Psychological Methods
2010, Vol. 15, No. 3, 250–267

1082-989X/10/$12.00 DOI: 10.1037/a0018719

The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies

Peter M. Steiner
Northwestern University and Institute for Advanced Studies

Thomas D. Cook
Northwestern University

William R. Shadish
University of California, Merced

M. H. Clark
Southern Illinois University Carbondale

The assumption of strongly ignorable treatment assignment is required for eliminating selection bias in observational studies. To meet this assumption, researchers often rely on a strategy of selecting covariates that they think will control for selection bias. Theory indicates that the most important covariates are those highly correlated with both the real selection process and the potential outcomes. However, when planning a study, it is rarely possible to identify such covariates with certainty. In this article, we report on an extensive reanalysis of a within-study comparison that contrasts a randomized experiment and a quasi-experiment. Various covariate sets were used to adjust for initial group differences in the quasi-experiment that was characterized by self-selection into treatment. The adjusted effect sizes were then compared with the experimental ones to identify which individual covariates, and which conceptually grouped sets of covariates, were responsible for the high degree of bias reduction achieved in the adjusted quasi-experiment. Such results provide strong clues about preferred strategies for identifying the covariates most likely to reduce bias when planning a study and when the true selection process is not known.

Keywords: strong ignorability, selection bias, hidden bias, within-study comparison, propensity score

When perfectly implemented, random assignment generates unbiased causal estimates because at pretest the treatment and control groups are equivalent on expectation over all possible covariates. However, random assignment cannot always be implemented, and in observational studies, differential selection processes can bias estimated treatment effects. This bias can be removed if all covariates determining the outcome are known (Steyer, Gabler, von Davier, & Nachtigall, 2000) or if the selection process is completely known, as with the perfectly implemented regression discontinuity design (Cook, 2008; Goldberger, 1972; Shadish, Cook, & Campbell, 2002). However, most other observational studies involve more complex treatment assignment processes characterized by some combination of self-, administrator, or other third-person selection. Such complexity raises doubt about how adequate statistical controls for selection can be because all the covariates simultaneously correlated with both treatment and the potential outcomes have to be fully observed for all the selection bias to be removed from the outcome measure and thus to claim that the "strong ignorability assumption" holds (Rosenbaum & Rubin, 1983). Without this assumption, or alternatively, similar other assumptions (Steyer et al., 2000), it is impossible to infer that all the selection bias has been removed from the estimated treatment effect.

Although the role of strong ignorability is clear in theory, in practice we rarely know whether the observed covariates suffice to justify the assumption. We do not even know if, or under which conditions, a practically important level of bias reduction can be achieved with whatever covariates are on hand from among all those imaginable—for instance, when demographic covariates are available or pretest measures of the main outcome, though we know the latter is usually better than the former (Glazerman, Levy, & Myers, 2003; Heckman, Ichimura, Smith, & Todd, 1998). In most observational studies, strong ignorability is assumed rather than directly tested, and the justifications offered to support the assumption are rarely convincing because there is usually no way to prove the absence of differential selection on unobserved covariates. Sensitivity analyses for hidden bias are available and sometimes undertaken (Rosenbaum, 2002); however, they cannot indicate whether such bias actually exists and is fully accounted for in the choice of covariates used to bound the causal estimate achieved.

Within-study comparisons can be used to explore which covariates reduce bias. Within-study comparisons contrast the effect

Peter M. Steiner, Institute for Policy Research, Northwestern University, and Institute for Advanced Studies, Vienna, Austria; Thomas D. Cook, Institute for Policy Research, Northwestern University; William R. Shadish, School of Social Sciences, Humanities, and Arts, University of California, Merced; M. H. Clark, Department of Psychology, Southern Illinois University Carbondale.

Peter M. Steiner, Thomas D. Cook, and William R. Shadish were supported in part by Institute for Educational Sciences, U.S. Department of Education Grant R305U070003. Peter M. Steiner was also supported by a grant from the Spencer Foundation. We thank Larry Hedges, Charles Manski, and Bruce Spencer for helpful discussions.

Correspondence concerning this article should be addressed to Peter M. Steiner, Institute for Policy Research, Northwestern University, 2040 Sheridan Road, Evanston, IL 60208. E-mail: p-steiner@northwestern.edu

sizes from a randomized experiment and an observational study, with the treatment group held constant. This design serves to focus attention on the contrast of maximal interest: whether control groups are formed at random. We call the randomized group the control group, and we call the nonrandomly formed group the comparison group. Covariates are used to adjust the comparison group data to see whether the resulting adjusted effect estimate is similar to the experimental effect estimate, the latter being the benchmark for the true causal effect. If they are similar, the conclusion is drawn that the selection mechanism in the observational study is ignorable given the covariates in the analysis.

A dozen within-study comparisons exist in the job training evaluation literature, and reviews of these comparisons have concluded that the selection models used there have almost always failed to remove all of the initial bias (Bloom, Michalopoulos, Hill, & Lei, 2002; Glazerman et al., 2003). The case is somewhat different in other social sciences (Cook, Shadish, & Wong, 2008) in which full bias reduction has been achieved in three carefully circumscribed contexts: when regression-discontinuity is used (Aiken, West, Schwalm, Carroll, & Hsuing, 1998; Black, Galdo, & Smith, 2007; Buddelmeyer & Skoufias, 2004), when full or extensive knowledge of the selection process is available (Diaz & Handa, 2006), and when intact and usually local but nonequivalent comparison groups are selected that heavily overlap with the treatment group on pretest measures of outcome (Aiken et al., 1998; Bloom et al., 2002; Diaz & Handa, 2006).

However, most past within-study comparisons entailed confounds, one of which stands out. The contrast between the control and comparison groups has often been correlated with third variables that could affect outcome, including variation in the geographical location of the control and comparison samples, in the time of their outcome measurement, and even in the content of each set of measures. Identifying confounds like these eventually led to designing better within-study comparisons than the original three-arm design with its treatment, control, and comparison groups. In the first example of a four-arm within-study comparison design, Shadish, Clark, and Steiner (2008) pretested participants in different domains before randomly assigning them to either a randomized experiment with two treatment arms (learning about vocabulary or mathematics) or to a quasi-experiment with self-selection into these same two arms. Participants in the randomly and nonrandomly formed groups experienced their vocabulary or mathematics treatment at the same sessions and were always tested in the same way at the same time so as to rule out third variable confounds.

Using this design, Shadish et al. (2008) were able to show that the bias due to self-selection was nearly completely reduced by a set of 23 constructs generated from theory and a common-sense appraisal of processes that might explain why individuals would self-select into the vocabulary or mathematics treatments. This result held whether they used propensity scores (PSs) based on all observed preintervention constructs or a simple analysis of covariance (ANCOVA) using the same constructs. Shadish et al. separately investigated the role only of demographic covariates, illustrating that they were not by themselves sufficient to reduce bias.

In this study, we decompose the 23 constructs into five homogeneous domains to examine their unique and joint influence on bias reduction. One domain involves individual-level demographic measures of the kind frequently available in both survey and quasi-experimental work. A second involves proxy-pretest measures of the outcomes. Shadish et al. (2008) measured two posttest outcomes: knowledge of mathematical exponents and a test of vocabulary. The proxy-pretest for the first was a measure of general math ability, and the proxy-pretest for the latter was a general test of verbal skills. The third construct domain involves prior academic performance in the same domain as the relevant knowledge outcomes—that is, GPA and SAT scores. The fourth involves motivational variables indicating liking for math and literature and preference for the one over the other. The final construct domain taps into general psychological dispositions, broad personality reasons that might explain why some students in the nonexperiment preferred math over vocabulary—for example, conscientiousness that leads to exposure to topics that might benefit one more even if they are not what one likes. A major purpose of this study is to identify just which individual domains, or sets of domains, are responsible for the bias reduction achieved in Shadish et al.'s study.

The yoked experiment and nonexperiment in Shadish et al. (2008) also allow us to explore the effectiveness of two different approaches to choosing covariates in observational studies. The first specifies in advance of data collection what one considers to be very good approximations to that part of the selection process that is correlated with outcome, given that it is never possible to know the selection process exactly. Subject matter experts, extant theories, and observations from pilot tests have to play a crucial role in this process of identifying the most important confounding variables. Here we use data from the within-study comparison of Shadish et al. to investigate whether the total bias reduction that they achieved with 23 constructs might have been possible with fewer, both with one or two of the five domains or even with one or two constructs from within the few domains that turn out to be important. At present we do not know how complex the "true" selection models are, and so we do not know how plausible it is to assume that they could be arrived at in study planning. It is important to identify whether selection bias can be successfully addressed at the level of domains or the constructs within them because we do not yet have much evidence on the issue. All the current evidence about bias reduction lumps all the domains and constructs together.

If we do identify single domains or constructs of such potency, this does not preclude other variables that are correlated with them from doing just as well when they are used together, even though they might not reduce much bias when used singly. The intuition here is that these proxies could together capture all or most of the selection bias in outcomes even when the best single proxies for selection are not available. This leads us to the second approach to choose covariates, which is based on sampling a large set of constructs from a very broad range of domains in the hope of bringing into the analysis more and more covariates that are nonredundantly correlated with both selection and potential outcomes. This approach is particularly relevant when researchers have only weak knowledge of the selection process while choosing covariates, when a rich set of covariates is available to them, and when they realize that the elusive true selection process will never be revealed to them. So the advice is to choose many covariate domains and constructs, even though some will turn out to be unrelated to selection or outcome, or only weakly related, or completely redundant with other covariates. The hope is to "some-how" capture all the important covariates under the assumption that it is generally impossible to know in advance that part of the true selection process that is related to outcome. Indeed, it might even be dangerous to think that one could know the selection process if this led to sampling only a narrow range of covariates.

The main purpose of this article is to conduct a secondary analysis of Shadish et al.'s (2008) data to identify (a) which covariate domains are responsible for the bias reduction achieved in Shadish et al.'s study, (b) which specific constructs within these domains most reduce bias, and (c) whether just as much bias reduction can be achieved when the most effective individual constructs for bias reduction are omitted. Answering these questions will identify after the fact those covariates that most likely measure the true selection process. However, because it is rare for researchers to know this in study planning and covariate choice, we use the results above to speculate whether researchers should select covariates by presuming they know true selection process or by consciously deciding to assess many covariate domains in hopes of capturing among them some that will reduce all or most of the selection bias.

Method

Data

We use data from a within-study comparison (Shadish et al., 2008) in which volunteer undergraduate students from introductory psychology classes at a large mid-south public university were randomly assigned to be in a randomized experiment ($N = 235$) or in a nonrandomized study ($N = 210$). The design is depicted in Figure 1. Students assigned to the randomized experiment were randomly assigned to mathematics training ($n = 119$) or to vocabulary training ($n = 116$); and those assigned to the nonrandomized experiment chose which training they wanted—79 students chose mathematics, and 131 chose vocabulary. All students then attended the same training sessions irrespective of whether they were assigned to it at random. Before the random assignment to the different assignment methods, all students were pretested in the same way with the same measures to ensure that all participants were treated identically except for assignment method. Note that pretesting before treatment assignment ensures that covariates are influenced neither by the assignment nor the treatment. Treatments consisted either of 50 advanced vocabulary terms or five algebraic concepts that were taught to students with a series of overhead transparencies. Shadish et al. (2008) compared two treatment conditions (rather than comparing treatment to no treatment) for two reasons: (a) Doing so created two independent effect estimates: one for vocabulary training and one for mathematics training; and (b) a no treatment control might have attracted a disproportionate number of participants to select what could have been the least time-consuming session in the nonrandomized experiment.

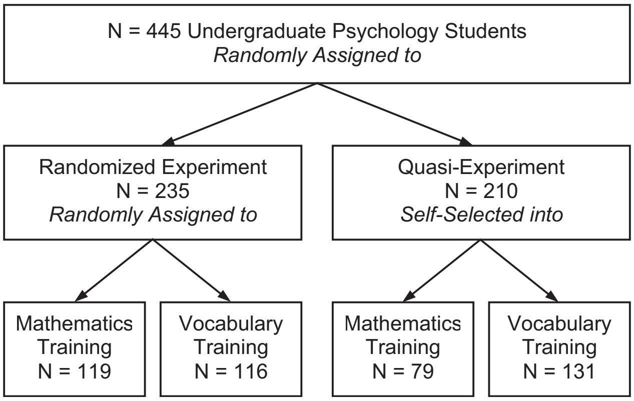

Figure 1. Overall design of the within-study comparison of the randomized experiment and the quasi-experiment.

After training, all participants were assessed on both mathematics and vocabulary outcomes. A 50-item posttest contained 30 vocabulary items and 20 mathematics items, presenting vocabulary first and mathematics second for all participants in all conditions. This posttest was given to all participants regardless of training. A more detailed description of the study's design and its implementation is given in Shadish et al. (2008).

Decomposition of Covariates

For this study, we decompose the complete set of covariates **X** into smaller, more homogeneous sets to investigate how well they establish strong ignorability and reduce bias. We identified five basic domains of constructs that are ranked below according to their typical availability to researchers:

1. *Demographics* (abbreviated to *dem*; six constructs): The measures included the student's age, gender, race/ethnicity (Caucasian, African American, Hispanic), marital status, and credit hours completed at the university. The first four of these measures are conventional, and the number of credit hours is included here because it assesses the quantity of academic performance and because it had the highest correlation with age (.26) among all covariates.

2. *Proxy-pretests* (abbreviated to *pre*; two multi-item constructs): The 36-item Vocabulary Test II (Educational Testing Service, 1962) measured vocabulary skills, and the 15-item Arithmetic Aptitude Test (Educational Testing Service, 1993) measured mathematics skills. Because both pretests differ from the study posttests in content and scale, they are proxy-pretest measures.

3. *Prior academic achievement* (abbreviated to *aca*; three multi-item constructs): Students' high school GPAs, current college GPAs, and the ACT college admission scores were used, and nearly all participants (84%) gave consent to access university records for these measures. However, the records contained ACT scores for only 62% of participants, though missingness was not significantly related to the four conditions (type of experiment and treatment condition) to which the participants were later assigned ($\chi^2 = 1.6$, $p = .20$). We substituted self-reported SAT scores, ACT scores, and GPAs for participants who did not consent or who had missing data in the university records, and we converted SAT

scores to ACT estimated scores using tables provided by ACT and Educational Testing Services (Dorans, Lyu, Pommerich, & Houston, 1997). Although the missing ACT scores could have been imputed (Hill, Reiter, & Zanutto, 2004), we believe that using self-reported ACT scores is probably more accurate and transparent. Finally, 12% of ACT scores and 4% of college GPAs were still missing.

4. *Topic preference* (abbreviated to *top*; six multi-item constructs): We assessed liking literature, liking mathematics, preferring mathematics over literature, math anxiety, the number of prior mathematics courses, and whether the major field of study was math intensive. Liking literature and liking mathematics consisted of two items each, preferring mathematics over literature of one item only. The number of prior mathematics courses is a composite of whether courses in algebra and calculus were taken in high school and college. Mathematics anxiety was measured according to the Short Mathematics Anxiety Rating Scale (Faust, Ashcraft, & Fleck, 1996), which consists of 25 items and assesses stress induced by mathematics. The major field of study is a dummy variable indicating whether their major field at the university is math intensive. Overall, this set of covariates mainly reflects a student's motivation for self-selection into vocabulary or mathematics treatment.

5. *Psychological predisposition* (abbreviated to *psy*; six multi-item constructs): We assessed depression according to the Short Beck Depression Inventory (Beck & Beck, 1972) and the Big Five personality factors (Extroversion, Emotional Stability, Agreeableness, Openness to Experience, and Conscientiousness) on the basis of the International Personality Item Pool of 50 items (Goldberg, 1992).

In sum, there were 23 single- and multi-item constructs based on a total of 156 questionnaire items. With the exception of the missing administrative data on prior academic achievement, for which self-reported alternatives were used as discussed above, missing values were rare. The overall incidence was 2.6% and 3.6% of all covariate measurements in the randomized experiment and in the quasi-experiment, respectively. For these last missing measures, Shadish et al. (2008) did not use multiple imputation, relying instead on maximum likelihood estimates using an expectation-maximization algorithm (Schafer & Graham, 2002). Information was also gathered on parental income and on mother's and father's educational attainment, but the restricted reliability of student reports on parental characteristics and the relatively large amount of missing data led us to exclude these covariates from further analysis.

Analytic Strategy

In tackling potential selection bias, we presume that selection into the vocabulary or mathematics group is a function of all five construct domains together: $P(Z = 1|\mathbf{X}) = f(\mathbf{X}) = f(\mathbf{X}^{dem}, \mathbf{X}^{pre}, \mathbf{X}^{aca}, \mathbf{X}^{top}, \mathbf{X}^{psy})$, where the treatment variable Z indicates vocabulary ($Z = 1$) or mathematics training ($Z = 0$). In a typical observational study, we would have to assume a treatment assignment that is strongly ignorable to get an unbiased treatment effect. More formally, treatment assignment is said to be strongly ignorable if the potential outcomes (Y^0, Y^1) are conditionally independent of treatment Z, given a set of observed covariates $\mathbf{X} = (X_1, \ldots, X_p)'$, that is, $(Y^0, Y^1) \perp Z|\mathbf{X}$ with $0 < P(Z = 1|\mathbf{X}) < 1$ (Rosenbaum & Rubin, 1983). Within the framework of the Rubin causal model, potential outcomes Y^0 and Y^1 refer to the outcomes that a subject would reveal if it would be in the control condition ($Z = 0$) or treatment condition ($Z = 1$), respectively. However, only one is actually observed in practice, that is, $Y = Y^0(1 - Z) + Y^1Z$ (Holland, 1986; Rubin, 1974). Note that the ignorability assumption refers to the pair of potential outcomes (Y^0, Y^1) but not to the observed outcome Y, which in general is dependent on treatment Z given $\mathbf{X}$. Ideally, all covariates $\mathbf{X}$ are measured before treatment assignment. Collecting them after assignment or treatment may lead to covariate measurements influenced by treatment. Including such covariates in a PS analysis could result in an under- or overestimation of the treatment effect.

Removing selection bias calls for a set of covariates that establishes the conditional independence $(Y^0, Y^1) \perp Z|\mathbf{X}$ (alternative conditions implying unbiasedness are given in Steyer et al., 2000). However, the strong ignorability assumption is silent about the structure and type of covariates. A minimal set of covariates would consist only of nonredundant covariates that are both correlated with treatment and with potential outcomes. With nonredundant we imply that a covariate needs to be conditionally correlated with both treatment and potential outcomes, given all other covariates included in the set. Such a set of covariates is not necessarily unique; several sets could establish the required conditional independence. Moreover, a nonminimal set of covariates might also include redundant covariates that are conditionally uncorrelated either with treatment assignment or with potential outcomes, or even with none of it. One such set, which is typical for the PS-approach, includes all covariates needed for correctly modeling the selection mechanism. Hence, it also includes covariates that are uncorrelated with potential outcomes. Given incomplete knowledge about the selection process and the "true" outcome model, it seems wise to collect a targeted and rich set of covariates, hoping that they will establish the strong ignorability assumption required for unbiased causal inference (Rubin, 2007; Rubin & Thomas, 1996). That is exactly the strategy that Shadish et al. (2008) applied. Although it is rarely possible to test the crucial strong ignorability assumption in actual research practice (Rosenbaum, 1984), the within-study design permits comparing the adjusted results of the quasi-experiment with the results from the randomized experiment to directly assess how well the covariates succeed in reducing selection bias.

We use a series of different selection models that systematically vary construct domains as well as single constructs. The first column of Table 1 lists all combinations of domains with their corresponding number of constructs. First, we separately considered the importance of each of the five construct domains alone. Second, we investigated all two-way combinations including the demographic and proxy-pretest measures because they are the most commonly available in actual research practice. Third, we reported on a selection of three- and four-way combinations, including those that do not include the two domains that are the most important for bias reduction, thus allowing us to test what

Table 1
Balance in Propensity Score (PS) and Observed Constructs Before and After PS Stratification

		Before PS-adjustment				After PS-stratification				Randomized experiment	
		PS-logit		Covariates		PS-logit		Covariates		Covariates	
Construct sets	No. of constructs	*d*	*v*	*d*%	*v*%	*d*	*v*	*d*%	*v*%	*d*%	*v*%
dem	6	0.45	0.64	50	100	0.01	0.86	100	100	67	67
pre	2	0.52	0.89	0	50	0.04	0.94	100	100	50	100
aca	3	0.45	0.90	100	67	0.05	0.99	67	100	67	100
top	6	0.95	0.61	17	17	0.08	0.82	83	83	83	100
psy	6	0.43	0.83	33	83	0.08	0.86	100	83	33	67
dem + pre	8	0.60	0.84	38	88	−0.02	0.89	100	100	63	75
dem + aca	9	0.61	0.81	67	89	0.04	0.93	100	100	67	78
dem + top	12	0.99	0.65	33	58	0.08	0.83	100	83	75	83
dem + psy	12	0.70	0.82	42	92	0.09	1.05	92	100	50	67
pre + top	8	1.00	0.63	13	25	0.05	0.89	100	88	75	100
pre + aca	5	0.64	0.91	60	60	−0.04	1.08	100	100	60	100
pre + psy	8	0.79	1.01	25	75	0.06	1.05	100	100	38	75
dem + pre + top	14	0.99	0.59	29	57	0.04	0.85	100	93	71	86
dem + aca + psy	15	0.72	1.06	53	87	0.02	1.04	100	93	53	73
dem + pre + aca	11	0.74	0.90	55	82	0.08	0.98	91	100	64	82
dem + pre + aca + psy	17	0.84	1.07	47	82	0.09	0.94	94	88	53	76
dem + aca + top + psy	21	1.09	0.67	43	67	0.02	1.03	100	90	62	81
dem + pre + aca + top	17	1.06	0.68	41	59	0.09	0.94	100	88	71	88
dem + pre + aca + top + psy	23	1.05	0.66	39	65	0.01	0.97	100	87	61	83

Note. Constructs sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). $d = (\bar{x}_t - \bar{x}_c)/\sqrt{(s_t^2 + s_c^2)/2}$, and $v = s_t^2/s_c^2$, where $\bar{x}_t$ and $\bar{x}_c$ are the sample means of a single construct x or the PS-logit in the treatment and comparison group; s_t^2 and s_c^2 are the corresponding sample variances. $d\% = \%\{x_j: -0.1 < d < 0.1\}$ is the portion of single constructs with an absolute bias d smaller than 0.1 standard deviations. $v\% = \%\{x_j: 4/5 < v < 5/4\}$ is the portion of single constructs with a variance ratio v larger than 4/5 and smaller than 5/4.

happens when we have many covariates but not the best single ones. Fourth, we considered all domains together to show that we are able to replicate the results of Shadish et al. (2008), though a slightly different procedure was used. Fifth, we investigated the importance of single constructs within the important topic preference and proxy-pretest domains to identify whether there are any single constructs that capture the outcome-related part of the selection process so well that they alone can be responsible for eliminating bias. Finally, we investigated whether a broad set of constructs can compensate for the omission of most important single constructs. We did so by excluding the two most important constructs from the remainder, thus creating an impoverished remainder that might nonetheless be collectively adequate for eliminating all or most of the initial selection bias.

We also varied the mode of data analysis, first by using ANCOVA with all covariates of the set under investigation and then by using the corresponding PSs in stratification, covariance, and weighted analyses as described in detail below. These particular analyses do not exhaust all possible models, but they are among the most frequently used today. The aim is to see which mode of data analysis most closely approximates the results of the experiment and to see whether analytic methods are as important in bias reduction as covariate choice. We used the statistical software package R to conduct all PS analyses with self-written functions (R Development Core Team, 2007).

Estimation of PSs

The PS is defined as the probability of treatment exposure given the observed covariates, $e(\mathbf{X}) = P(Z = 1|\mathbf{X})$, where we assume that $0 < e(\mathbf{X}) < 1$. Rosenbaum and Rubin (1983) proved that if treatment selection is strongly ignorable given observed covariates $\mathbf{X}$, it is also strongly ignorable given the PS, that is, $(Y^0, Y^1) \perp Z|e(\mathbf{X})$. For the quasi-experimental data, we estimated the PS for the vocabulary training group with logistic regression, and we denote it by $\hat{e}(\mathbf{X})$. Consequently, the PS for being in the mathematics group is given by $1 - \hat{e}(\mathbf{X})$. PS-models are not specified to optimize prediction or a model fit criterion but to balance pretreatment group differences in observed covariates. The balance property states that treated and control subjects with the same PS have the same distribution of observed covariates $\mathbf{X}$ (Rosenbaum & Rubin, 1983), that is, a PS-adjustment should equate the distribution of $\mathbf{X}$ of the treatment and comparison groups. Because the assessment of the similarity of the multivariate distributions of $\mathbf{X}$ is nearly impossible with a large set of covariates, most balancing criteria focus on univariate distributions of covariates and PS-logits only. We assess balance in observables and PS-logits using Cohen's $d = (\bar{x}_t - \bar{x}_c)/\sqrt{(s_t^2 + s_c^2)/2}$ and variance ratio $v = s_t^2/s_c{}^2$ between treatment and comparison groups. After PS-adjustment, standardized mean differences d should be close to zero, and variance ratios v should be close to one (Rubin, 2001). For each of our PS-models, we started by including all covariates of the covariate set under investigation as main effects. For some PS-models it was necessary to include interaction and higher order terms to achieve balance. For others, better balance was obtained by dropping predictors. Although the specification of a PS-model is usually a stepwise procedure of entering and dropping predictors, it does not inflate Type I and II errors in testing the treatment effect because the PS-model is specified independent of the outcome. Only after estimating the balancing PS is the outcome analyzed.

Table 1 depicts these balancing statistics for PS-logits and covariates before and after PS-stratification, in which balance is assessed only on the covariate set under investigation. Before any adjustment, topic preference shows the worst balance between treatment and comparison groups, followed by proxy-pretests. For these two covariate domains, standardized mean differences d in the PS-logit are 0.95 and 0.52 standard deviations, respectively, and the corresponding variance ratios v are 0.61 and 0.89. With regard to the underlying covariates, only one of six topic preference covariates (= 17%) and none of two proxy-pretest covariates exhibit a standardized mean close to zero (i.e., $-0.1 < d < 0.1$); furthermore, only 17% and 50% of covariates of the two sets show a nearly balanced variance ratio ($4/5 < v < 5/4$).

After adjustment via PS-stratification, all PS-models estimated in this study achieve balance in PS-logits and on the covariate set under investigation. Only minor differences in means and variances remain. With the exception of two variance ratios, all models show balance in observables at least as good as in the randomized experiment (the last two columns of Table 1 contain balancing statistics for the covariate-unadjusted randomized experiment). We also conducted F tests to assess overall and within-PS-quintile mean differences in PS-logits and covariates. Only for a few PS-models was it not possible to achieve equal means within quintiles—particularly the first and fifth quintile. This is mainly due to the fact that we always included all observations, even observations slightly outlying the common support region. Though the slight imbalance might result in less bias reduction than with a perfect balancing score, the additional covariate adjustment in the outcome model—as discussed below—very likely compensates for it. Thus, with a few exceptions, PS-stratification yields good balance by all the usual criteria. For PS-models investigating the importance of single covariates, we achieved very similar results on balance (not presented here).[1] However, as we show in the section on bias reduction, excellent balance in observed covariates does not inevitably lead to bias reduction. To the latter requires meeting the often elusive strong ignorability assumption—a set of covariates that establishes conditional independence of treatment assignment and potential outcomes.

Estimation of Average Treatment Effects

To estimate the average treatment effect and subsequently the degree of bias reduction over different covariate sets and adjustment methods, we conducted four different analyses (for an overview, see Lunceford & Davidian, 2004; Morgan & Winship, 2007; Rubin, 2006; Schafer & Kang, 2008): (a) stratification based on PS-quintiles, (b) ANCOVA including PS-logits, (c) PS-weighting, and (d) simple ANCOVA without PSs. We did not use PS-matching because our comparison groups were too small for matching treated and untreated.

Each of the three PS-methods tries to control for group differences in the PS in a different way. PS-stratification uses PS to subclassify observations into $q = 1, \ldots, Q$ strata—typically quintiles—with index sets $I_q = \{i\text{: observation } i \in \text{stratum } q\}$ indicating each observation's stratum membership. The classical PS-stratification approach estimates treatment effects for each of the Q strata by $\hat{\tau}_q = \{\sum_{i \in I_q} z_i\}^{-1} \sum_{i \in I_q} z_i y_i - \{\sum_{i \in I_q} (1 - z_i)\}^{-1} \sum_{i \in I_q} (1 - z_i) y_i$ and then pools them across strata, that is, $\hat{\tau} = \sum_{q=1}^{Q} w_q \hat{\tau}_q$ with weights $w_q = n_q/n$. PS-ANCOVA simply uses the logit of the PS and higher order terms thereof as covariates in an ANCOVA model. In the PS-weighting approach, inverse probability weights $w_i = z_i/\hat{e}_i + (1 - z_i)/(1 - \hat{e}_i)$ are directly obtained from the estimated PS and used to estimate the average treatment effect $\hat{\tau} = \{\sum_{i=1}^{n} z_i w_i\}^{-1} \sum_{i=1}^{n} z_i w_i y_i - \{\sum_{i=1}^{n} (1 - z_i) w_i\}^{-1} \sum_{i=1}^{n} (1 - z_i) w_i y_i$. All three approaches have their advantages and disadvantages (Schafer & Kang, 2008): PS-ANCOVA relies on the correct specification of the functional form, PS-weighting is rather sensitive to extreme weights, and PS-stratification cannot remove all bias if some imbalance remains within strata—under some assumptions we can at least expect that five PS-strata remove approximately 90% of selection bias (Cochran, 1968; Rosenbaum & Rubin, 1984).

In following Hirano and Imbens (2001), we combined PS-adjustments described above with covariate adjustments in a regression framework (see also Bang & Robins, 2005; Robins & Rotnitzky, 1995; Rubin & Thomas, 1996). Assuming a linear relation between predictors and outcome, the basic regression model for estimating the covariance adjusted treatment effect is given by $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$, where β_0 is a constant, Z_i is the treatment indicator variable, τ is the average treatment effect, $\mathbf{X}_i$ is a column vector of k predictors (including polynomials or interaction terms) for $i = 1, \ldots, n$ subjects, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is the corresponding vector of regression coefficients. This model assumes that treatment and comparison groups differ only with respect to the treatment effect τ, which is constant across all values of predictors. However, if this assumption does not hold, interaction terms between treatment and the centered predictors need to be included to get unbiased average treatment effects with simple ANCOVA (Schafer & Kang, 2008). PS-weighting is integrated into the basic regression approach by using weighted least squares with PS-weights $w_i = z_i/\hat{e}_i + (1 - z_i)/(1 - \hat{e}_i)$. For PS-stratification, individual weights are derived from PS-strata (see also Hong & Hong, 2009): For observation $i \in I_q$ with treatment z_i, the weight is given by $w_{zq} = (n_z \cdot n_q/n)/n_{zq}$, where n_{zq} is the number of subjects in treatment group z and stratum q, $n_z = \sum_{q=1}^{Q} n_{zq}$, $n_q = \sum_{z=0}^{1} n_{zq}$, and n is the total number subjects.[2] Given a strongly ignorable treatment assignment, these weighted least squares estimators of the average treatment effect are consistent, and using PS-balancing together with covariance adjustment makes the estimation of the treatment effect "doubly robust" because it protects against misspecification of either the PS-model or the outcome model (Robins & Rotnitzky, 1995, 2001). However, if both models are misspeci-

[1] We estimated a PS even if the model consisted of a single covariate only.

[2] These weights for the weighting and stratification approach refer to the average treatment effect for the population of treated and untreated subjects together. If the interest is on the conditional average treatment effect for the treated only, weights have to be adapted with regard to the treated population.

fied, estimates remain biased, and there is no guarantee that the combined approach results in less bias than regression or PS-analyses on their own (Kang & Schafer, 2007). For all regression adjustments, we included all constructs under investigation (but as main effects only).

We also covariance-adjusted the treatment effects of the randomized experiment for all 23 constructs to control for random group differences. As we would expect with an experiment, covariance adjustment only marginally changed treatment effects by 2% and −4% for vocabulary and mathematics, respectively. In following Rubin (2008a), we increased the comparability of the randomized experiment and the quasi-experiment by restricting the analysis of the randomized experiment to those cases whose PS-logit fell into the range of the quasi-experimental PS-logits. To do this, we first used the PS-model of the quasi-experiment to estimate the PS-logits for the randomized experiment and then discarded all cases scoring 0.1 standard deviations below the minimum or above the maximum of the observed quasi-experimental PS-logits. This was done separately for each set of constructs under investigation—between 0 and 24 (11%) nonoverlapping cases were discarded. However, results were rather insensitive to the deletion of these cases; they were substantially the same as for the full sample.

Comparison of Effect Sizes

The effect sizes from experiments and observational studies have to be directly compared. This is not as easy as it sounds because both the experimental and quasi-experimental effect estimates are liable to sampling error. Each of the two experiments had enough power to detect small effect sizes, but we do not have sufficient power to reveal statistically significant differences between the randomized and quasi-experiment, given the modest initial bias and moderate sample sizes. To draw conclusions about differences in estimates, we rely on replication across the two independent treatments—one in vocabulary and the other in mathematics. Though the selection process is identical in both cases, the covariates' relationship with the outcomes can still differ, and, as we shall see, they do so. Hence, we infer the relative importance of construct domains (or single constructs) for removing bias from the consistency in replication across the math and vocabulary treatments and outcomes. Therefore, our presentation of results is basically descriptive and treats the covariate-adjusted experimental results as the true treatment effect that the adjusted quasi-experiment is trying to recreate.

Results

The Impact of Construct Domains on Bias Reduction

The ability of each method and set of constructs to reduce selection bias is assessed by the fraction of the initial selection bias remaining after adjustment, that is, $\hat{b}\% = (\hat{\tau}_Q^{adj} - \hat{\tau}_E)/(\hat{\tau}_Q^{unadj} - \hat{\tau}_E) \cdot 100$, where $\hat{\tau}_Q$ is the adjusted or unadjusted average treatment effect in the quasi-experiment, and $\hat{\tau}_E$ is the estimated average treatment effect in the randomized experiment. Given a positive initial bias (i.e., $\hat{\tau}_E < \hat{\tau}_Q^{unadj}$), a positive sign of $\hat{b}\%$ indicates an underadjustment with respect to the experimental effect ($\hat{\tau}_E < \hat{\tau}_Q^{adj}$), and a negative sign indicates overadjustment ($\hat{\tau}_Q^{adj} < \hat{\tau}_E$), meaning that actual bias adjustment removes even more than the initial bias and introduces bias of opposite direction. In our case, overadjustments are very likely due to the sampling error in both the experimental and quasi-experimental treatment effects.

Remaining bias in vocabulary. Table 2 and Figure 2 depict the results for vocabulary training. The estimated treatment effect for the randomized experiment is 8.18 points, which corresponds to an effect size of 2.4 standard deviations (Cohen's *d*). Because the unadjusted effect for the quasi-experiment amounts to 9.00 points, the estimated selection bias is 9.00 – 8.18 = 0.82 points, or 0.24 standard deviations.

When construct domains are examined singly, topic preference leads to the highest reduction in bias, with the remaining bias ranging from 31% for PS-ANCOVA to 50% for PS-stratification, being an average of 39% across all four methods. Proxy-pretests do the next best in reducing bias, with an average remaining bias of 42% (with a range from 36% for PS-weighting to 49% for PS-stratification). After adjustment for just the demographic covariates, 53%–66% of the initial bias is still present. Controlling for academic achievement or psychological measures also fails to substantially reduce bias—with 66% and 72% of the initial bias remaining on average.

Combining two construct domains improves bias reduction. Indeed, having both the proxy-pretest and topic preference measures in the model removes almost all bias. Thus, PS-weighting leaves only 7% of the initial bias, whereas PS-ANCOVA and traditional ANCOVA overadjusted by 9%–18%. Averaged across methods, only 4% overadjusted bias remains. All other combinations of two covariate sets reduce considerably less of the initial bias, though typically more bias than each single covariate set when used on its own (see Figure 2). The best pairing of three covariate sets also includes proxy-pretests and topic preference—together with demographics they reduce nearly all bias (1% overadjustment). In contrast, considerable bias remains if neither proxy-pretests nor measures on topic preference are taken into account—remaining bias amounts to 31%–53% when covariates on demographics, prior academic achievement, and psychological predisposition are combined. Using four domains reduces almost all selection bias because either topic preference or proxy-pretests are necessarily involved in the model. Finally, the inclusion of all domains leads to a modest overadjustment of the treatment effect in comparison with the experimental effect size by 3% on average; depending on the adjustment method, the remaining bias varies between −9% and 5%. Thus, bias is completely removed if just the two most important domains—proxy-pretests and topic preference—are taken into account; bias reduction increases as more and more covariates are available; however, proxy-pretests or motivational factors have to be among these covariate sets because the three other sets were not adequate by themselves for meaningful bias reduction.

The choice of covariates has a much stronger impact on bias reduction than the choice of a specific adjustment method. No adjustment method is uniformly or on average significantly better than the others, and simple ANCOVA does as well as any PS-analysis. As a sole exception, ANCOVA may on average perform slightly worse when prior academic achievement measures are

Table 2
Adjustment for Constructs Domains: Effect Estimates and Bias for Vocabulary Training

	PS-stratification[a]			PS-ANCOVA[b]			PS-weighting[c]			ANCOVA[d]			
Vocabulary constructs sets	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\overline{\hat{b}\%}$
Adjusted randomized experiment[e]										8.18	0.39		
Unadjusted quasi-experiment										9.00	0.51		
Adjusted quasi-experiments													
dem	8.57	0.51	53	8.69	0.48	66	8.68	0.48	65	8.64	0.48	60	61
pre	8.58	0.45	49	8.56	0.43	47	8.47	0.43	36	8.48	0.44	37	42
aca	8.68	0.48	60	8.70	0.46	63	8.69	0.47	61	8.83	0.44	78	66
top	8.54	0.57	50	8.36	0.55	31	8.42	0.56	37	8.41	0.54	36	39
psy	8.78	0.49	74	8.75	0.47	70	8.70	0.47	64	8.83	0.48	80	72
dem + pre	8.55	0.46	48	8.48	0.45	39	8.41	0.44	31	8.40	0.43	30	37
dem + aca	8.54	0.49	42	8.49	0.45	36	8.52	0.47	40	8.58	0.43	47	41
dem + top	8.52	0.54	41	8.36	0.51	22	8.41	0.53	27	8.36	0.51	22	28
dem + psy	8.51	0.50	49	8.47	0.45	46	8.56	0.45	54	8.57	0.47	55	51
pre + top	8.30	0.49	4	8.21	0.48	−9	8.33	0.47	7	8.15	0.47	−18	−4
pre + aca	8.46	0.45	27	8.37	0.43	14	8.26	0.43	0	8.44	0.42	24	16
pre + psy	8.41	0.44	31	8.45	0.42	36	8.33	0.44	22	8.40	0.44	30	30
dem + pre + top	8.32	0.53	7	8.25	0.48	−2	8.32	0.48	7	8.17	0.46	−14	−1
dem + aca + psy	8.48	0.46	44	8.46	0.43	42	8.37	0.44	31	8.56	0.43	53	43
dem + pre + aca	8.25	0.47	13	8.28	0.44	17	8.21	0.45	9	8.35	0.42	26	16
dem + pre + aca + psy	8.18	0.45	2	8.19	0.43	3	8.09	0.44	−8	8.35	0.43	22	5
dem + aca + top + psy	8.32	0.50	25	8.13	0.50	5	8.13	0.50	5	8.25	0.46	17	13
dem + pre + aca + top	8.17	0.53	5	8.02	0.49	−13	8.18	0.54	6	8.14	0.45	1	0
dem + pre + aca + top + psy	8.19	0.49	5	8.07	0.46	−9	8.13	0.45	−2	8.12	0.45	−4	−3

Note. Construct sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). All treatment effect estimates $\hat{\tau}$ are based on (weighted) regression analyses with constructs included as main effects. Standard errors for the propensity score (PS) methods are based on 1,000 bootstrap samples (separate samples for each group with replacement) with refitted PSs, quintiles, and weights for each sample (predictors remained unchanged). Bias in quasi-experimental estimates is defined with respect to experimental estimates ($\hat{\tau}_Q - \hat{\tau}_E$). To account for differences in the randomized and quasi-experiment, we discarded cases outlying the range of the corresponding quasi-experimental PS-logit by 0.1 standard deviation. Hence, $\hat{\tau}_E$ slightly differs for each set of constructs. $\hat{b}\% = (\hat{\tau}_Q^{adj} - \hat{\tau}_E)/(\hat{\tau}_Q^{unadj} - \hat{\tau}_E) \cdot 100$ is the fraction of bias remaining after PS and/or covariance adjustment. A positive sign indicates an underadjustment with respect to the experimental effect, and a negative sign indicates an overadjustment. $\overline{\hat{b}\%}$ represents the average across all four adjustment methods. ANCOVA = analysis of covariance.

[a] Weighted least squares (WLS) regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{L}_i'\boldsymbol{\lambda} + \mathbf{D}_i'\boldsymbol{\delta} + \varepsilon_i$ with stratum weights; $\mathbf{L}_i$ are predictors based on PS-logits (linear, quadratic, and cubic term); $\mathbf{D}_i$ are four dummy variables representing PS-quintiles. [b] Ordinary least squares (OLS) regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{L}_i'\boldsymbol{\lambda} + \varepsilon_i$. [c] WLS regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ with PS-weights. [d] OLS regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ (also for the randomized experiment but not for the unadjusted quasi-experiment). [e] The estimated experimental treatment effect refers to the quasi-experiment including all 23 constructs, that is, after discarding cases outlying the range of the quasi-experimental PS-logit.

included—misspecified functional forms are very likely the reason for it (Schafer & Kang, 2008).

Remaining bias in mathematics. The estimated mathematics treatment effect for the randomized experiment is 4.06 points, making an effect size of 1.3 standard deviations. Initial bias in mathematics amounts to 5.01 – 4.06 = 0.95 points, corresponding to 0.26 standard deviations.

Mathematics training clearly demonstrates the necessity of adequate selection modeling (see Table 3 and Figure 3). Constructs assessing psychological predisposition, demographics, prior academic achievement, and proxy-pretests hardly reduce bias at all either by themselves or when combined. In contrast, topic preference actually results in an overadjustment ranging from 3% to 26% of initial bias, with an average of 15% across methods. Adding other covariates to topic preference leads to no noticeable change in bias reduction. If all five domains are used, bias is overadjusted by 10% on average—varying from −3% for PS-stratification to −17% for ANCOVA.

The sole combination of other domains that comes close to the reduction achieved by topic preference is the combined set of demographics, proxy-pretests, and prior academic achievement. However, even this, with an average of 25% bias remaining across methods, is more than when topic preference is used by itself (−15%). Adding constructs of psychological predisposition does not further reduce bias, as 23% of initial bias remains. Once again, covariates are much more important than the choice of an adjustment method. No adjustment method is uniformly and on average best or worst, and simple ANCOVA shows similar results to PS-methods (except when prior academic achievement is involved, as was the case for the vocabulary outcome).

The analyses presented above indicate that the actual covariates used make a major difference in reducing bias. There is more to bias reduction than merely adding more and more covariates. Thus, for mathematics, the topic preference domain played a unique role; it had to be included for an adequate degree of bias reduction. For vocabulary, there was no single dominant domain, but either topic preference or proxy-pretests had to be included to remove nearly all bias. Considerable bias remained if these two domains were omitted. Although the choice of covariates mattered

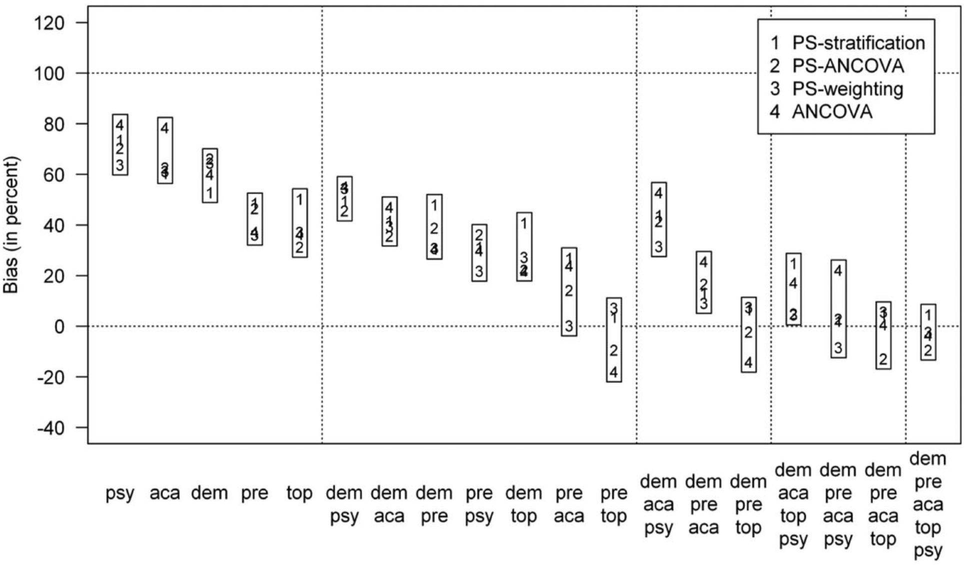

Figure 2. Remaining bias in vocabulary by construct set and analytic method (in the order of average bias). Construct sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). PS = propensity score; ANCOVA = analysis of covariance.

for bias reduction, the mode of analysis did not. One PS-method did as well as another, and no PS-method did uniformly better than ANCOVA.

The Impact of Single Constructs on Bias Reduction

We now analyze within the topic preference and proxy-pretest domains to discover whether there were unique constructs responsible for bias reduction whose inclusion was necessary for significant bias reduction.

Remaining bias in vocabulary. Table 4 and Figure 4 show the results for constructs belonging to proxy-pretests or topic preference. It is striking that two constructs are sufficient for reducing almost all the selection bias—vocabulary proxy-pretest and preferring mathematics over literature. An average of 20% bias remains across methods when controlling for the former, and an average of 22% bias remains for the latter. If both constructs are used together, remaining bias drops to −8% (not shown in Table 4). No other single construct came close to reducing all the bias. Some even seem to increase bias, possibly because of model misspecification resulting from the omission of important variables (Heckman & Navarro-Lozano, 2004; see also Morgan & Winship, 2007; Pearl, 2009): Bias increases from 20% to 42% when the mathematics proxy-pretest is added to the vocabulary one, thus thwarting the bias-reducing capacity of the vocabulary proxy-pretest on its own (see Tables 2 and 4). Including additional covariates does not necessarily reduce bias; it may also sometimes increase it. However, the bias increasing effect of the math proxy-pretest vanishes the more covariates are included in the model.

The relative importance of both vocabulary proxy-pretest and also preferring mathematics over literature is best evaluated when they are both left out of analyses that include all the other constructs from all five domains. In comparison with the case in which all 23 constructs are included, omitting just the vocabulary pretest increases average bias from −3% (all 23 constructs) to 21%, whereas 21% bias also remains if preferring mathematics over literature is left out. Omitting both constructs increases the post-adjustment bias to 31%. However, this is still better than omitting all the measures from the proxy-pretest and topic preference domains, whatever their individual efficacy, for then 43% of the initial bias still remains (see Table 2). Hence, the individually less efficacious items from the two crucial domains do compensate to some extent for omitting the two most important single constructs from the most important domains. Having a pretest measurement both of the outcome and of the most important part of the selection process seems well-nigh indispensable for removing all selection bias for the vocabulary outcome.

Remaining bias in mathematics. As for the mathematics outcome, Table 5 and Figure 5 reveal that two single construct items are sufficient for reducing almost all the bias, both within the topic preference domain. One is liking mathematics, and the other is, once again, preferring mathematics over literature. Each removes almost all bias, −3% remains on average for the former and 6% for the latter. No other single construct is able to remove a major part of the bias. Controlling for the mathematics proxy-pretest alone leaves on average 66% bias, for instance, in contrast with the vocabulary outcome in which the vocabulary proxy-pretest was more successful.

When all 23 constructs are included in the analyses, almost all the bias in mathematics is reduced, and only −10% remains. Excluding preferring mathematics over literature from the overall set of covariates leads to an overadjustment by 24%, whereas the exclusion of liking mathematics results in 1% remaining bias. Omitting both covariates increases bias to 10% on average. This is still much better than when all six covariates of topic preference are omitted—bias then averages to 23% (see Table 3). This

Table 3

Adjustment for Constructs Domains: Effect Estimates and Bias for Mathematics Training

	PS-stratification[a]			PS-ANCOVA[b]			PS-weighting[c]			ANCOVA[d]			
Mathematics constructs sets	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\overline{\hat{b}\%}$
Adjusted randomized experiment[e]										4.06	0.36		
Unadjusted quasi-experiment										5.01	0.55		
Adjusted quasi-experiments													
dem	5.18	0.52	118	5.02	0.50	101	4.94	0.48	94	4.94	0.53	93	102
pre	4.83	0.46	82	4.75	0.45	74	4.69	0.45	68	4.73	0.50	72	74
aca	4.99	0.46	98	4.72	0.43	68	4.88	0.43	87	5.00	0.45	99	88
top	3.91	0.53	−6	3.73	0.52	−23	3.94	0.50	−3	3.70	0.56	−26	−15
psy	5.13	0.55	113	5.15	0.54	115	5.18	0.54	119	5.06	0.55	106	113
dem + pre	4.40	0.45	35	4.61	0.45	58	4.65	0.43	62	4.67	0.48	64	55
dem + aca	4.62	0.48	58	4.61	0.45	57	4.66	0.43	62	4.72	0.45	69	62
dem + top	3.91	0.51	−6	3.72	0.50	−24	3.96	0.48	−1	3.77	0.53	−20	−13
dem + psy	4.91	0.56	89	5.01	0.52	100	4.94	0.52	92	4.89	0.54	87	92
pre + top	3.91	0.47	−9	3.77	0.44	−23	3.92	0.45	−8	3.84	0.51	−16	−14
pre + aca	4.44	0.46	42	4.36	0.41	35	4.51	0.42	50	4.51	0.45	50	44
pre + psy	4.78	0.47	74	4.82	0.45	79	4.69	0.45	65	4.75	0.51	71	72
dem + pre + top	3.92	0.42	−10	3.82	0.44	−19	4.01	0.42	0	3.89	0.49	−12	−10
dem + aca + psy	4.57	0.46	57	4.52	0.46	51	4.57	0.44	56	4.66	0.46	65	57
dem + pre + aca	4.28	0.44	19	4.27	0.43	18	4.35	0.40	27	4.43	0.45	35	25
dem + pre + aca + psy	4.28	0.44	19	4.24	0.43	15	4.31	0.40	23	4.40	0.45	33	23
dem + aca + top + psy	3.74	0.46	−32	3.71	0.45	−35	3.89	0.42	−15	3.84	0.46	−21	−26
dem + pre + aca + top	3.67	0.42	−31	3.64	0.43	−33	3.74	0.39	−24	3.79	0.45	−19	−27
dem + pre + aca + top + psy	3.98	0.39	−3	3.86	0.41	−15	3.94	0.38	−6	3.83	0.46	−17	−10

Note. Construct sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). All treatment effect estimates $\hat{\tau}$ are based on (weighted) regression analyses with constructs included as main effects. Standard errors for the propensity score (PS) methods are based on 1,000 bootstrap samples (separate samples for each group with replacement) with refitted PSs, quintiles, and weights for each sample (predictors remained unchanged). Bias in quasi-experimental estimates is defined with respect to experimental estimates ($\hat{\tau}_Q - \hat{\tau}_E$). To account for differences in the randomized and quasi-experiment, we discarded cases outlying the range of the corresponding quasi-experimental PS-logit by 0.1 standard deviation. Hence, $\hat{\tau}_E$ slightly differs for each set of constructs. $\hat{b}\% = (\hat{\tau}_Q^{adj} - \hat{\tau}_E)/(\hat{\tau}_Q^{unadj} - \hat{\tau}_E) \cdot 100$ is the fraction of bias remaining after PS and/or covariance adjustment. A positive sign indicates an underadjustment with respect to the experimental effect, and a negative sign indicates an overadjustment. $\overline{\hat{b}\%}$ represents the average across all four adjustment methods. ANCOVA = analysis of covariance.

[a] Weighted least squares (WLS) regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{L}_i'\boldsymbol{\lambda} + \mathbf{D}_i'\boldsymbol{\delta} + \varepsilon_i$ with stratum weights; $\mathbf{L}_i$ are predictors based on PS-logits (linear, quadratic, and cubic term); $\mathbf{D}_i$ are four dummy variables representing PS-quintiles. [b] Ordinary least squares (OLS) regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{L}_i'\boldsymbol{\lambda} + \varepsilon_i$. [c] WLS regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ with PS-weights. [d] OLS regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ (also for the randomized experiment but not for the unadjusted quasi-experiment). [e] The estimated experimental treatment effect refers to the quasi-experiment including all 23 constructs, that is, after discarding cases outlying the range of the quasi-experimental PS-logit.

indicates that having multiple measures within important domains like topic preference can at least partially protect against the omission of important single constructs and still remove almost all selection bias.

Why Did Different Covariates Reduce Bias for Vocabulary and Mathematics?

Results on bias reduction revealed that constructs on topic preference are sufficient and necessary to remove selection bias in the mathematics outcome. For the vocabulary outcome, the combination of topic preference and proxy-pretests reduced most successfully initial bias. The question is why did different constructs remove bias in vocabulary and mathematics? Table 6 shows that the topic preference domain is most strongly related to treatment selection. The multiple correlation of all six topic preference constructs with treatment is .43. The two most highly correlated single constructs with treatment are preferring mathematics over literature (−.38) and liking mathematics (−.36). The next important construct is major field of study; however, at −.19, it is only moderately correlated with treatment. Proxy-pretest measurements are not strongly correlated with treatment selection at all—for the vocabulary pretest, the correlation is .17, and for the mathematics pretest, it is −.09.

The high correlation of mathematics-related constructs suggests that in this study, self-selection was mainly driven by the math alternative. Students self-selecting into the mathematics training condition seem to have been influenced by a stronger desire to take or to avoid mathematics than to take or to avoid vocabulary training. Indeed, more students in the quasi-experiment opted for vocabulary than mathematics training, implying an active avoidance of the latter. Further data support this interpretation. Shadish et al. (2008) asked students why they chose their treatment conditions, and they coded their open-ended answers into four categories (liking for it, avoidance of it, self-efficacy, and self-improvement). Of all students in the quasi-experiment, 30% who chose mathematics said that they did so because they liked it, whereas only 18% of vocabulary participants said they did so

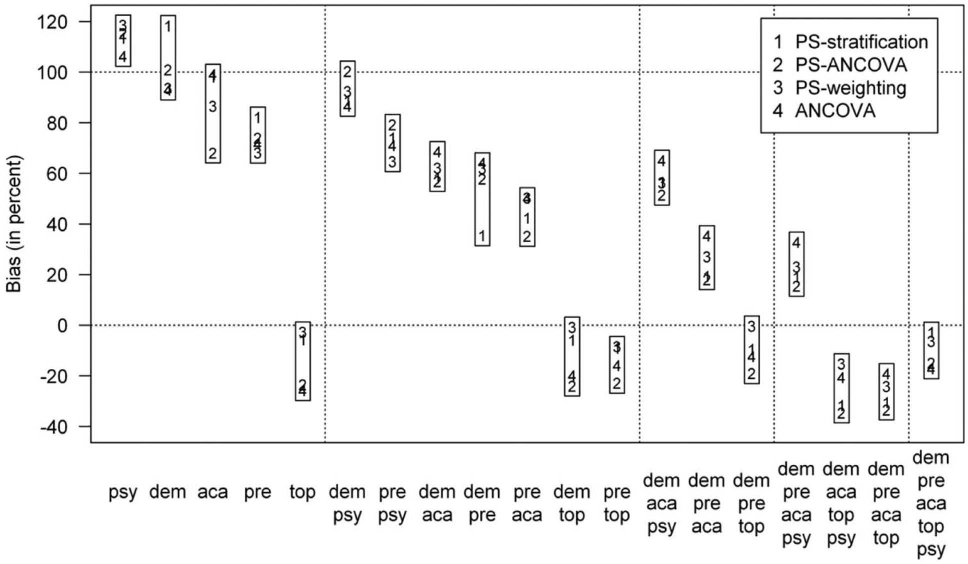

Figure 3. Remaining bias in mathematics by construct set and analytic method (in the order of average bias). Construct sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). PS = propensity score; ANCOVA = analysis of covariance.

because of liking vocabulary. Only 8% of the students in mathematics chose it to avoid vocabulary, but 21% of participants in the vocabulary training group reported taking vocabulary training to avoid mathematics. The other reasons for choosing a condition are more balanced across the mathematics and vocabulary groups—47% and 42% reported choosing their training for self-improvement, and 11% and 17% for self-efficacy, that is, they were good at the chosen topic or found it easy (see Table 7).

The salience of positive and negative feelings about mathematics in the selection process is also reflected in the multiple point-biserial correlations of basic covariate sets with the four reasons for choosing conditions. As Table 7 shows, all correlation coefficients except one are higher for the mathematics group than for the vocabulary group. Moreover, topic preference exhibits by far the highest correlations in the mathematics group (r = .60 with self-improvement, and r = .54 with liking). Hence, self-selection was more strongly dominated by the mathematics training condition than the vocabulary one.

However, this selection difference cannot by itself explain the obtained differences between vocabulary and mathematics results, for the selection model alone is the same for the vocabulary and mathematics outcomes. It is the differential correlation of the constructs to these two outcomes that makes the difference. The math-focused constructs of topic preference are more strongly correlated with the mathematics than the vocabulary outcome. Table 6 shows that liking mathematics is related to the math outcome (r = .36) but not to the vocabulary outcome (r = −.04). We also found considerable differences in correlations between outcome scores and the student's major field of study (r = .24 for mathematics outcomes, and r = −.05 for vocabulary) or the number of math courses (r = .26 for mathematics outcomes, and r = .11 for vocabulary). The strong correlations between these selection-relevant constructs and the mathematics outcome explain the unique role of topic preference in reducing bias for the mathematics condition of the quasi-experiment. The proxy-pretest measures were less influential in selecting into the mathematics training condition because the correlation between the mathematics proxy-pretest and treatment was weak (r = −.09), and the correlations between the vocabulary proxy-pretest and both treatment selection (r = .17) and mathematics outcome (r = .24) were only moderate compared with what we found with the topic preference covariates.

If we turn now to the vocabulary outcome, both topic preference and the vocabulary proxy-pretest played important roles in reducing bias. Though the correlation of vocabulary proxy-pretest with treatment is only moderate (r = .17), the pretest is highly related to the outcome (r = .53). Thus, the crucial covariates for removing bias are those that are both highly correlated with treatment and with potential outcomes. Furthermore, because the covariates strongly correlated with treatment show different relations to the vocabulary and mathematics outcomes across the two treatment groups, different constructs are necessary and sufficient for reducing bias in the vocabulary and the mathematics treatment conditions.

Discussion

Our reanalysis of Shadish et al.'s (2008) data identified the specific contributions made by each domain, by specific sets of domains, and by single constructs within some domains. At the domain level, selection bias in the math outcome was completely removed by topic preference alone. Within this domain, two of six constructs sufficed to remove all the bias: preferring mathematics over literature and liking mathematics. Preferring mathematics is based on a single question asked prior to the experiment, and it offered participants the very same choices they were to confront

Table 4
Adjustment for Single Constructs: Effect Estimates and Bias for Vocabulary Training

	PS-stratification[a]			PS-ANCOVA[b]			PS-weighting[c]			ANCOVA[d]			
Vocabulary constructs	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\overline{\hat{b}\%}$
Proxy-pretests													
Vocabulary pretest	8.44	0.42	26	8.46	0.42	29	8.35	0.43	15	8.32	0.44	11	20
Mathematics pretest	9.26	0.46	131	9.24	0.46	128	9.21	0.46	125	9.23	0.48	128	128
Topic preference													
Liking mathematics	8.85	0.46	82	9.00	0.46	100	8.80	0.46	76	8.94	0.55	93	88
Liking literature	9.00	0.51	100	8.81	0.48	76	8.79	0.50	74	8.79	0.51	75	81
Prefer math over literature	8.33	0.54	19	8.37	0.53	24	8.33	0.53	19	8.37	0.54	24	22
Mathematics anxiety scale	9.20	0.51	124	9.05	0.51	105	9.00	0.50	100	9.00	0.51	100	107
No. of math courses	9.09	0.51	112	9.11	0.51	115	9.07	0.50	109	9.13	0.51	117	113
Major field of study	8.93	0.49	92	8.95	0.49	94	8.93	0.49	92	8.95	0.52	94	93
All covariates of all five domains except for													
Vocabulary pretest	8.37	0.49	32	8.19	0.47	13	8.26	0.46	20	8.23	0.46	17	21
Prefer math over literature	8.06	0.53	14	8.09	0.49	16	8.15	0.50	22	8.26	0.46	33	21
Vocabularly pretest and prefer math	8.38	0.51	35	8.27	0.49	24	8.28	0.49	25	8.43	0.46	41	31

Note. Construct sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). All treatment effect estimates $\hat{\tau}$ are based on (weighted) regression analyses with constructs included as main effects. Standard errors for the propensity score (PS) methods are based on 1,000 bootstrap samples (separate samples for each group with replacement) with refitted PSs, quintiles, and weights for each sample (predictors remained unchanged). Bias in quasi-experimental estimates is defined with respect to experimental estimates ($\hat{\tau}_Q - \hat{\tau}_E$). To account for differences in the randomized and quasi-experiment, we discarded cases outlying the range of the corresponding quasi-experimental PS-logit by 0.1 standard deviation. Hence, $\hat{\tau}_E$ slightly differs for each set of constructs. $\hat{b}\% = (\hat{\tau}_Q^{adj} - \hat{\tau}_E)/(\hat{\tau}_Q^{unadj} - \hat{\tau}_E) \cdot 100$ is the fraction of bias remaining after PS and/or covariance adjustment. A positive sign indicates an underadjustment with respect to the experimental effect, and a negative sign indicates an overadjustment. $\overline{\hat{b}\%}$ represents the average across all four adjustment methods. ANCOVA = analysis of covariance.

[a] Weighted least squares (WLS) regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{L}_i'\boldsymbol{\lambda} + \mathbf{D}_i'\boldsymbol{\delta} + \varepsilon_i$ with stratum weights; $\mathbf{L}_i$ are predictors based on PS-logits (linear, quadratic, and cubic term); $\mathbf{D}_i$ are four dummy variables representing PS-quintiles. [b] Ordinary least squares (OLS) regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{L}_i'\boldsymbol{\lambda} + \varepsilon_i$. [c] WLS regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ with PS-weights. [d] OLS regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ (also for the randomized experiment but not for the unadjusted quasi-experiment).

later when they had to choose between training in math or vocabulary. The other important construct for bias reduction—liking mathematics—reflects the fact that math factors were more important for the selection process than vocabulary factors. None of the four other domains or the other four constructs within topic preference were sufficient to eliminate all of the initial bias in the quasi-experiment.

The selection bias in vocabulary was somewhat more complex. At the domain level, the topic preference and proxy-pretest domains had to be combined for maximal bias reduction. Within domains, the largest roles were played by the vocabulary rather than the math proxy-pretest measure and by preferring mathematics over literature. Removing bias in the vocabulary training condition did not require any of the three other domains, the proxy-pretest about mathematics ability, or five other constructs within topic preference. Therefore, for both interventions it would have been possible to establish strong ignorability and reduce almost all the bias with a minimal set of just one or two constructs had it been possible to know in advance which ones these were.

However, when taken together, all the constructs also removed all the selection bias. The inclusion of additional covariates neither improved nor diminished bias reduction; it only resulted in more efficient estimates of the treatment effect because of the additional covariance adjustment of PS-estimates. Of special interest is that three of the five domains—demographics, psychological predisposition, and prior academic achievement—did not succeed in meaningfully reducing the bias at all, whether analyzed singly or in any combination. The topic preference and proxy-pretest domains were the only ones necessary (and sufficient) for bias reduction.

The story is not quite the same at the construct level. Across both outcomes, two constructs were sufficient for bias reduction, but they were not necessary for it. Analyses omitting the two constructs most responsible for bias reduction showed that the remaining constructs within proxy pretests and topic preference were sufficient for a meaningful bias reduction even though they were not individually effective in reducing bias. This is likely due to these constructs being jointly correlated with both treatment selection and potential outcomes at levels high enough to substitute for the deliberately omitted constructs that had turned out to be the most direct single approximations to the true selection process. Including the two most important constructs was not necessary for reducing bias in this study; however, including the two crucial domains was necessary.

These results demonstrate that strong ignorability can be established with quite different sets of covariates: (a) with a minimal set including only those nonredundant covariates that are both highly correlated with treatment and with potential outcomes, or (b) with a rich data set that may also include redundant covariates that are conditionally uncorrelated with treatment or potential outcomes. Within the PS-approach, redundant covariates are typically unrelated to potential outcomes but related to treatment selection.

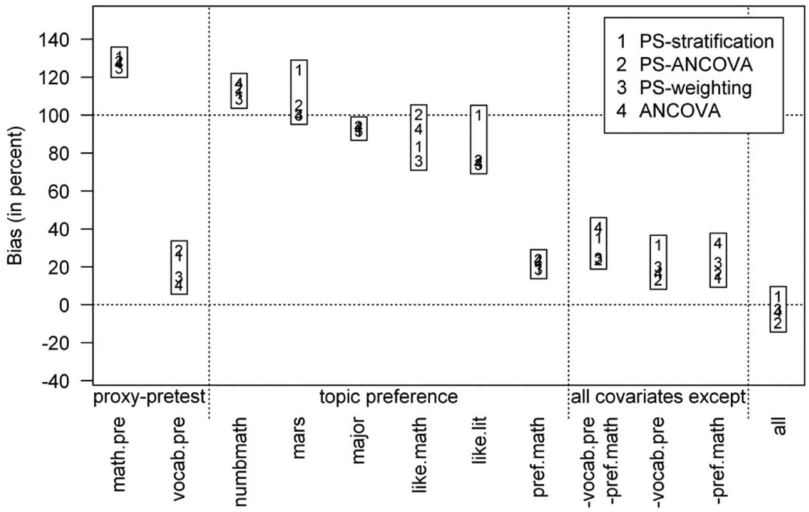

Figure 4. Remaining bias in vocabulary by single construct and analytic method (in the order of average bias). The single constructs are the mathematics and vocabulary proxy-pretests (math.pre, vocab.pre), number of math courses (numbmath), mathematics anxiety scale (mars), major field of study (major), liking mathematics (like.math), liking literature (like.lit), and preferring math over literature (pref.math). PS = propensity score; ANCOVA = analysis of covariance.

Though their inclusion is not required, it is nonetheless advisable because in practice we barely know whether they are really unrelated to potential outcomes. Moreover, a rich set of covariates might even lack the most effective single constructs, given that close substitutes for them are available from within the same construct domain.

The benchmark provided by the randomized experiment allowed us to investigate the relative importance of various covariate constructs and domains in the observational study. However, in practice we do not have an experimental benchmark. So what can be learned from this case study about how to select covariates? In prospectively planning a quasi-experiment, three different strategies for selecting covariates are possible. First, if all of the covariates that influence the selection process are measured and correctly modeled, then both theory and one clear empirical example tell us that all the selection bias can be removed (Diaz & Handa, 2006). The difficulty here is that complete knowledge of the selection process is rare in practice.

The second best strategy consists in theoretically and empirically grounded investigations of the selection mechanism before any quasi-experimental data are collected. This entails (a) theoretical reasoning regarding the expected selection mechanism and its relation to potential outcomes, (b) directly investigating selection in a pilot study, and (c) incorporating expert knowledge from substantive experts and third persons on the ground, such as administrators directly involved in the selection process (Rubin, 2007, 2008b). Such knowledge permits identifying the most crucial construct domains and single constructs within each domain and obviates wasting resources in the measurement of ineffective covariates. The results of our analysis indicate that researchers should primarily aim for constructs that are highly correlated with treatment (preferring mathematics over literature and liking mathematics in our study) and that are highly correlated with potential outcomes (real pretest measures or the probably less effective proxy-pretests as in our case). The most effective covariates for reducing selection bias are those that best predict both treatment selection and the outcome under consideration.

The third strategy is least desirable but should be considered as a fallback in situations with little or no information on the selection mechanism. This approach requires measuring a rich set of covariates covering a broad range of construct domains and constructs within each domain. One might then be lucky enough to get a set of covariates that establishes strong ignorability, though there is no independent justification for claiming that the ignorability assumption actually holds. In carefully measuring 23 constructs with 156 questionnaire items, Shadish et al. (2008) had luck. However, without the benchmark of the randomized experiment, who would have trusted them in the claim that their covariates sufficed for completely removing selection bias—particularly because they only had proxy-pretest measures instead of real pretest measures of the outcome? How could they otherwise have justified that their observed set of covariates rendered the selection mechanism ignorable?

The uncertainty about the ignorability of the assignment mechanism is even larger for observational studies that rely on archival data or other forms of secondary data. Such data are typically collected independently of the researchers' specific causal questions and without thought about the selection processes corresponding with the treatment at this time and in this setting. Although measures of demographics, prior academic achievement, and personality are often available from national or local archives, one implication of the present results is that they will sometimes, perhaps often, fail to remove all selection bias. This is certainly true of demographic variables when used alone because they have failed to be effective in two reviews of different within-study comparisons (Cook et al., 2008; Glazerman et al., 2003). Yet, many researchers make causal claims from the available covariates in archives that were rarely collected for the explicit purpose of

Table 5
Adjustment for Single Constructs: Effect Estimates and Bias for Mathematics Training

	PS-stratification[a]			PS-ANCOVA[b]			PS-weighting[c]			ANCOVA[d]			
Mathematics constructs	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\hat{\tau}$	*SE*	$\hat{b}\%$	$\overline{\hat{b}\%}$
Proxy-pretests													
Vocabulary pretest	5.28	0.50	128	5.24	0.48	123	5.31	0.49	131	5.33	0.55	133	129
Mathematics pretest	4.73	0.48	71	4.63	0.46	61	4.68	0.45	66	4.67	0.49	64	66
Topic preference													
Liking mathematics	4.16	0.48	11	3.92	0.46	−14	4.10	0.45	4	3.93	0.55	−14	−3
Liking literature	4.57	0.55	54	4.79	0.52	77	4.80	0.52	78	4.79	0.55	77	72
Prefer math over literature	4.17	0.51	12	4.07	0.53	1	4.16	0.50	11	4.06	0.58	1	6
Mathematics anxiety scale	5.00	0.53	99	4.90	0.52	89	5.00	0.51	100	5.00	0.55	100	97
No. of math courses	4.80	0.53	80	4.81	0.52	81	4.73	0.51	73	4.72	0.54	72	77
Major field of study	4.67	0.53	65	4.64	0.53	62	4.67	0.53	65	4.64	0.55	62	64
All covariates of all five domains except for													
Liking mathematics	3.99	0.47	3	3.88	0.42	−8	3.99	0.40	3	4.00	0.46	4	1
Prefer math over literature	3.72	0.41	−28	3.67	0.40	−32	3.82	0.38	−17	3.81	0.45	−19	−24
Liking math and prefer math	4.14	0.42	13	4.00	0.43	−1	4.14	0.39	13	4.17	0.44	16	10

Note. Construct sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). All treatment effect estimates $\hat{\tau}$ are based on (weighted) regression analyses with constructs included as main effects. Standard errors for the propensity score (PS) methods are based on 1,000 bootstrap samples (separate samples for each group with replacement) with refitted PSs, quintiles, and weights for each sample (predictors remained unchanged). Bias in quasi-experimental estimates is defined with respect to experimental estimates ($\hat{\tau}_Q - \hat{\tau}_E$). To account for differences in the randomized and quasi-experiment, we discarded cases outlying the range of the corresponding quasi-experimental PS-logit by 0.1 standard deviation. Hence, $\hat{\tau}_E$ slightly differs for each set of constructs. $\hat{b}\% = (\hat{\tau}_Q^{adj} - \hat{\tau}_E)/(\hat{\tau}_Q^{unadj} - \hat{\tau}_E) \cdot 100$ is the fraction of bias remaining after PS and/or covariance adjustment. A positive sign indicates an underadjustment with respect to the experimental effect, and a negative sign indicates an overadjustment. $\overline{\hat{b}\%}$ represents the average across all four adjustment methods. ANCOVA = analysis of covariance.

[a] Weighted least squares (WLS) regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{L}_i'\boldsymbol{\lambda} + \mathbf{D}_i'\boldsymbol{\delta} + \varepsilon_i$ with stratum weights; $\mathbf{L}_i$ are predictors based on PS-logits (linear, quadratic, and cubic term); $\mathbf{D}_i$ are four dummy variables representing PS-quintiles. [b] Ordinary least squares (OLS) regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{L}_i'\boldsymbol{\lambda} + \varepsilon_i$. [c] WLS regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ with PS-weights. [d] OLS regression $Y_i = \beta_0 + \tau Z_i + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$ (also for the randomized experiment but not for the unadjusted quasi-experiment).

contributing to a selection model for a specific causal agent. In the present case, bias reduction required measures of two domains: motivation (topic preference) or pretest performance on a knowledge test that could serve as a proxy for the missing pretest. Without these, the bias reduction achieved was minimal. Of course, it might have been more had we had a standard pretest, as occurs in longitudinal surveys with two or more waves of measurement. The absence of such a measure is a problem with the present study because we do not know whether it alone would have obscured the role of the proxy-pretest and perhaps even of the motivational domain. Future within-study comparisons should include such a measure because it reflects better causal practice in longitudinal survey research work. Without a strong theoretical and empirical justification that the available covariates establish an ignorable treatment mechanism, we should probably all be humble about the validity of whatever causal evidence and conclusions we want to claim.

Skepticism and humility are also appropriate for prospectively planned and carefully implemented quasi-experiments if selection mechanisms are likely to change just before or during the quasi-experiment (e.g., because of interventions by stakeholders or rumors concerning treatments). Even with grounded prior knowledge of selection-relevant constructs, it is still advisable to cover important construct domains with multiple measures and even to collect information from construct domains considered to be of less importance. These additional measures should also be carefully selected and supplement the targeted selection of covariates. Such a strategy minimizes the risk of unadjusted selection bias due to unobserved constructs but also to the unreliable measurement of observed ones. The reliable measurement of covariates is important whenever selection is on latent constructs (Steiner, Cook, & Shadish, in press), and using structural equation models might be a viable alternative to standard PS- and regression models, though they invoke further crucial assumptions (Kaplan, 1999; Steyer, 2005).

Our results also lead us to another note of caution about practice in observational study design and analysis. The balance achieved for the combinations of construct domains that were ineffective in bias reduction was as good as that achieved for the effective ones. This implies that balance is a necessary but insufficient condition for bias reduction. This is because achieving balance on observed covariates does not necessarily make the groups equivalent on unobserved but important covariates. Yet, without a yoked randomized experiment, how does one know that the strong ignorability assumption is met? Sensitivity analysis may help (Rosenbaum, 2002), but it is currently limited to assessing the impact of an unobserved covariate that is assumed to be as strongly correlated with treatment and outcome as the most influential single covariate other than a pretest measure of outcome (e.g., Hong & Raudenbush, 2006; Rosenbaum, 1986). This is reasonable but obviously no guarantee of assessing all hidden bias. Had topic preference not been measured in the present study, we would likely have underestimated hidden bias in mathematics with sensitivity tests based on any other covariates available. The sad reality is that

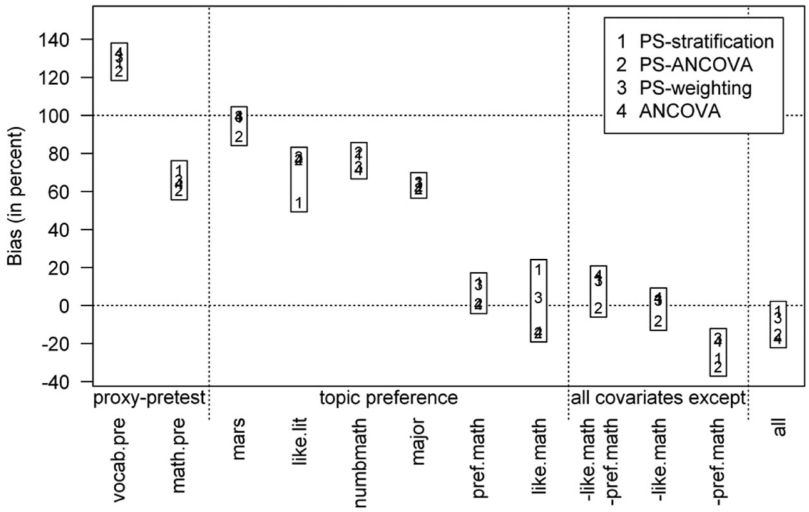

Figure 5. Remaining bias in mathematics by single construct and analytic method (in the order of average bias). The single constructs are the mathematics and vocabulary proxy-pretests (math.pre, vocab.pre), number of math courses (numbmath), mathematics anxiety scale (mars), major field of study (major), liking mathematics (like.math), liking literature (like.lit), and preferring math over literature (pref.math). PS = propensity score; ANCOVA = analysis of covariance.

balance in observed covariates alone does not ensure ruling out hidden bias (Rosenbaum, 1984) other than in studies with regression discontinuity or an independently and completely known selection process (e.g., Diaz & Handa, 2006).

The bias reduction achieved in this study depended not just on the measured covariates but probably also on well-matched locally and focally similar comparison groups in the experimental and nonexperimental conditions (Cook et al., 2008). The quasi-experiment involved treatment and comparison groups from the same introductory psychology courses in the same university who were then treated and measured in the same place by the same research personnel. Any treatment and comparison differences at pretest were induced by individual treatment selection preferences, not by any differences in location, procedure, or participant characteristics save for the preference for treatment in the quasi-experiment. These protections against confounding were more extensive than in existing within-study comparisons in which the nonrandom comparison group was mostly selected from a different locale and sometimes at a different time and with different dynamics for measuring the outcome (e.g., Agodini & Dynarski, 2004; Bloom et al., 2002; Dehejia & Wahba, 1999; Hotz, Imbens, & Mortimer, 1999). Focal local controls may do a better job of holding constant many potential confounds and of maximizing the area of common support between the different populations under analysis. If the treatment and comparison groups are inherently quite different, additional covariates that control for local and focal differences will likely have to be observed if strong ignorability is to be established. Even if these covariates are observed, creating a successfully balanced comparison group might fail to create a situation comparable with the treatment group because, say, the mean in the matched treatment sample is far different from the mean in the population at large. Situations in which the treatment and comparison group overlap little—that is, in which only a fraction of treatment and comparison cases share the common support region on the PS-logit—typically require the nonoverlapping cases to be deleted. This limits the generalizability of the causal estimates obtained unless it is reasonable to assume constant treatment effects.

The present results demonstrate that the choice of covariates is more important than the choice of the analytic method, assuming the analysis is competent and sensitive to the assumptions required. There was no uniformly best performing PS-method, and traditional ANCOVA nearly always did as well as any PS-method. The same lack of meaningful differences by mode of analysis is apparent from reviews of within-study comparisons in economics and social sciences (Bloom et al., 2002; Cook et al., 2008; Glazerman et al., 2003), meta-analytic reviews in epidemiology (Shah, Laupacis, Hux, & Austin, 2005; Stürmer et al., 2006), as well as in simulation studies (e.g., Schafer & Kang, 2008). All reveal that ANCOVA does as well as PS-methods when the ANCOVA model is correctly specified. However, PS-models have their own assumptions too, particularly as regards balance. Nonetheless, PS-methods have two clear advantages over ANCOVA methods: they are not so dependent on functional form assumptions, and pretreatment group differences can be balanced without knowing the outcome data, thereby ensuring a more objective estimation of causal treatment effects (Rubin, 2007, 2008b) and avoiding inflated Type I and II errors.

The findings we have presented are not generalizable without restriction. Singular features of this within-study comparison include the laboratory-like setting, a very specific educational treatment, the moderate complexity of the selection process, the modest degree of initial bias, and the specificity of the correlations achieved among covariates and between covariates and outcomes. The importance of covariate sets in reducing selection bias may be quite different for, say, labor market or health programs because of more complex selection processes and greater initial bias. However, a close replication of Shadish et al. (2008) in Germany by

Table 6
Correlation of Constructs With Treatment Selection, Vocabulary, and Mathematics Outcomes

	Correlation with		
Constructs	Treatment selection $R_{pb}(x, z)$[a]	Vocabulary $R(x, y_v)$[b]	Mathematics $R(x, y_m)$[b]
1. Demographics	0.22	0.45	0.42
Age	0.02	0.12	−0.26
Gender (male)	−0.06	0.19	0.13
Ethnicity (Caucasian)	0.18	0.30	0.04
Ethnicity (African American)	−0.14	−0.30	−0.10
Marital status (married)	0.00	−0.11	−0.24
Credit hours completed at university	0.03	0.19	0.17
2. Proxy-pretests	0.24	0.55	0.49
Vocabulary proxy-pretest (vocab.pre)	0.17	0.53	0.24
Mathematics proxy-pretest (math.pre)	−0.09	0.34	0.48
3. Prior academic achievement	0.07	0.51	0.60
ACT score	0.03	0.49	0.53
High school GPA	−0.04	0.05	0.47
College GPA	−0.03	0.13	0.25
4. Topic preference	0.43	0.37	0.46
Liking mathematics (like.math)	−0.36	−0.04	0.36
Liking literature (like.lit)	0.16	0.18	−0.17
Preferring math over literature (pref.math)	−0.38	−0.23	0.28
Mathematics anxiety scale (mars)	0.00	−0.05	−0.18
No. of math courses (numbmath)	−0.13	0.11	0.26
Major field of study (major)	−0.19	−0.05	0.24
5. Psychological predisposition	0.18	0.42	0.30
Extroversion (Big 5)	0.09	−0.11	−0.15
Agreeableness (big 5)	0.10	0.10	−0.03
Conscientiousness (Big 5)	−0.13	−0.16	−0.15
Emotional Stability (Big 5)	−0.02	−0.14	−0.16
Openness to Experience (Big 5)	0.05	0.24	0.08
Beck Depression Inventory	−0.01	0.14	0.18

[a] (Multiple) point-biserial correlation. [b] (Multiple) correlation averaged across vocabulary and mathematics groups, that is, $R = (R_v + R_m)/2$.

Pohl, Steiner, Eisermann, Soellner, and Cook (2009) also succeeded in removing all the selection bias in the quasi-experiment despite a different selection mechanism and different degrees of initial bias. Additional analyses of this data set also indicated that proxy-pretest measures and topic preference were once again necessary and sufficient for removing all the selection bias (Cook, Steiner, & Pohl, 2009).

Notwithstanding the limitations of our case study for generalization, it has shown that total bias reduction in an observational study can be achieved when (a) the outcome-related part of the

Table 7
Distribution of Reported Reasons for Treatment Choice and Multiple Point-Biserial Correlation With Construct Domains

	Vocabulary				Mathematics			
Covariate set	Liking	Avoiding	Self-improvement	Self-efficacy	Liking	Avoiding	Self-improvement	Self-efficacy
				Distribution[a] (%)				
	18	21	42	17	30	8	47	11
				Correlation				
dem	.21	.21	.21	.23	.33	.34	.39	.29
pre	.24	.25	.36	.12	.29	.16	.41	.15
aca	.06	.21	.13	.18	.27	.31	.16	.22
top	.31	.36	.47	.23	.54	.43	.60	.26
psy	.31	.29	.25	.11	.37	.42	.31	.24

Note. Construct sets are demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy).
[a] Missing percentages of distribution to 100% belong to the "other reasons" category for treatment choice.

selection process is quite specific, in this case limited to two constructs from two domains; (b) a set of constructs is available that is individually less successful in bias reduction but comes from within the most crucial domains for bias reduction, thus implying the need to measure crucial domains but not necessarily crucial single constructs within these domains; and (c) a combination of expert judgment, theory, observation, and common sense is used to arrive at the rich set of domains, constructs within these domains, and even items within these constructs that might explain the selection process and be correlated with the outcome. This pushes the design of quasi-experiments to a new appreciation of the value of broad coverage when sampling covariates.

References

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics, 86,* 180–194.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22,* 207–244.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61,* 962–972.

Beck, A. T., & Beck, R. W. (1972). Screening depressed patients in family practice: A rapid technic. *Postgraduate Medicine, 51,* 81–85.

Black, D., Galdo, J., & Smith, J. A. (2007). *Evaluating the regression discontinuity design using experimental data.* Available from http://economics.uwo.ca/newsletter/misc/2009/smith_mar25.pdf

Bloom, H. S., Michalopoulos, C., Hill, C. J., & Lei, Y. (2002). *Can nonexperimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs?* Washington, DC: Manpower Demonstration Research Corporation.

Buddelmeyer, H., & Skoufias, E. (2004). *An evaluation of the performance of regression discontinuity design on PROGRESA* (World Bank Policy Research Working Paper No. 3386; IZA Discussion Paper No. 827). Available from http://ssrn.com/abstract=434600

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24,* 295–313.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics, 142,* 636–654.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27,* 724–750.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). Assessing how bias reduction is influenced by covariate choice, unreliability, and data analytic mode: An analysis of different kinds of within-study comparisons in different substantive domains. *Multivariate Behavioral Research, 44,* 828–847.

Dehejia, R., & Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association, 94,* 1053–1062.

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator. *The Journal of Human Resources, 41,* 319–345.

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University, 73,* 24–33.

Educational Testing Service. (1962). *Vocabulary Test II (V-2): Kit of factor referenced cognitive tests.* Princeton, NJ: Author.

Educational Testing Service. (1993). *Arithmetic Aptitude Test (RG-1): Kit of factor referenced cognitive tests.* Princeton, NJ: Author.

Faust, M. W., Ashcraft, M. H., & Fleck, D. E. (1996). Mathematics anxiety effects in simple and complex addition. *Mathematical Cognition, 2,* 25–62.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy, 589,* 63–93.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4,* 26–42.

Goldberger, A. S. (1972). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion Paper No. 123). Madison, WI: Institute for Research on Poverty, University of Wisconsin—Madison.

Heckman, J. J., Ichimura, H., Smith, J. C., & Todd, P. (1998). Characterizing selection bias. *Econometrica, 66,* 1017–1098.

Heckman, J. J., & Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics, 86,* 30–57.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete data perspectives* (pp. 51–60). New York, NY: Wiley.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology, 2,* 259–278.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–970.

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis, 31,* 54–81.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101,* 901–910.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (1999). *Predicting the efficacy of future training programs using past experiences* (NBER Technical Working Paper 238). Cambridge, MA: National Bureau of Economic Research.

Kang, J., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating population means from incomplete data. *Statistical Science, 26,* 523–539.

Kaplan, D. (1999). An extension of the propensity score adjustment method for the analysis of group differences in MIMIC models. *Multivariate Behavioral Research, 34,* 467–492.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via propensity score in estimation of causal treatment effects: A comparative study. *Statistical Medicine, 23,* 2937–2960.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge, England: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, England: Cambridge University Press.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis, 31,* 463–479.

R Development Core Team. (2007). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90,* 122–129.

Robins, J. M., & Rotnitzky, A. (2001). Comment on "Inference for semiparametric models: Some questions and answers" by Bickel and Kwon. *Statistica Sinica, 11,* 920–936.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79,* 41–48.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11,* 207–224.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2,* 169–188.

Rubin, D. B. (2006). *Matched sampling for causal effects.* Cambridge, England: Cambridge University Press.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26,* 20–36.

Rubin, D. B. (2008a). The design and analysis of gold standard randomized experiments. Discussion of "Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment" by W. R. Shadish, M. H. Clark, and P. M. Steiner. *Journal of the American Statistical Association, 103,* 1350–1353.

Rubin, D. B. (2008b). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2,* 808–840.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52,* 249–264.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13,* 279–313.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103,* 1334–1343.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton-Mifflin.

Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. *Journal of Clinical Epidemiology, 58,* 550–559.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (in press). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics.*

Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology, 1,* 39–64.

Steyer, R., Gabler, S., von Davier, A. A., & Nachtigall, C. (2000). Causal regression models: II. Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online, 5,* 55–87.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology, 59,* 437–447.

Received May 12, 2008
Revision received November 28, 2009
Accepted December 2, 2009 ■