certain treatment. Hong and Raudenbush (2006) refer to this assumption as that of "intact clusters." We also assume that there is no interference between schools, that is, the treatment received at one school does not affect the outcomes of the children at any other school. Hence our discussion considers only interference within schools, not interference between schools.

With this notation we can define the effect of the treatment on the outcome, $E[Y_{ijk}(1) - Y_{ijk}(0)]$, and the effect of the treatment on the mediator, $E[M_{jk}(1) - M_{jk}(0)]$. These effects correspond to the difference in average child outcomes and that in classroom quality scores had all schools received the 4Rs interventions as compared with if none of the schools had received the intervention. Because the treatment is randomized, these causal effects can be estimated using a simple intent-to-treat analysis. In this paper, we are not, however, interested simply in the overall effect of the treatment on classroom quality and on child social-emotional outcomes but rather also in the extent to which the effect of the 4Rs intervention on child outcomes is mediated by classroom quality. In order to define the direct and indirect effects of interest, we will need to introduce additional counterfactual quantities below.

2.2 Potential Outcomes Incorporating Interference

In studying mediation, we need to consider potential outcomes if both the treatment and the mediator had been set to values possibly other than what they in fact were. Because child outcomes may depend not only on the quality of the child's own classroom but also on the quality of other classrooms, we need to incorporate such within-school interference into our

potential outcomes notation. Let $Y_{ijk}(t_k, m_{jk}, \mathbf{m}_{-jk})$ denote the counterfactual outcome that child i in classroom j and school k would have obtained if the school-level treatment in school k were set to t_k, if the quality in classroom j of school k were set to m_{jk} and if the quality of all classrooms in school k other than classroom j were set to the vector $\mathbf{m}_{-jk} = (m_{1k}, ..., m_{j-1k}, m_{j+1k}, ..., m_{J_k k})$. As we discussed in the Introduction section, the theory underlying the 4Rs intervention suggests that such interference would indeed be present. Note that here, for a particular school, either the entire school is treated ($T_k = 1$) or the entire school is untreated ($T_k = 0$), whereas in Hudgens and Halloran (2008) the treatment is at the individual level so different clusters may have a proportion treated anywhere between 0 and 1.

Following Hong and Raudenbush (2006) and Hudgens and Halloran (2008), we assume that the potential outcome $Y_{ijk}(t_k, m_{jk}, \mathbf{m}_{-jk})$ depends on $\mathbf{m}_{-jk}$ through some scalar function $G(\mathbf{m}_{-jk})$ of $\mathbf{m}_{-jk}$ so that we may express the potential outcome $Y_{ijk}(t_k, m_{jk}, \mathbf{m}_{-jk})$ as $Y_{ijk}(t_k, m_{jk}, G(\mathbf{m}_{-jk}))$. For example, $G(\mathbf{m}_{-jk})$ may denote the average classroom quality for all classrooms in school k other than classroom j. In some social-emotional outcome settings, it may be more reasonable for the function $G(\mathbf{m}_{-jk})$ to be the geometric mean of classroom quality for all classrooms in school k other than classroom j so that if there is greater dispersion in classroom quality, the mean will be less than would be the case if the quality of the various classrooms were uniformly consistent. The assumption that $G(\mathbf{m}_{-jk})$ is a scalar function of $\mathbf{m}_{-jk}$ will not in fact be necessary for any of our identification results below but simplifies modeling considerably. In practice, it is unlikely there would be sufficient data to avoid assumptions on $G(\mathbf{m}_{-jk})$ in modeling; but our identification results themselves will hold even if $G(\mathbf{m}_{-jk})$ is the entire vector $\mathbf{m}_{-jk}$. Here we let $Y_{ijk}(t, m, g)$ denote the outcome for child i in classroom j and school k if the school received treatment t, the child's classroom was at the quality level m, and the scalar function of the quality of other classrooms, $G(\mathbf{m}_{-jk})$, took the value g. In our discussion of mediation below we will also use the notation $\mathbf{M}_{-jk}(t_k)$ to denote $(M_{1k}(t_k), ..., M_{j\,1k}(t_k), M_{j+1k}(t_k), ..., M_{J_k k}(t_k))$.

2.3 Controlled Direct Effects

With this notation in place, we can draw upon the literature on causal inference for direct and indirect effects (Greenland and Robins, 1992; Pearl, 2001) and on mediation in a multilevel context (VanderWeele, 2010) to define the direct and indirect effects of interest in a group-randomized trial. Note that we essentially have two mediators of interest here, the quality of a child's own classroom and the quality of other classrooms at the school. The causal contrast

$$E\left[Y_{ijk}(1, m, g) - Y_{ijk}(0, m, g)\right]$$

provides a measure of the direct effect of the 4Rs program but also intervening to fix the classroom quality of the child's own classroom to level m and intervening to fix the average quality of other classrooms to g. This quantity is referred to as a controlled direct effect of treatment intervening to fix the values of the mediators to a particular level irrespective of treatment T_k. The contrast at different levels of m and g could also be used to assess whether the effect of treatment on the outcome varies with a child's own classroom quality or the quality of other classrooms. Likewise the contrast

$$E\left[Y_{ijk}(t, m, g) - Y_{ijk}(t, m^*, g)\right]$$

could be used to assess the effect of a child's own classroom quality (comparing levels m and m^*) on a child's outcome. We may examine whether the effect of classroom quality depends on the presence of the 4Rs intervention or whether it depends on the quality of classrooms other than the child's own (i.e. whether the contrast varies with t or g). Similarly, the contrast

$$E\left[Y_{ijk}(t,m,g) - Y_{ijk}(t,m,g^*)\right]$$

could be used to assess the spillover effect on a child's outcome of the quality of classrooms other than the child's own. We may examine whether this spillover effect depends on the presence of the 4Rs intervention or whether it depends on the child's own classroom quality (i.e. whether the contrast varies with t or m). For example, it may turn out that a child's social-emotional outcome is relatively unaffected by such spillover effects when a child's own classroom quality is high but that behavior is more susceptible to spillover effects when a child's classroom quality is low.

2.4 Natural Direct and Indirect Effects

An alternative conceptualization of a direct effect is what is sometimes referred to as a "natural direct effect" (Pearl, 2001) which provides a measure of the direct effect of treatment when the classroom quality mediators are set to the levels they would have been at under the control condition. The natural direct effect in this setting is defined as $E[Y_{ijk}(1, M_{jk}(0), G(\mathbf{M}_{-jk}(0))) - Y_{ijk}(0, M_{jk}(0), G(\mathbf{M}_{-jk}(0)))]$. We can also define a natural indirect effect as $E[Y_{ijk}(1, M_{jk}(1), G(\mathbf{m}_{-jk}(1))) Y_{ijk}(1, M_{jk}(0), G(\mathbf{M}_{-jk}(0)))]$. As in the case of non-clustered treatments without interference (Pearl, 2001), the total effect of the intervention on the outcome, $E[Y_{ijk}(1) - Y_{ijk}(0)]$, decomposes into a natural direct and indirect effect:

$$\begin{aligned}
&E\left[Y_{ijk}(1) - Y_{ijk}(0)\right] \\
&\quad = E\left[Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(1)\right)\right) - Y_{ijk}\left(0, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right] \\
&\quad = E\left[Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(1)\right)\right) - Y_{ijk}\left(1, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right] \\
&\quad + E\left[Y_{ijk}\left(1, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right) - Y_{ijk}\left(0, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right]
\end{aligned}$$

where the first expression in the sum is the natural indirect effect and the second expression is the natural direct effect. Note that the decomposition is achieved simply by adding and subtracting the term $E[Y_{ijk}(1, M_{jk}(0), G(\mathbf{M}_{-jk}(0)))]$.

The decomposition will hold irrespective of the functional form of $Y_{ijk}(t, m, g)$. In particular, the decomposition will hold even if there are interactions between the effects of the treatment and the mediator on the outcome, that is, if the controlled direct effect $E[Y_{ijk}(1, m, g) - Y_{ijk}(0, m, g)]$ is not constant across m and g. Robins and Greenland (1992) refer to the decomposition above as one of a total effect into a pure direct effect and a total indirect effect; a decomposition is also possible into a total direct effect and a pure indirect effect (Robins and Greenland, 1992; Robins, 2003):

$$\begin{aligned}
&E\left[Y_{ijk}(1) - Y_{ijk}(0)\right] \\
&\quad = E\left[Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(1)\right)\right) - Y_{ijk}\left(0, M_{jk}(1), G\left(\mathbf{M}_{-jk}(1)\right)\right)\right] \\
&\quad + E\left[Y_{ijk}\left(0, M_{jk}(1), G\left(\mathbf{M}_{-jk}(1)\right)\right) - Y_{ijk}\left(0, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right];
\end{aligned}$$

in other words, the decomposition is not unique. The pure direct effect need not equal the total direct effect and similarly the pure indirect effect need not equal the total indirect effect unless the effect of classroom quality on the outcome does not depend on treatment.

2.5 Mediated Spillover Effect and Within-Class Mediated Effect

In fact, the natural indirect effect itself decomposes into an effect mediated through the quality of a child's own classroom and a spillover effect through the quality of other classrooms at a school. Consider the contrast

$$E\left[Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(1)\right)\right) - Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right].$$

This quantity compares child outcomes under treatment, with the quality of a child's own classroom set to what it would be under treatment but then contrasts what would happen if the quality of other classrooms were to set to the level they would be with treatment to what they would be without treatment. The contrast is one way to formally capture the effect of the treatment on a child's outcome as mediated through the quality of classrooms other than the one the child is in. We will refer to this contrast as the spillover mediated effect. Similarly, we could also consider the contrast $E[Y_{ijk}(1, M_{jk}(1), G(\mathbf{M}_{-jk}(0))) - Y_{ijk}(1, M_{jk}(0), G(\mathbf{M}_{-jk}(0)))]$; this quantity compares child outcomes under treatment, with the classroom quality in other classrooms set to what they would have been without treatment but then contrasts what would happen if the quality of a child's own classroom were to set to the level it would be either with treatment to what it would be without treatment. The contrast is one way to conceive of the effect of the treatment on a child's outcome as mediated through the quality of the child's own classroom. We will refer to this contrast as the within-classroom mediated effect. Having defined these effects, we can now decompose the natural indirect effect into the spillover mediated effect and the within-classroom mediated effect: $E[Y_{ijk}(1, M_{jk}(1), G(\mathbf{M}_{-jk}(1))) - Y_{ijk}(1, M_{jk}(0), G(\mathbf{M}_{-jk}(0)))]$

$$\begin{aligned}
&= E\left[Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(1)\right)\right) - Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right] \\
&\quad + E\left[Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(0)\right)\right) - Y_{ijk}\left(1, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right].
\end{aligned}$$

As with the decomposition of a total effect into a natural direct and indirect effect, the decomposition of a natural indirect effect into a mediated spillover effect and a within-classroom mediated effect is not unique.

From our discussion above it follows that we can decompose a total causal effect into the sum of (i) a spillover mediated effect, (ii) a within-classroom mediated effect and (iii) a natural direct effect as follows: $E[Y_{ijk}(1) - Y_{ijk}(0)]$

$$\begin{aligned}
&= E\left[Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(1)\right)\right) - Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right] \\
&\quad + E\left[Y_{ijk}\left(1, M_{jk}(1), G\left(\mathbf{M}_{-jk}(0)\right)\right) - Y_{ijk}\left(1, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right] \\
&\quad + E\left[Y_{ijk}\left(1, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right) - Y_{ijk}\left(0, M_{jk}(0), G\left(\mathbf{M}_{-jk}(0)\right)\right)\right].
\end{aligned}$$

In the next section we consider the identification of these various direct, mediated and spillover effects.

3. Identification of Direct, Indirect and Spillover Effects

In this section we extend the identification results of VanderWeele (2010) on mediation to allow for the spillover effects and within-classroom mediated effects, described above, which arise because of interference. We also give results on the consequences of ignoring interference when it is in fact present. Let $\mathbf{X}_{ijk}$ denote child-level baseline covariates, $\mathbf{W}_{jk}$ denote classroom level baseline covariates and $\mathbf{V}_k$ denote school level baseline covariates. Let $\mathbf{X}_{-ijk}$ denote the vector of child-level baseline covariates for children in school k other than child i in classroom j, $\mathbf{W}_{-jk}$ to denote classroom-level baseline covariates for classrooms in school k other than classroom j. Let $\mathbf{M}_k = (M_{1k}, ..., M_{J_k k})$ and $\mathbf{M}_k(t) = (M_{1k}(t), ..., M_{J_k k}(t))$. For sets of random variables **A, B** and **C**, we will use $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}$ to denote that **A** is independent of **B** conditional on **C**. We will consider certain functions of the baseline covariates of other children in the classroom (or even at the school), $\mathbf{h}_1(\mathbf{X}_{-ijk})$, and of baseline covariates of classrooms other than a child's own $\mathbf{h}_2(\mathbf{W}_{-jk})$. In practice, $\mathbf{h}_1(\mathbf{X}_{-ijk})$ will likely be fairly constant within a classroom which may have twenty or more children (since only one child is different in the computation of each of these means or functions) and thus may plausibly be replaced by the aggregate summary of $\mathbf{X}_{ijk}$ for all children in classroom i and school k. To simplify notation we let $\mathbf{C}_{ijk} = (\mathbf{X}_{ijk}, \mathbf{W}_{jk}, \mathbf{V}_k, \mathbf{h}_1(\mathbf{X}_{-ijk}), \mathbf{h}_2(\mathbf{W}_{-jk}))$. We present five identification results, one for controlled direct effects, one for natural direct and indirect effects, one for the spillover mediated effect and within-classroom mediated effect, and two for the consequences of ignoring interference when it is in fact present. Several independence conditions will follow immediately from the randomization of treatment; however, we will state the assumptions in such a way so as to allow the results to also be applicable to observational settings. As we move through the presentation of these result, we will see that each subsequent result requires stronger and stronger assumptions. In our actual analysis of the 4Rs data we evaluate the plausibility of each of these assumptions and choose to rely on some of the less strong assumptions

3.1 Total and Controlled Direct Effects Under Interference

Since treatment T_k is randomized, it will be independent of all pretreatment covariates and also of the potential outcomes as these can be viewed as (unobserved) pretreatment covariates. We thus have that $T_k \perp\!\!\!\perp \{Y_{ijk}(t, m, g), \mathbf{C}_{ijk}\}$ and thus also that for all t, m, g,

$$Y_{ijk}(t, m, g) \perp\!\!\!\perp T_k | \mathbf{C}_{ijk}. \quad (1)$$

which is sufficient for identifying the total effect. We can now state our first result concerning controlled direct effects. The result requires not only that treatment is randomized so that (1) holds but also that selection into mediator levels of classroom quality is effectively random conditional on treatment, T_k, and the covariates $\mathbf{C}_{ijk} = (\mathbf{X}_{ijk}, \mathbf{W}_{jk}, \mathbf{V}_k, \mathbf{h}_1(\mathbf{X}_{-ijk}), \mathbf{h}_2(\mathbf{W}_{-jk}))$. Proofs of all theorems are given in the online supplement.

Theorem 1. If treatment T_k is randomized then (1) will hold. If, in addition, we have that for all t, m, g,

$$Y_{ijk}(t, m, g) \perp\!\!\!\perp \left\{M_{ij}, G\left(\mathbf{M}_{-jk}\right)\right\} | T_k, \mathbf{C}_{ijk} \quad (2)$$

then the average counterfactual outcome $Y_{ijk}(t, m, g)$ conditional on $\mathbf{C}_{ijk}$ is identified and is given by: $E[Y_{ijk}(t, m, g)|\mathbf{C}_{ijk} = \mathbf{c}] =$

$$E\left[Y_{ijk}|T_k{=}t, M_{jk}{=}m, G\left(\mathbf{M}_{-jk}\right){=}g, \mathbf{C}_{ijk}{=}\mathbf{c}\right]. \quad (3)$$

Under assumption (2), we could apply Theorem 1 twice (once for $t = 1$ and once for $t = 0$) and sum (3) over $\mathbf{C}_{ijk}$ to obtain the controlled direct effect of treatment, $E[Y_{ijk}(1, m, g) - Y_{ijk}(0, m, g)] =$

$$\sum_{\mathbf{c}} \left\{ E\left[Y_{ijk} | T_k = 1, M_{jk} = m, G\left(\mathbf{M}_{-jk}\right) = g, \mathbf{C}_{ijk} = \mathbf{c} \right] - E\left[Y_{ijk} | T_k = 0, M_{jk} = m, G\left(\mathbf{M}_{-jk}\right) = g, \mathbf{C}_{ijk} = \mathbf{c} \right] \right\} \times P\left(\mathbf{C}_{ijk} = \mathbf{c}\right). \quad (4)$$

Similarly we could obtain expressions for the average effect on the outcome of a change in the quality of a child's classroom, while holding the school-level intervention fixed and the classroom quality of the other classrooms fixed, $E[Y_{ijk}(t, m, g) - Y_{ijk}(t, m^*, g)] =$

$$\sum_{\mathbf{c}} \left\{ E\left[Y_{ijk} | T_k = t, M_{jk} = m, G\left(\mathbf{M}_{-jk}\right) = g, \mathbf{C}_{ijk} = \mathbf{c} \right] - E\left[Y_{ijk} | T_k = t, M_{jk} = m^*, G\left(\mathbf{M}_{-jk}\right) = g, \mathbf{C}_{ijk} = \mathbf{c} \right] \right\} P\left(\mathbf{C}_{ijk} = \mathbf{c}\right). \quad (5)$$

We could also obtain expressions for the average effect on the outcome of a change in the measure of the quality of other classrooms from level $G(\mathbf{M}_{-jk}) = g$ to level $G(\mathbf{M}_{-jk}) = g^*$ while holding the school-level intervention fixed and the classroom quality a child's own classroom fixed, $E[Y_{ijk}(t, m, g) - Y_{ijk}(t, m, g^*)] =$

$$\sum_{\mathbf{c}} \left\{ E\left[Y_{ijk} | T_k = t, M_{jk} = m, G\left(\mathbf{M}_{-jk}\right) = g, \mathbf{C}_{ijk} = \mathbf{c} \right] - E\left[Y_{ijk} | T_k = t, M_{jk} = m, G\left(\mathbf{M}_{-jk}\right) = g^*, \mathbf{C}_{ijk} = \mathbf{c} \right] \right\} P\left(\mathbf{C}_{ijk} = \mathbf{c}\right). \quad (6)$$

The conclusion of Theorem 1 still applies if the independence assumptions (1) and (2) hold only in mean rather than in distribution.

We now turn to the interpretation of condition (2). The quality of a classroom, M_{jk}, will in general depend on the particular teacher in classroom j and on which children are assigned to classroom j. For a particular child i in classroom j and school k, the potential outcomes $Y_{ijk}(t, m, g)$, as t, m, and g vary, can be thought of as how well a child would fare under different scenarios concerning treatment, classroom quality, and the classroom quality of other classrooms. Condition (2) is contingent on the process by which children are organized into classrooms and assigned to teachers which thereby gives rise to the quality of the classroom. Condition (2) requires that, for a particular child, conditional on that child's baseline covariates, X_{ijk}, certain characteristics of the school and teachers, $\mathbf{W}_{jk}$, $\mathbf{V}_k$, $\mathbf{h}_2(\mathbf{W}_{-jk})$, and some measure of the baseline covariates of other children, $\mathbf{h}_1(\mathbf{X}_{-ijk})$, this process of class assignment is independent of how well the child would fare (i.e. of the potential outcomes $Y_{ijk}(t, m, g)$). If the schools' assignment mechanisms use only information on $\mathbf{C}_{ijk} = (\mathbf{X}_{ijk}, \mathbf{W}_{jk}, \mathbf{V}_k, \mathbf{h}_1(\mathbf{X}_{-ijk}), \mathbf{h}_2(\mathbf{W}_{-jk}))$ to place children into classrooms, or if the information they use beyond $\mathbf{C}_{ijk}$ is unrelated to the potential outcomes, then this may be a reasonable assumption. However, even if children are randomly assigned to classrooms, assumption (2) may not hold exactly because a child's own characteristics may have an effect on classroom quality; the value of the mediator is partially caused by the child's membership in the class. This issue may be partially circumvented by stratification or control for covariates that relate to child characteristics that may affect classroom quality; the issue will also be partially mitigated in settings with larger class sizes.

For the analysis described below to be valid the covariates $\mathbf{C}_{ijk}$ for which control is made must not be affected by treatment. If covariates required for assumption (2) to hold are affected by treatment then an alternative identification approach for direct effects would be needed (Pearl, 2001; VanderWeele, 2009). Note that assumption (2) is effectively also presupposed by the conventional mediation analysis, though sometimes left unstated; moreover with conventional mediation analyses, effort is often not devoted to trying to control for covariates that may render assumption (2) more plausible. Finally the

conventional analysis in addition to making these assumptions does not accommodate treatment-mediator interactions.

Even under assumption (2) the conditional expectation $E[Y_{ijk}|T_k = t, M_{jk} = m, G(\mathbf{M}_{-jk}) = g, \mathbf{C}_{ijk} = \mathbf{c}]$ must be modeled. This could be done by a multi-level regression model. Covariate overlap for the variables in **C** across strata defined by T, M, and G should be checked to avoid extrapolation of the regression model to regions where data is not available. We discuss modeling further below.

3.2 Natural Direct and Indirect Effects Under Interference

The expressions above for the various controlled direct effects $E[Y_{ijk}(1, m, g) - Y_{ijk}(0, m, g)]$, $E[Y_{ijk}(t, m, g) - Y_{ijk}(t, m^*, g)]$ and $E[Y_{ijk}(t, m, g) - Y_{ijk}(t, m, g^*)]$ can be useful in assessing how important each individual factor - treatment, quality of a student's own classroom, and the quality of the other classrooms - is in determining child-level outcomes while holding the other factors fixed. However, in order to assess mediation, it is also of interest to consider the extent to which the effect of treatment, T_k, on child-level outcomes, Y_{ijk}, is mediated through the other two factors, quality of a student's own classroom, M_{jk}, and quality of the other classrooms, $G(\mathbf{M}_{-jk})$. For this we need to consider the effect of treatment on classroom quality in addition to assessing the effect of treatment and classroom quality on the outcome. Because treatment is randomized, it will be the case that $T_k \perp\!\!\!\perp \{M_{jk}(t), G(\mathbf{M}\text{-}_{jk}(t)), \mathbf{C}_{ijk}\}$ and also that for all t,

$$\left(M_{jk}(t), G\left(\mathbf{M}_{-jk}(t)\right)\right) \perp\!\!\!\perp T_k | \mathbf{C}_{ijk}. \quad (7)$$

This brings us to our second identification result which allows for the identification of the natural direct effect and the total natural indirect effect. In a third identification result below we will consider the identification of spillover mediated effects and within-classroom mediated effects.

Theorem 2. If treatment T_k is randomized then (1) and (7) will hold. If, in addition, (2) holds and we furthermore have that for all t, t^*, m, g,

$$Y_{ijk}(t, m, g) \perp\!\!\!\perp \left\{M_{jk}(t^*), G\left(\mathbf{M}_{-jk}(t^*)\right)\right\} | \mathbf{C}_{ijk} \quad (8)$$

then the average counterfactual outcome $Y_{ijk}(t, M_{jk}(t^*), G(\mathbf{M}_{-jk}(t^*)))$ conditional on $\mathbf{C}_{ijk}$ is identified and is given by $E[Y_{ijk}(t, M_{jk}(t^*), G(\mathbf{M}_{-jk}(t^*)))|\mathbf{C}_{ijk} = \mathrm{c}] =$

$$\sum_m \sum_g E\left[Y_{ijk} | T_k = t, M_{jk} = m, G\left(M_{-jk}\right) = g, \mathbf{C}_{ijk} = \mathbf{c}\right] \times P\left(M_{jk} = m, G\left(\mathbf{M}_{-jk}\right) = g | T_k = t^*, \mathbf{C}_{ijk} = \mathbf{c}\right). \quad (9)$$

Assuming (2) holds, assumption (8) then essentially requires that there is no post-treatment covariate that is an effect of treatment T_k that in turn affects both the mediator levels of the various classrooms and also the child-level outcomes (Pearl, 2001). This may be plausible if the mediator occurs not long after the treatment. See also Pearl (2001) and VanderWeele (2010) for further discussion. We note that without assumption (8), there is no data to identify $E[Y_{ijk}(t, M_{jk}(t^*), G(\mathbf{M}_{-jk}(t^*)))|\mathbf{C}_{ijk} = \mathrm{c}]$; it is assumption (8) that allows us to draw inferences about the counterfactual conditional expectation $E[Y_{ijk}(t, M_{jk}(t^*), G(\mathbf{M}_{-jk}(t^*)))|\mathbf{C}_{ijk} = \mathbf{c}]$ from the observed conditional expectation $E[Y_{ijk}|T_k = t, M_{jk} = m, G(\mathbf{M}_{-jk}) = g, \mathbf{C}_{ijk} = \mathbf{c}]$. Of course, even under assumption (8), sufficient data would have to be available to model $E[Y_{ijk}|T_k = t, M_{jk} = m, G(\mathbf{M}_{-jk}) = g, \mathbf{C}_{ijk} = \mathbf{c}]$ so as to be able to extrapolate from the observed data. As above, covariate overlap for the variables in **C** across strata defined by T,

M, and G should be checked to avoid extrapolation of the regression model to regions where data is not available.

The conclusion of Theorem 2 also holds in an observational setting under independence assumptions (1), (2), (7) and (8). Note that if T_k is randomized then (8) implies (2). Note also that the conclusion of Theorem 2 still applies if the independence assumptions (1), (2) and (8) hold only in mean rather than in distribution; independence assumption (7), however, must hold in distribution.

We can apply Theorem 2 to obtain empirical expressions for the natural direct effect and the natural indirect effect. Under assumptions (2) and (8), the natural direct effect described in the previous section is given by $E[Y_{ijk}(1, M_{jk}(0), G(\mathbf{M}_{-jk}(0))) - Y_{ijk}(0, M_{jk}(0), G(\mathbf{M}_{-jk}(0)))] =$

$$\sum_{\mathbf{c}}\sum_{m}\sum_{g}\left\{E\left[Y_{ijk}|T_k=1, M_{jk}=m, G\left(\mathbf{M}_{-jk}\right)=g, \mathbf{C}_{ijk}=\mathbf{c}\right] - E\left[Y_{ijk}|T_k=0, M_{jk}=m, G\left(\mathbf{M}_{-jk}\right)=g, \mathbf{C}_{ijk}=\mathbf{c}\right]\right\} \times P\left(M_{jk}=m, G\left(\mathbf{M}_{-jk}\right)=g|T_k=0, \mathbf{C}_{ijk}=\mathbf{c}\right) P\left(\mathbf{C}_{ijk}=\mathbf{c}\right) \quad (10)$$

and the natural indirect effect described in the previous section is given by $E[Y_{ijk}(1, M_{jk}(1), G(\mathbf{M}_{-jk}(1))) - Y_{ijk}(1, M_{jk}(0), G(\mathbf{M}_{-jk}(0)))] =$

$$\sum_{\mathbf{c}}\sum_{m}\sum_{g}E\left[Y_{ijk}|T_k=1, M_{jk}=m, G\left(\mathbf{M}_{-jk}\right)=g, \mathbf{C}_{ijk}=\mathbf{c}\right] \times\left\{P\left(M_{jk}=m, G\left(\mathbf{M}_{-jk}\right)=g|T_k=1, \mathbf{C}_{ijk}=\mathbf{c}\right) - P\left(M_{jk}=m, G\left(\mathbf{M}_{-jk}\right)=g|_{T_k}=0, \mathbf{C}_{ijk}=\mathbf{c}\right)\right\} P\left(\mathbf{C}_{ijk}=\mathbf{c}\right). \quad (11)$$

We also note, employing an idea in Petersen et al. (2006), that in order to identify the natural direct effect in (10), assumption (8) could be further weakened to require only that $\{Y_{ijk}(1, m, g) - Y_{ijk}(0, m, g)\} \perp\!\!\!\perp \{M_{jk}(0), G(\mathbf{M}_{-jk}(0))\}|\mathbf{C}_{ijk}$ for all m, g.

3.3 Within-classroom and spillover mediated effects

As noted above the natural indirect effect can be further decomposed into a spillover mediated effect and a within-classroom mediated effect. Our next theorem gives an identification result for these spillover and within-classroom mediated effects. In addition to assumptions (2) and (8), the result will require one further assumption in order to identify these effects.

Theorem 3. If treatment T_k is randomized then (1) and (7) will hold. If, in addition (2) and (8) hold and we furthermore have that for $t' \neq t^*$,

$$M_{jk}\left(t'\right) \perp\!\!\!\perp G\left(\mathbf{M}_{-jk}\left(t^*\right)\right)|\mathbf{C}_{ijk} \quad (12)$$

then the average counterfactual outcome $Y_{ijk}(t, M_{jk}(t'), G(\mathbf{M}_{-jk}(t^*)))$ conditional on $\mathbf{C}_{ijk}$ is identified and is given by $E[Y_{ijk}(t, M_{jk}(t'), G(\mathbf{M}_{-jk}(t^*)))|\mathbf{C}_{ijk} = \mathbf{c}] =$

$$\sum_{m}\sum_{g}E\left[Y_{ijk}|T_k=t, M_{jk}=m, G\left(M_{-jk}\right)=g, \mathbf{C}_{ijk}=\mathbf{c}\right]\times P\left(M_{jk}=m|\ T_k=t', \mathbf{C}_{ijk}=\mathbf{c}\right) P\left(G\left(\mathbf{M}_{-jk}\right)=g|\ T_k=t^*, \mathbf{C}_{ijk}=\mathbf{c}\right). \quad (13)$$

Theorem 3 still applies if the independence assumptions (1), (2) and (8) hold only in mean rather than in distribution; independence assumptions (7) and (12), however, must in general hold in distribution.

The interpretation of assumption (12) is that conditional on a child's baseline covariates, $\mathbf{X}_{ijk}$, certain characteristics of the school and teachers, $\mathbf{W}_{jk}$, $\mathbf{V}_k$, $\mathbf{h}_2(\mathbf{W}_{-jk})$, and some measure of the baseline covariates of other children, $\mathbf{h}_1(\mathbf{X}_{-ijk})$, information on the classroom quality of other classrooms had there been no intervention, $G(\mathbf{M}_{-jk}(0))$, gives no information on the quality the child would have received under the intervention, $M_{jk}(1)$. This is a fairly strong assumption and it may not hold in many settings. Using twin network diagrams (Pearl, 2009), it can be shown that this assumption would be violated if an intervention to change classroom quality in one class would also have an effect on classroom quality in other classrooms. Theorem 3 still applies if the independence assumptions (1), (2) and (8) hold only in mean rather than in distribution; independence assumptions (7) and (12), however, must in general hold in distribution.

As noted above, assumption (12) requires that an intervention to change the mediator (classroom quality) in one class not affect the mediator (classroom quality) in other classrooms. Essentially, this would only hold if there is no spillover effect of an intervention on the mediator to the mediator value of other classrooms. Note that this would still be compatible with there being a spillover effect of the mediator in one class on the outcomes of children in another class. The former spillover concerns spillover of an intervention on the mediator; the latter concerns spillover of the mediator on the outcome. In the substantive context of the 4Rs study, the assumption that there is no spillover effect of an intervention on the mediator (classroom quality) in one classroom to the mediator values of other classrooms would effectively require that the teachers from different classrooms in the same school do not interact with one another by way of sharing information on teaching techniques, classroom management and so forth. In this educational context, this assumption will likely not be plausible. A context in which the assumption may be more plausible would be if the randomized treatment were a community-building grant administered at a district level, the mediator were new youth clubs at the neighborhood level, and the outcome were child delinquency behaviors. An intervention creating a youth club in one neighborhood arguably would not change the creation of a youth club in another neighborhood (i.e. no spillover of an intervention on the mediator to the mediators of other neighborhoods) even though there may well still be spillover effects of the mediator of one neighborhood on the outcomes of individual children in other neighborhoods because children often form peer groups across neighborhood boundaries.

If the assumptions of Theorem 3 are satisfied then it may be applied to obtain empirical expressions for spillover mediated effects and within-classroom mediated effects. Under randomization and assumptions (2), (8) and (12), the within-classroom mediated effect described in the previous section is given by $E[Y_{ijk}(1, M_{jk}(1), G(\mathbf{M}_{-jk}(0))) - Y_{ijk}(1, M_{jk}(0), G(\mathbf{M}_{-jk}(0)))] =$

$$\begin{aligned}\sum_{\mathbf{c}}\sum_{m}\sum_{g} & E\left[Y_{ijk}|T_k=1, M_{jk}=m, G\left(\mathbf{M}_{-jk}\right)=g, \mathbf{C}_{ijk}=\mathbf{c}\right] \\ & \times\left\{P\left(M_{jk}=m|\, T_k=1, \mathbf{C}_{ijk}=\mathbf{c}\right)\right. \\ & \left.-P\left(M_{jk}=m|\, T_k=0, \mathbf{C}_{ijk}=\mathbf{c}\right)\right\} \\ & \times P\left(G\left(\mathbf{M}_{-jk}\right)=g|\, T_k\right. \\ & \left.=0, \mathbf{C}_{ijk}\right. \\ & \left.=\mathbf{c}\right) P\left(\mathbf{C}_{ijk}=\mathbf{c}\right)\end{aligned} \tag{14}$$

and the spillover mediated effect described in the previous section is given by $E[Y_{ijk}(1, M_{jk}(1), G(\mathbf{M}_{-jk}(1))) - Y_{ijk}(1, M_{jk}(1), G(\mathbf{M}_{-jk}(0)))]$

$$\sum_{\mathbf{c}}\sum_{m}\sum_{g}E\left[Y_{ijk}|T_k=1, M_{jk}=m, G\left(\mathbf{M}_{-jk}\right)=g, \mathbf{C}_{ijk}=\mathbf{c}\right] \times P\left(M_{jk}=m|\, T_k=1, \mathbf{C}_{ijk}=\mathbf{c}\right)\left\{P\left(G\left(\mathbf{M}_{-jk}\right)=g|\, T_k=1, \mathbf{C}_{ijk}=\mathbf{c}\right)-P\left(G\left(\mathbf{M}_{-jk}\right)=g|\, T_k=0, \mathbf{C}_{ijk}=\mathbf{c}\right)\right\}P\left(\mathbf{C}_{ijk}=\mathbf{c}\right). \quad (15)$$

3.4 Consequences of Ignoring Interference

Our final results are concerned with the consequences of ignoring interference and the conditions under which interference can be ignored while still obtaining direct and indirect effect estimates with a meaningful causal interpretation. When interference is absent so that $Y_{ijk}(t, m, g) = Y_{ijk}(t, m)$ for all t, m, g then the natural indirect effect would simplify to $E[Y_{ijk}(t, M_{jk}(t))]–E[Y_{ijk}(t, M_{jk}(t^*))]$ which, under assumptions (1), (2), (7), (8) with $Y(t, m, g) = Y(t, m)$, one could estimate by

$$\sum_{\mathbf{c}}\sum_{m}E\left[Y_{ijk}|T_k=t, M_{jk}=m, \mathbf{C}_{ijk}=\mathbf{c}\right]P\left(M_{jk}=m|T_k=t, \mathbf{C}_{ijk}=\mathbf{c}\right)P\left(\mathbf{C}_{ijk}=\mathbf{c}\right) -\sum_{\mathbf{c}}\sum_{m}E\left[Y_{ijk}|T_k=t, M_{jk}=m, \mathbf{C}_{ijk}=\mathbf{c}\right]P\left(M_{jk}=m|T_k=t^*, \mathbf{C}_{ijk}=\mathbf{c}\right)P\left(\mathbf{C}_{ijk}=\mathbf{c}\right). \quad (16)$$

Likewise, without interference the natural direct effect reduces to $E[Y_{ijk}(t, M_{jk}(t^*))]–E[Y_{ijk}(t^*, M_{jk}(t^*))]$ which under assumptions (1), (2), (7), (8) with $Y(t, m, g) = Y(t, m)$, one could estimate by

$$\sum_{\mathbf{c}}\sum_{m}E\left[Y_{ijk}|T_k=t, M_{jk}=m, \mathbf{C}_{ijk}=\mathbf{c}\right]P\left(M_{jk}=m|T_k=t^*, \mathbf{C}_{ijk}=\mathbf{c}\right)P\left(\mathbf{C}_{ijk}=\mathbf{c}\right) -\sum_{\mathbf{c}}\sum_{m}E\left[Y_{ijk}|T_k=t^*, M_{jk}=m, \mathbf{C}_{ijk}=\mathbf{c}\right]P\left(M_{jk}=m|T_k=t^*, \mathbf{C}_{ijk}=\mathbf{c}\right)P\left(\mathbf{C}_{ijk}=\mathbf{c}\right). \quad (17)$$

Note, we have in general by definition that $E[Y_{ijk}(t, M_{jk}(t^*))] \equiv E[Y_{ijk}(t, M_{jk}(t^*), G(\mathbf{M}_{-jk}(t)))]$. However, with interference present, the empirical quantity used in expressions (16) and (17) to estimate $E[Y_{ijk}(t, M_{jk}(t^*))]$, namely,

$$\sum_{\mathbf{c}}\sum_{m}E\left[Y_{ijk}|T_k=t, M_{jk}=m, \mathbf{C}_{ijk}=\mathbf{c}\right]P\left(M_{jk}=m|T_k=t^*, \mathbf{C}_{ijk}=\mathbf{c}\right)P\left(\mathbf{C}_{ijk}=\mathbf{c}\right)$$

will no longer necessarily equal $E[Y_{ijk}(t, M_{jk}(t^*))] \equiv E[Y_{ijk}(t, M_{jk}(t^*), G(\mathbf{M}_{-jk}(t)))]$ under assumptions (1), (2), (7) and (8) alone. Theorem 4 gives assumptions under which the equality will hold in the presence of interference. Theorem 4 will allow us to consider what the interpretation of the quantities in (16) and (17) in fact is when interference is present but is not taken into account.

Theorem 4. Suppose that (1), (2), (7), (8) and (12) hold (i.e. the conditions used to identify within-classroom mediated effects and spillover mediated effects) and suppose moreover that (12) holds not just for $t' \neq t^*$ but also for $t' = t^*$ i.e. we also require for all t,

$$M_{jk}(t) \perp\!\!\!\perp G\left(\mathbf{M}_{-jk}(t)\right)|\mathbf{C}_{ijk}. \quad (18)$$

then

$$\sum_{\mathbf{c}}\sum_{m} E\left[Y_{ijk}|T_k=t, M_{jk}=m, \mathbf{C}_{ijk}=\mathbf{c}\right] P\left(M_{jk}=m|T_k=t^*, \mathbf{C}_{ijk}=\mathbf{c}\right) P\left(\mathbf{C}_{ijk}=\mathbf{c}\right) = E\left[Y_{ijk}\left(t, M_{jk}\left(t^*\right), G\left(\mathbf{M}_{-jk}\left(t\right)\right)\right)\right].$$

It follows from Theorem 4 that if we attempted to estimate the natural indirect effect, ignoring interference, using (16) we would in fact obtain

$$E\left[Y_{ijk}\left(t, M_{jk}\left(t\right), G\left(\mathbf{M}_{-jk}\left(t\right)\right)\right)\right] - E\left[Y_{ijk}\left(t, M_{jk}\left(t^*\right), G\left(\mathbf{M}_{-jk}\left(t\right)\right)\right)\right]$$

i.e. a within-classroom mediated effect. It also follows immediately from Theorem 4 that if we attempted to estimate the natural direct effect, ignoring interference, using (17) we would in fact obtain

$$\begin{aligned}
&E\left[Y_{ijk}\left(t, M_{jk}\left(t^*\right), G\left(\mathbf{M}_{-jk}\left(t\right)\right)\right)\right] \\
&\quad - E\left[Y_{ijk}\left(t^*, M_{jk}\left(t^*\right), G\left(\mathbf{M}_{-jk}\left(t^*\right)\right)\right)\right] \\
&\quad = E\left[Y_{ijk}\left(t, M_{jk}\left(t^*\right), G\left(\mathbf{M}_{-jk}\left(t\right)\right)\right)\right] \\
&\quad - \left[Y_{ijk}\left(t, M_{jk}\left(t^*\right), G\left(\mathbf{M}_{-jk}\left(t^*\right)\right)\right)\right] \\
&\quad + \left\{E\left[Y_{ijk}\left(t, M_{jk}\left(t^*\right), G\left(\mathbf{M}_{-jk}\left(t^*\right)\right)\right)\right] - E\left[Y_{ijk}\left(t^*, M_{jk}\left(t^*\right), G\left(\mathbf{M}_{-jk}\left(t^*\right)\right)\right)\right]\right\}
\end{aligned}$$

i.e. the sum of a spillover mediated effect and the actual natural direct effect.

Intuitively, this result concerning the natural direct effect is not so surprising. If part of the effect of treatment on the outcome is mediated by the classroom quality in other classrooms and we ignore this and only consider the effect mediated through a child's own classroom then the direct effect will then capture both the true natural direct effect, $E[Y_{ijk}(t, M_{jk}(t^*), G(\mathbf{M}_{-jk}(t^*)))] - E[Y_{ijk}(t^*, M_{jk}(t^*), G(\mathbf{M}_{-jk}(t^*)))]$, and also the spillover mediated effect since the mediator being considered in the analysis ignoring interference is just the quality of the child's own classroom.

Several important points merit attention here. First, the assumptions under which this intuitive result holds are rather strong. To ignore interference in the way described above and still obtain meaningful causal interpretations of the estimators in (16) and (17), one makes all of the assumptions used to identify the within-classroom mediated effect and the spillover mediated effect but moreover one also assumes (18). Second, even if these assumptions hold, if the substantive question of interest is whether classroom quality mediates the effect of treatment and one uses the estimator in (16), this will essentially be an underestimate of the actual importance of classroom quality since it will not include an assessment of the effect mediated through the quality of other classrooms (which will instead be captured by the estimate of the natural direct effect ignoring interference). Third, if these assumptions, (1), (2), (7), (8), (12) and (18), do not hold then it is not clear that what is being estimated as a natural direct and indirect effect ignoring interference have any meaningful causal interpretation at all.

We have discussed the interpretation of assumptions (1), (2), (7), (8) and (12) above, we now consider assumption (18). Assumption (18) states that conditional on a child's baseline covariates, $\mathbf{X}_{ijk}$, certain characteristics of the school and teachers, $\mathbf{W}_{jk}$, $\mathbf{V}_k$, $\mathbf{h}_2(\mathbf{W}_{-jk})$, and perhaps some measure of the baseline covariates of other children, $\mathbf{h}_1(\mathbf{X}_{-ijk})$, information on the classroom quality of the other classrooms under treatment t, i.e. $G(\mathbf{M}_{-jk}(t))$, gives no

information on the classroom quality of the child's own classroom also under treatment t, $M_{jk}(t)$, beyond that of the baseline covariates, $\mathbf{C}_{ijk} = (\mathbf{X}_{ijk}, \mathbf{W}_{jk}, \mathbf{V}_k, \mathbf{h}_1(\mathbf{X}_{-ijk}), \mathbf{h}_2(\mathbf{W}_{-jk}))$. This assumption could in some sense be interpreted as one of an absence of resource constraints and lack of communication among teachers within schools. A similar result to Theorem 4 also holds for controlled direct effects which we state as our next theorem.

Theorem 5. Suppose (1), (2), (7), (8) and (18) hold then the estimator for the controlled direct effect that would be used ignoring interference, namely,

$$E\left[Y_{ijk}|T_k=t, M_{jk}=m, \mathbf{C}_{ijk}=\mathbf{c}\right] - E\left[Y_{ijk}|T_k=t^*, M_{jk}=m, \mathbf{C}_{ijk}=\mathbf{c}\right]$$

in actual fact identifies

$$E\left[Y_{ijk}\left(t, m, G\left(\mathbf{M}_{-jk}(t)\right)\right)|\mathbf{C}_{ijk} - \mathbf{c}\right] - E\left[Y_{ijk}\left(t^*, m, G\left(\mathbf{M}_{-jk}(t^*)\right)\right)|\mathbf{C}_{ijk}=\mathbf{c}\right].$$

Theorem 5's interpretation is essentially that under assumptions (1), (2), (7), (8) and (18), if we attempt to proceed with estimating controlled direct effects ignoring interference then what we think we are estimating as the direct effect not through the mediator will in fact also pick up the effect that is mediated through classroom quality in classrooms other than the child's own.

As already noted, the assumptions employed in Theorems 4 and 5 which allow one to ignore interference are even stronger than the assumptions used to identify the effects allowing for interference. We finally also note that if interference is ignored for the treatment then the covariate aggregates for other children and classrooms, namely $\mathbf{h}_1(\mathbf{X}_{-ijk})$, $\mathbf{h}_2(\mathbf{W}_{-jk})$ will likely also be ignored in which case (1), (2), (7), (8), (12) and (18) would all have to hold without conditioning on $\mathbf{h}_1(\mathbf{X}_{-ijk})$ and $\mathbf{h}_2(\mathbf{W}_{-jk})$.

In summary, (1) and (7) will hold under randomization. Under (1) and (2) we can identify the controlled direct effects of the treatment, classroom quality, and quality of other classrooms. Under (1), (2), (7) and (8) we can identify natural direct and indirect effects. Under (1), (2), (7), (8) and (12) we can furthermore identify the within-classroom mediated effect and the spillover mediated effect. Under (1), (2), (7), (8), (12) and (18), we could ignore interference and yet still identify the within-classroom mediated effect and also the sum of the natural direct effect and the spillover-mediated effect.

4. Analysis of the 4Rs Educational Intervention

4.1 Sample and Data

We apply the theoretical results in the previous section to a study of the 4Rs educational intervention. The study (Jones et al., 2010) involved a 3-year, 6 wave longitudinal experimental design with measurements in the fall and spring semester of each year. The eighteen New York City elementary schools in the study were fairly representative of the demographic composition of New York City schools generally. The schools were pair matched at baseline to minimize within-pair multivariate distance based on twenty school characteristics including size, average reading achievement, race/ethnic composition, mobility/two-year stability, school lunch receipt, expenditures, attendance and organizational readiness. Within each pair, schools were randomly assigned to either the 4Rs treatment or the control group. The intervention is implemented school-wide, grades K-6 for 3 years; all 3rd grade children in each school were followed over three years through 5th grade. In the application here we will consider the first year of the study for the children

beginning in third grade. The sample includes 82 3rd grade classrooms and 942 children. Schools in both treatment and control conditions had an average of 3 classrooms per grade level, ranging from smaller schools in both groups with between 2 and 3 classrooms in each grade level to larger schools such as a school with between 4 and 9 classrooms in the treatment group and a school with between 5 and 7 classrooms in the control group.

Classroom quality was measured using the CLASS scoring system (Pianta, La Paro, and Hamre, 2005) which assesses instructional support, emotional support and organizational climate with an overall score between 1 and 7. The internal reliability for this scale was 0.93. The measure assesses the quality of the interactions among students and teachers through which students are afforded opportunities to experience positive connections to their peers and teachers in well-regulated, organized classroom settings. Note that this is a measure of classroom quality and cannot in any sense be taken as "true" classroom quality. Our analyses must therefore be interpreted with respect to "measured classroom quality" as the mediator. In prior research (Buchinal et al., 2010), the overall measure of class quality showed a threshold effect on child outcomes. We observed a similar threshold effect at a cut-off of approximately 4.4. Hence, for simplicity in this example, the quality of a child's classroom mediator, M_{jk}, was dichotomized so that $M_{jk} = 1$ if the class measure was greater than or equal to 4.4 and $M_{jk} = 0$ otherwise. Likewise, the average quality of classrooms other than the child's own, $G_{jk} = G(M_{-jk})$, was dichotomized with respect to the same threshold: we define $G_{jk} = 1$ if the average score was greater than or equal to 4.4 and $G_{jk} = 0$ otherwise. Table 1 shows the number of classes in each M-by-G category within each treatment group. Covariate overlap was assessed for the covariates by strata defined by T and by dichotomized values of M and G.

The outcome was child depressive symptoms scored on a scale of 0 to 1. Child level covariates included child race/ethnicity, gender, baseline depression and baseline literacy skills; teacher covariates included teacher race/ethnicity and baseline confi-dence in behavior management. Because randomization to the 4Rs program occurred within matched pairs of schools, the models also included pair fixed effects to control for pair specific factors. Table 2 lists the distribution of the outcome and of each covariate by treatment and class quality. Unsurprisingly, children attending low-quality classes in control schools displayed the highest level of depressive symptoms at the end of the treatment year.

4.2 Assumptions

In attempting to draw causal conclusions, a critical evaluation of the assumption is crucial in assessing the validity of the study results. In this sub-section we consider the assumptions employed in our analysis of the 4Rs data and relate the assumptions considered in section 3 to the 4Rs study. As the treatment is randomized, assumption (1) holds; assumption (7) would likewise hold by randomization. By controlling for various child, teacher and school level covariates we hope to render assumption (2) somewhat plausible so that the effects of the quality in a child's own classroom and the effect of the quality of classrooms other than the child's own are unconfounded. Classroom quality arises from a combination of the particular teacher and the particular students in that class. Controlling for the afore-mentioned child and teacher characteristics within matched pairs of schools removes a considerable amount of bias associated with selection into different class quality levels and helps render assumption (2) somewhat plausible. Conditioning on these observed covariates, if the assignment of students to teachers within a school is unsystematic, as indicated by evidence from field reports in the 4Rs study, this would also render assumption (2) somewhat more plausible. However, we are unaware of any comprehensive study on the dynamics of the assignment process. Under assumptions (1) and (2) we can estimate the controlled direct effects of the 4Rs intervention, the classroom quality mediator, and the spillover controlled direct effect of classroom quality in other classroom.

To carry out the effect decomposition above so as to decompose a total effect into a natural direct effect, a within-classroom mediated effect and spillover mediated effect, we would also need to employ assumptions (8) and (12). Assumption (8) was that there were no mediator-outcome confounders affected by treatment; assumption (8) may be somewhat plausible insofar as we have used as our mediator the earliest available measurements of classroom quality after treatment was administered (cf. VanderWeele and Vansteelandt, 2009), but it may be contestable in this application in that class quality was measured in the second half, rather than the first half, of the school year. Assumption (12) is also problematic in the 4Rs study. Assumption (12) would require that an intervention to change classroom quality in one class must not affect classroom quality in other classrooms, i.e., that there is no spillover effect of an intervention on the mediator to the mediator value of other classrooms. This assumption is compatible with there being a spillover effect of the mediator in one class on the outcomes of children in other classes. In the context of the 4Rs study, this assumption would effectively require that the teachers in each school do not interact with one another by way of sharing information on teaching techniques, classroom management, etc. Because teacher autonomy has been the norm rather than the exception in most US schools (Lortie, 1975), it might be argued that, at least for the control schools, a lack of communication among teachers may prevail. However, given that the 4Rs program was designed deliberately to enhance the professional skills of all teachers within a school in part through collaboration among teachers, spillover of the quality of one class to the quality of other classes arguably would occur in a 4Rs school. Assumption (12) may not be plausible in this application.

Because of the problematic nature of both assumptions (8) and (12), we do not in our application proceed with the estimation of natural direct effects, within-classroom mediated effects and spillover mediated effects. We restrict the analysis to the estimation of the controlled direct effects of the 4Rs intervention and of the classroom quality mediator, and to the spillover controlled direct effect of classroom quality in other classrooms. Identification and estimation of these effects only require assumptions (1) and (2).

4.3 Models and Estimation

In section 3, we described how, under certain assumptions, the various causal effects of interest were identified in terms of conditional expectations. Here we will describe a modeling approach for estimating the controlled direct effects of the treatment, of one's own class quality as a mediator, and of the quality of other classes in the same school as a concurrent mediator. In the online supplement, we discuss a potential modeling approach for estimating natural direct and indirect effects and within-classroom and spillover mediated effects when these effects are identified. Appropriate modeling will depend in part on the study design, on the data available, and on the causal quantities of interest.

With data from this group randomized trial, we may estimate the total effect of the 4Rs program on child depressive symptoms through analyzing a multilevel model reflecting the data structure in which third-graders were nested within classes that were in turn nested within schools:

$$Y_{ijk}=\alpha+\beta T_k+\gamma f\left(\mathbf{C}_{ijk}\right)+\epsilon_k+v_{jk}+u_{ijk} \quad \text{with} \quad \epsilon_k \sim N(0,\psi), v_{jk} \sim N(0,\tau), u_{ijk} \sim N\left(0,\sigma^2\right). \quad (19)$$

The model includes a school-specific random component ε_k, a class-specific random component v_{jk}, and a student-specific random component u_{ijk}. These are assumed normally distributed and mutually independent. Because assumption (1) holds under randomization, treatment assignment is independent of these random effects. In model (19), $f(\mathbf{C}_{ijk})$ is a function of covariates including child gender, race, baseline depressive symptoms and

baseline literacy skills, teacher race and baseline confidence in class management, along with fixed effects for matched school pairs. All covariates were centered at their sample means. To compare the 4Rs school and the control school within matched pairs, one might have alternatively included in the multilevel model random effects for matched pairs and then centered all the predictors at their respective means within matched pairs. This would be equivalent to our model using pair fixed effects (Raudenbush, 2009).

Preliminary examination of the data revealed nonlinear and non-additive associations between treatment and teacher baseline confidence in class management in predicting child depressive symptoms. Teacher baseline confidence was measured on a 1 – 7 scale. We identified a cutoff at 5 and specified different slopes below and above the cutoff under each treatment. This was the only covariate for which there was clear evidence of different slopes (details available from the authors upon request). Substantive results were, however, generally similar even when not allowing for differing slopes for high and low teacher baseline confidence. Parameter values and model-based standard errors were estimated via maximum likelihood in HLM 7.0. The sample size was insufficient to consider robust standard errors.

Likewise, to estimate the effect of the 4Rs program on class quality measured on a continuous scale, we specified a two level model with classes nested within schools:

$$M_{jk}+\mu+\kappa T_k+\lambda f\left(\mathbf{C}_{jk}\right)+\epsilon_k+v_{jk} \quad \text{with} \quad \epsilon_k \sim N(0,\phi), v_{jk} \sim N(0,\zeta). \quad (20)$$

Here ε_k and v_{jk} are assumed normally distributed and mutually independent. Given the randomization of treatment within matched pair of schools, treatment assignment is independent of these random effects. In model (20), $f(\mathbf{C}_{jk})$ is a function of covariates including the aforementioned teacher baseline covariates and pair fixed effects.

Finally we fit a multilevel model for the effects of treatment, classroom quality, and quality of other classrooms on child depressive symptoms with the interactions between these variables saturated:

$$\begin{aligned} Y_{ijk}=&\alpha_{00}I\left(M_{jk}=0,G_{jk}=0\right)\\ &+\alpha_{01}I\left(M_{jk}=0,G_{jk}=1\right)\\ &+\alpha_{10}I\left(M_{jk}=1,G_{jk}=0\right)\\ &+\alpha_{11}I\left(M_{jk}=1,G_{jk}=1\right)\\ &+\beta_{00}T_kI\left(M_{jk}=0,G_{jk}=0\right)\\ &+\beta_{01}T_kI\left(M_{jk}=0,G_{jk}=1\right)\\ &+\beta_{10}T_kI\left(M_{jk}=1,G_{jk}=0\right)\\ &+\beta_{11}T_kI\left(M_{jk}=1,G_{jk}=1\right)\\ &+\gamma f\left(\mathbf{C}_{ijk}\right)\\ &+\epsilon_k+v_{jk}+u_{ijk} \quad \text{with} \quad \epsilon_k \sim N(0,\psi), v_{jk} \sim N(0,\tau), u_{ijk} \sim N\left(0,\sigma^2\right). \end{aligned} \quad (21)$$

Here ε_k, v_{jk} and u_{ijk} are assumed normally distributed and mutually independent and $f(\mathbf{C}_{ijk})$ is a function of the aforementioned child and teacher baseline covariates and pair fixed effects. Under identification assumptions (1) and (2), the treatment indicator, class quality indicators, and their interactions are assumed independent of these random effects after adjustment for $f(\mathbf{C}_{ijk})$; $I(M_{jk} = m, G_{jk} = g)$ are indicator function that take value 1 if $M_{jk} = m$

and $G_{jk} = g$ and 0 otherwise. The class quality indicators could be associated with the random error terms if the identification assumptions do not hold. For example, after statistical adjustment for the covariates, if class quality is generally higher in schools in which students entering third grade with a lower level of depressive symptoms on average, the effects of class quality would be estimated with bias. This is unlikely to be a problem here given that our covariate list has included individual student's baseline depressive symptoms. Unlike most conventional methods that assume no treatment-mediator interaction, we explicitly model the direct effect of the 4Rs program as a function of one's own class quality and of the quality of other classes. We checked covariate overlap for each of the variables in $\mathbf{C}_{ijk}$ by strata defined by T_k, M_{jk}, G_{jk} and found reasonably good covariate overlap. Details available upon request from the authors. For simplicity, the model assumes that the direct effect does not depend on the covariate values. This assumption could also be relaxed.

In model (21), under assumptions (1) and (2), α_{mg} correspond to mean potential outcomes $E[Y_{ijk}(0, m, g)]$; also $\alpha_{mg} + \beta_{mg}$ correspond to mean potential outcomes $E[Y_{ijk}(1, m, g)]$. The coefficients β_{mg} correspond to the controlled direct effects of treatment, $E[Y_{ijk}(1, m, g) - Y_{ijk}(0, m, g)]$. The controlled direct effects of the quality of an individual's own classroom, $E[Y_{ijk}(t, 1, g) - Y_{ijk}(t, 0, g)]$, are $\alpha_{1g} - \alpha_{0g}$ when $t = 0$ and $(\alpha_{1g} + \beta_{1g}) (\alpha_{0g} + \beta_{0g})$ when $t = 1$, the controlled direct effects of the quality of other classrooms, $E[Y_{ijk}(t, m, 1) - Y_{ijk}(t, m, 0)]$, are $\alpha_{m1} - \alpha_{m0}$ when $t = 0$ and $(\alpha_{m1} + \beta_{m1}) - (\alpha_{m0} + \beta_{m0})$ when $t = 1$. The parameters from the model in (22), saturated for T_k, M_{jk}, G_{jk}, directly maps onto the controlled direct effects of interest. This mapping can be used as a guide for researchers interested in these effects. We estimate these parameters and effects with the 4Rs data below.

Note that the conventional analysis for mediation would replace model (21) with the following:

$$Y_{ijk} = \alpha + \beta T_k + \delta M_{jk} + \gamma f\left(\mathbf{C}_{ijk}\right) + \epsilon_k + v_{jk} + u_{ijk} \quad \text{with} \quad \epsilon_k \sim N(0, \psi), v_{jk} \sim N(0, \tau), u_{ijk} \sim N\left(0, \sigma^2\right). \quad (22)$$

where ε_k, v_{jk} and u_{ijk} are again assumed normally distributed and mutually independent, ignoring interactions and ignoring potential interference. The conventional analysis would moreover often also ignore control for the covariate vector $f(\mathbf{C}_{ijk})$. The conventional approach would take β in model (22) as the direct effect and the difference between β in model (22) and β in model (19) as the indirect or mediated effect. As noted above, for linear models, this “difference method” for the indirect effect gives the same results as would a “product-of-coefficients” method (MacKinnon, 2008). For comparison, we will present also the results of this conventional analysis using model (22).

4.4 Results

Fitting model (19) to the data from the 4Rs study, we obtained an estimated total treatment effect of the 4Rs intervention on depressive symptom outcomes of –0.052 ($se = 0.022$, $t = -2.305$, $p = 0.05$), which suggests a marginally significant effect of the treatment in reducing child depressive symptoms. Likewise, fitting model (20) to the data from the 4Rs study, we obtained an estimated treatment effect of the 4Rs intervention on classroom quality of 0.45 ($se = 0.20$, $t = 2.28$, $p = 0.05$), indicating an improvement in classroom quality due to the 4Rs intervention.

Fitting model (21), we obtained estimates of the average controlled direct effects of T, M, and G, these results are summarized in Table 3. We applied t tests to the controlled direct effects of T and chi square tests with 1 degree of freedom each to the controlled direct effects of M and G. Among the four possible combinations of m and g, the 4Rs program

showed a statistically significant direct effect on child depressive symptoms only when a child attended a low-quality class surrounded by high-quality classes (i.e., β_{01}). None of the controlled direct effects of M and of G appear to be different from zero; however, our sample size may be too small to detect these.

We tested the null hypothesis of no interaction of treatment with either a child's own classroom quality or the quality of other classrooms, i.e. $\beta_{00} = \beta_{01} = \beta_{10} = \beta_{11}$. The chi square test result indicated that we could reject the null hypothesis ($\chi^2 = 10.28$, $df = 4$, $p = 0.035$). Furthermore, according to the result of a likelihood ratio test, a parsimonious model constraining $\beta_{00} = \beta_{10} = \beta_{11}$ appeared to be equivalent to model (25). The parsimonious model gave an estimate of the average controlled direct effect of treatment –0.022 (*s.e.* = 0.029, $t = 0.743$, $p = 0.461$) for children whose own class quality and the quality of other classes in school were both low or both high or one's own class quality was high while the quality of other classes was low. The estimated average controlled direct effect of treatment for children whose own class quality was low while the quality of other classes in school was high remained statistically significant –0.154 (*s.e.* = 0.051, $t = –3.006$, $p = 0.004$). For these students, without changing class quality, simply being under the 4Rs program seemed to have the potential of reducing child depressive symptoms by about 60 percent of a standard deviation. The 95% confidence interval for the effect size ranged from about 22 percent to 98 percent of a standard deviation.

We now consider the results of fitting model (22) to the 4Rs data using the conventional approach to mediation analysis, and ignoring potential interference and potential treatment-mediator interaction. With the conventional analysis, fitting model (22) the direct effect of the treatment would have appeared to be –0.051 (*se* = 0.023, $t = –2.233$, $p = 0.056$), suggesting a marginally significant beneficial effect for reducing child depressive symptoms regardless of class quality. This “direct effect”, –0.051, is very similar to the estimate of the total effect reported above, –0.052, and so the conclusion of the conventional approach would be that almost none of the effect of the 4Rs intervention on depressive symptoms is mediated by classroom quality. Note, however, that even under the best case scenario, that all of the assumptions of Theorem 5 hold, this “direct effect” under the conventional approach would still capture part of the effect of the 4Rs intervention that is mediated through classroom quality in classrooms other than the child's own; in the presence of spillover, without the assumptions of Theorem 5, the estimate under the conventional approach does not have a clear causal interpretation. The conventional approach misses the potential interaction and spillover that was suggested by our analyses above.

5. Concluding Remarks

In this paper we have developed an approach for analyzing spillover effects in mediation analysis with data from group randomized trials. By relaxing the stable unit treatment value assumption (SUTVA) typical in causal inference, we have been able to precisely define spillover effects that will often be of substantive and theoretical interest. The 4Rs program evaluation presented in this paper has provided an important case study in which interference is not simply a problem that must be addressed but in fact may be an important pathway mediating the treatment effect on individual outcomes. By taking into account potential interference among students from different classrooms in the same school, we were able to reveal that, for children attending low-quality classes in schools in which other classes have relatively high quality, the 4Rs intervention could possibly reduce child depressive symptoms through means other than improving class quality.

We have presented theoretical results for identifying the controlled direct effects of the treatment, of a student's own class quality as a mediator, and of the quality of other classes

in the same school as a concurrent mediator. In addition, the paper has provided assumptions for identifying natural direct and indirect effects, within-classroom mediated effects, and spillover mediated effects. The identification assumptions for these latter effects are particularly strong and the estimation of these effects was thus not pursued in our analysis of the 4Rs data. We hope that by presenting the assumptions, researchers will be more aware of what is required in the design and analysis studies which seek to estimate these various effects. For example, assumption (8) might have been more plausible had the mediator been measured much earlier in the year.

However, even for the controlled direct effects of the treatment, the mediator, and the value of the mediator in other classrooms, the approach that we have described here constitutes an important advance over the conventional approach to analyzing direct and indirect effects that is often used in group-randomized trials. This is because our approach (i) allows an individual's potential outcomes to depend on one's own mediator and on other individuals' mediator as well as on the treatment (i.e. for interference), (ii) accommodates possible treatment-by-mediator interactions, (iii) disentangles the causal effects of interest using a model that is saturated for the treatment, mediator and mediator of other classrooms, and (iv) explicates the no-unmeasured-confounding assumptions required for identification. We have also importantly documented the consequences of ignoring interference when it is present. We saw that even in a best case scenario under which our various identification assumptions hold, estimates of direct effects ignoring interference actually also capture effects through the mediator values in classrooms other than a child's own classroom.

Our analysis of the 4Rs data made some strong assumptions and is therefore subject to some important limitations. First, the no-unmeasured-confounding assumptions required for identification are quite strong, we have tried to make them more plausible by conditioning on relevant pretreatment covariates. Further research could also develop sensitivity analysis techniques to assess the extent to which an unobserved variable affecting both the mediator and the outcome might invalidate inference about direct, indirect, and spillover effects. Such unobserved variables might give rise to confounding of the effects related to both the quality of a child's own classroom and that of the quality of other classrooms. Second, we have made the simplifying assumption in the current application that spillover effects depend on a mean scalar function; the assumption is not necessary in principal. However, without a very large dataset it would be difficult to make progress without some assumption on the form of interference. Future work could consider using friendship network data to inform make alternative assumptions on the form of interference. Third, we have used a measurement of classroom quality; our analyses need to be interpreted with respect to measured classroom quality as the mediator; future work could consider consequences of measurement error.

The analyses here have important implications for discerning in what contexts the 4Rs intervention is most effective. We saw a relatively large controlled direct effect of the 4Rs program when the child's own classroom quality is low in a school in which the overall classroom quality of other classrooms is relatively high. This direct effect of the treatment may be mediated by other variables, possibly related to conflict resolution or reduction in bullying among children in the hallways or on the playground, raising new questions for further investigation. The analysis and results here are also important in that they provide a template and framework for other such analyses in education research and the social sciences more broadly. In group-randomized trials, interference and spillover effects are often important to consider for understanding the causal mechanisms of a group-level intervention.

Acknowledgments

The authors thank Stephen Raudenbush for his support and helpful comments on the manuscript. Tyler J. VanderWeele was supported by NIH grants R03 HD060696 and R01 ES017876. Guanglei Hong was supported by the Spencer Foundation and a Scholars Award from the William T. Grant Foundation. The evaluation of the 4Rs Program, conducted as the New York City Study of Social and Literacy Development is supported by grants from the U.S. Department of Education in collaboration with the Centers for Disease Control (R305L030003) and the William T. Grant Foundation (# 2618) to J. Lawrence Aber (PI), Stephanie Jones and Joshua Brown (co-PIs) and by a grant from the William T. Grant Foundation (#7520) to Stephanie Jones and Joshua Brown (co-PIs).

References

Brown JL, Jones SM, LaRusso M, Aber JL. Improving classroom quality: teacher influences and experimental impacts of the 4Rs program. Journal of Educational Psychology. 2010; 102(1):153–167.

Buchinal M, Vandergrift N, Pianta R, Mashburn A. Threshold Analysis of Association between Child Care Quality and Child Outcomes for Low-Income Children in Pre-Kindergarten Programs. Early Childhood Research Quartlerly. in press.

Cox, DR. Planning of Experiments. John Wiley & Sons; New York: 1958.

Hong, G. 2010 Proceedings of the American Statistical Association, Biometrics Section. American Statistical Association; Alexandria, VA: 2010. Ratio of mediator probability weighting for estimating natural direct and indirect effects.; p. 2401-2415.

Hong G, Raudenbush SW. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. Journal of the American Statistical Association. 2006; 101:901–910.

Hudgens MG, Halloran ME. Towards causal inference with interference. Journal of the American Statistical Association. 2008; 103:832–842. [PubMed: 19081744]

Jones SM, Brown JL, Aber JL. The longitudinal impact of a universal school-based social-emotional and literacy intervention: an experiment in translational developmental research. Child Development: Special Issue on Raising Healthy Children: Translating Child Development Research into Practice. 2010 in press.

Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. Evaluation Review. 1981; 5:602–619.

Lortie, D. Schoolteacher: a Sociological Study. University of Chicago Press; Chicago: 1975.

MacKinnon, DP. An Introduction to Statistical Mediation Analysis. Lawrence Erlbaum Associates; New York: 2008.

Pearl, J. Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence. Morgan Kaufmann; San Francisco: 2001. Direct and indirect effects.; p. 411-420.

Pearl, J. Causality: Models, Reasoning, and Inference. 2nd edition. Cambridge University Press; Cambridge: 2009.

Peterson ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. Epidemiology. 2006; 17:276–84. [PubMed: 16617276]

Pianta, RC.; La Paro, K.; Hamre, B. Unpublished measure. University of Virginia; Charlottesville, VA: 2005. Classroom Assessment Scoring System (CLASS)..

Raudenbush SW. Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. Journal of Education Finance and Policy. 2009; 4:468–491.

Robins, JM. Semantics of causal DAG models and the identification of direct and indirect effects.. In: Green, P.; Hjort, NL.; Richardson, S., editors. Highly Structured Stochastic Systems. Oxford University Press; New York: 2003. p. 70-81.

Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992; 3:143–155. [PubMed: 1576220]

Rosenbaum PR. Interference between units in randomized experiments. Journal of the American Statistical Association. 2007; 102:191–200.

Rubin DB. Comment on: "Randomization analysis of experimental data in the fisher randomization test" by D.Basu. Journal of the American Statistical Association. 1980; 75:591–593.

Rubin DB. Which ifs have causal answers? Comment on: "Statistics and causal inference" by P. Holland. Journal of the American Statistical Association. 1986; 81:961–962.

Sobel ME. What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. Journal of the American Statistical Association. 2006; 101:1398–1407.

Tchetgen Tchetgen EJ, VanderWeele TJ. On causal inference in the presence of interference. Statistical Methods in Medical Research – Special Issue on Causal Inference. 2011 in press.

van der Laan MJ, Petersen ML. Direct effect models. International Journal of Biostatistics. 2008; 4 Article 23.

VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiology. 2009; 20:18–26. [PubMed: 19234398]

VanderWeele TJ. Direct and indirect effects for neighborhood-based clustered and longitudinal data. Sociological Reserach and Methods. 2010; 38:515–544.

Table 1

Sample of Classes in M-by-G Categories in Each Treatment Group

	T=1	T=0
M=0, G=0	12	17
M=0, G=1	7	6
M=1, G=0	9	13
M=1, G=1	17	1
Total	45	37

Table 2

Distributions of Outcome and Covariates by Treatment and Class Quality

	T=1				T=0			
	M=1		M=0		M=1		M=0	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Child Outcome: Depressive Symptoms	0.49	0.27	0.46	0.25	0.48	0.26	0.54	0.25
Child Baseline Depressive Symptoms	0.48	0.26	0.51	0.25	0.50	0.27	0.47	0.20
Child Baseline Literacy Skills	3.22	1.06	3.06	1.01	3.03	1.03	3.24	1.14
Child Gender (0=boy; 1=girl)	0.54	---	0.49	---	0.53	---	0.49	---
Child Hispanic	0.48	---	0.41	---	0.54	---	0.44	---
Child Black	0.38	---	0.47	---	0.29	---	0.43	---
Child Other Non-White Racial Identity	0.09	---	0.08	---	0.13	---	0.07	---
Teacher Baseline Confidence	6.02	0.90	4.99	1.84	5.89	0.94	5.43	1.47
Teacher Hispanic	0.19	---	0.22	---	0.00	---	0.13	---
Teacher Black	0.19	---	0.39	---	0.21	---	0.30	---

Table 3

Average Controlled Direct Effects of Treatment (*T*), Quality of One's Own Class (*M*), and Quality of Other Classes in School (*G*)

Controlled Direct Effect of T			**Coefficient**	**s.e.**	***t***	***p***
	M=0, G=0	β_{00}	-0.035	0.046	-0.762	0.450
	M=0, G=1	β_{01}	-0.154	0.051	-2.980	0.004
	M=1, G=0	β_{10}	-0.022	0.046	-0.470	0.640
	M=1, G=1	β_{11}	0.016	0.076	0.211	0.834
Controlled Direct Effect of M			Coefficient	s.e.	χ^2	*p*
	T=0, G=0	$\alpha_{10} - \alpha_{00}$	0.019	0.047	0.168	>0.500
	T=0, G=1	$\alpha_{11} - \alpha_{01}$	-0.118	0.086	1.874	0.167
	T=1, G=0	$(\alpha_{10} + \beta_{10}) - (\alpha_{00} + \beta_{00})$	0.033	0.047	0.480	>0.500
	T=1, G=1	$(\alpha_{11} + \beta_{11}) - (\alpha_{01} + \beta_{01})$	0.053	0.046	1.319	0.249
Controlled Direct Effect of G						
	T=0, M=0	$\alpha_{01} - \alpha_{00}$	0.085	0.056	2.345	0.122
	T=0, M=1	$\alpha_{11} - \alpha_{10}$	-0.052	0.083	0.392	>0.500
	T=1, M=0	$(\alpha_{01} + \beta_{01}) - (\alpha_{00} + \beta_{00})$	-0.034	0.057	0.359	>0.500
	T=1, M=1	$(\alpha_{11} + \beta_{11}) - (\alpha_{10} + \beta_{10})$	-0.014	0.054	0.069	>0.500