Abstract Title Page

Title: Sensitivity Analysis for Multivalued Treatment Effects: An Example of a Cross-country Study of Teacher Participation and Job Satisfaction

Authors and Affiliations: Chi Chang, Michigan State University

Abstract Body

Background / Context

It is known that interventions are hard assign randomly to subjects in social psychological studies, because randomized control is difficult to implement strictly and precisely. We might be able to control how the treatment is implemented, but to what degree that the treatment is implemented might vary, based on different kinds of conditions. For example, if little children are involved, the degree to which the treatment is implemented might be altered or adapted based on how long they can hold their attention to the examiner. Thus, in nonexperimental studies and observational studies, controlling the impact of covariates on the dependent variables and addressing the robustness of the inferences from the results is very important.

While ways to obtain accurate average treatment effects of a single and dichotomous treatment in research designs has been studied prudently in decades, average treatment effects of multiple treatments or a multivalued treatment are rarely studied in causal inference of educational research in recent years (Feng, Zhou, Zou, Fan, & Li, 2012; Hong, 2012; Landsman & Pfeiffer, 2013). There are relatively more research studies regarding multiple treatments in health science, epidemiology and the biology area, since it is necessary to give treatment trials according to patients' conditions. For example, dosage and a combination of medicines are great examples of using multivalued treatments and multiple treatments (Imbens, 2000; Shapiro, Kazdin, & McGonigle, 1982).

A sensitivity framework developed by Frank (2000) quantified the impact of a potential confounding variable on statistical inference with regard to a regression coefficient. According to Frank's (2000) approach, an index of how large is large enough for the impact of a potential confounding variable to change the inference can be identified by calculating the threshold of the inference, which is named the Impact Threshold of a Confounding Variable (ITCV) index. Also, for external validity concerns, the robustness of inference can be justified by examining what proportion of the sample size has to be replaced for altering the inference (Frank & Min, 2007).

Purpose / Objective / Research Question / Focus of Study

This study firstly focused on examining the sensitivity of a multivalued treatment effect. Multivalued treatment is often seen in practice, but it is rarely discussed for its causation association possibility in previous research. Second, a small simulation study of ITCV was conducted to examine to what degree the ITCV index is affected by test statistics and sample size. Third, whether the different propensity score weighting methods were sensitive enough to examine different levels of the treatment, and how robust the inference can be made on each value treatment, were demonstrated using TALIS data.

The purpose of the study was to examine the sensitivity of the inference when: 1) the ITCV index is applied to a multivalued treatment, and 2) different propensity score weighting is applied to a multivalued treatment scheme. To put the relationships in the causal inference framework, the three research questions of interest were: 1) Does teacher participation affect their job satisfaction? 2) What must be the conditions in the alternative sample to invalidate the inference? 3) Which propensity score weighting methods can help to identify the effect of teacher participation the most?

Population / Participants / Subjects

66,434 teachers in 367 schools of the 21 countries that participated in the Teaching and Learning International Survey (TALIS) 2008 dataset from the Organization for Economic Co-Operation and Development (OECD) were used. We involved all countries that were available for teacher level and country level. The list of countries and frequency table of how each country was involved can be found in Table 1.

Intervention / Program / Practice

The treatment variable is a 4- level Likert scale at never (1,948 responses), seldom (9,721 responses), quite often (35,279 responses), and very often (19,897 responses). These multiple levels were categorized with different cut-points into three dichotomous variables of interventions using different thresholds: 1) never, vs. seldom and above (TP1: 1 vs. 2, 3, 4); 2) never and seldom vs. quite often and very often (TP2: 1, 2 vs. 3, 4); 3) quite often and below vs. very often (TP3: 1, 2, 3 vs. 4).

Significance / Novelty of study

This study first focused on examining the sensitivity of a multivalued treatment effect. Multivalued treatment is often seen in practice, but its causation association possibility in previous research is rarely discussed. Second, a small simulation study of ITCV is provided to examine to what degree the ITCV index is affected corresponding to the test statistics and sample size. Third, whether the different propensity score weighting methods were sensitive enough to examine different levels of the treatment, and how robust the inference can be made on each value of the treatment, are demonstrated using TALIS data.

The importance of validity issues cannot be overemphasized. Propensity score methods have been recently applied to educational, sociological, and psychological research to make causal inferences; however, these research studies were rarely followed by sensitivity analysis to validate the results they proposed from the data. Propensity score methods tend to approximate a non-experimental study as a randomized assignment condition; however, all in all, it is not always the panacea. Sensitivity analysis helps raise concerns based on given data and the association of the variables of interest, and insure the validity and the credibility of the causal results. Therefore, with this research, the authors would like to demonstrate the importance of this issue and suggest including sensitivity analysis as a final procedure when causal inferences are made in application studies.

Statistical, Measurement, or Econometric Model

Sensitivity analysis results from three methods were compared: unweighted least square, propensity score weighting method, and Estimate of Treatment for People at the Margin of Indifference (EOTM) weighting (Hirano & Imbens, 2001; Robins, Rotnitzky, & Zhao, 1995). Available confounding variables included their classroom disciplinary climate (CCLIMATE), teacher-student relations (TSRELAT), teachers' self-efficacy (SELFEF), structuring practices (TPSTRUC), student-oriented practices (TPSTUD), enhanced activities (TPACTIV), direct transmission beliefs about teacher (TBTRAD), constructivist beliefs about teaching (TBCONS), exchange and coordination for teaching (TCEXCHAN), professional collaboration, percentage of professional development that is compulsory (TCCOLLAB), age group (AgeGrp), employment time (EmplyTime), whether they are working in multiple schools (WorkASchl), their highest education (HighEdu), and their experience of teaching (ExpTr). These were included for multivariate extension. Since the treatment variables were dichotomized based on different thresholds, their effects were examined individually with three kinds of propensity score methods. SAS and R were used to conduct this study.

Usefulness / Applicability of Method

Sensitivity study is important since validity issues cannot be overemphasized in causal inference studies. This study examined the effect of the multivalued treatment, and the sensitivity of the inference made in the research can serve as the evidence for researchers to support the robustness of the inference for generalization.

TALIS data was used in examining the approach of dichotomizing a multivalued treatment and the sensitivity of the inference. It can easily be applied to different levels of a multivalued treatment since subjects can be assumed to choose levels of the treatment independently and exclusively. While the effects of different levels of the treatment were compared across different propensity score weighting methods, a relatively better method was identified, and the strength of the inference was validated. For

facilitating the process of sensitivity, an R function was developed based on Frank's (2000) approach and is available upon request.

Research Design

This study first focused on examining the sensitivity of a multivalued treatment effect. Second, a small simulation study of ITCV was developed to examine to what degree the ITCV index is affected corresponding to the test statistics and sample size. Third, whether three propensity score weighting methods were sensitive enough to examine different levels of treatments, and how robust the inference could make on each value treatment was demonstrated using TALIS data.

In the empirical illustration, teachers were surveyed through a questionnaire with respect to their classroom practice, professional competence, related beliefs and attitudes, professional activities, classroom environment, school environment, and overall job-related attitudes, such as self-efficacy and job satisfaction. In this study, the treatment variable was the item "In this school, the principal and teachers act to ensure that education quality issues are a collective responsibility," and the outcome variable was job satisfactory.

Findings / Results

When the analysis was extended to a multivariate situation, with all possible confounding variables included, the impact of the unmeasured confounding variable had to be larger than .090 to invalid the inference. Table 2 and Table 3 indicate the ITCV index for different treatment levels in the univariate scenario and the multivariate scenario. A similar phenomenon in both scenarios is that the ITCV index increases when the cut scores shift upward. The inference goes stronger once the cut score is chosen between level 3 (quite often) and level 4 (very often). However the magnitude of the difference between TP2 and TP3 is smaller than that between TP1 and TP2. In addition, the ITCV index in univariate scenario is consistently larger than multivariate scenario across all three treatment variables. The effects of 21 countries and 367 schools were also taken into consider in these analyses for controlling purpose.

Because the threshold for inference is affected by the t-test's critical value and the sample size, and the ITCV index is affected by the sample size, the threshold of the inference, and the observed t, a small simulation was affected and is shown in Table 4. While the critical value was fixed at 1.96, the observed t statistic was fixed at 2, and the sample size was the only variable changing from 100 to 1000, ITCV dropped from 0.004 to 0.001. While the sample size was fixed and the observed t statistics was the only variable changing from 2 to 9, the ITCV index had a greater change, from 0.118 to 0.593. These two results are within expectations; as the sample size change, the power usually increases. While the t statistic is fixed along with increasing sample size, but the t statistics still stay the same, the threshold of inference drops and correspondingly the ITCV index drops accordingly. In contrast, if the sample size is fixed, and the observed t statistic increases, even though the thresholds of inference do not change, a greater impact would be needed to invalidate the inference. That is, the inference would be stronger.

As presented in Table 5, the largest impact of the confounding variable, teacher-student relations, for TP1 was 0.049, for TP2 was 0.086, for TP3 was 0.102, which were smaller than the threshold respectively. To put it simply, the unmeasured confound would have to be around one and half as much as the strongest observed confounding covariates here to invalid the statistic inference; otherwise, the causal inference statement that teacher participation, or we can say, teachers' collective efficacy, affects their job satisfaction can hold.

As for the robustness of the inference, more than 97% of teachers would have to be replaced with others for whom teacher participation has no effect to invalid the inference in a combined sample. In other words, 97% of the estimate must be due to bias to invalidate the inference so that the inference can be changed. In addition, if half of the teachers in the sample were replaced with a different inference, then the correlation between teacher participation and job satisfaction of replaced sample would have to be less than -.2268 to alter the inference statement.

Teachers' self-efficacy is commonly viewed as the proxy of job satisfaction. Table 6 indicates the results of absorption of teachers' self-efficacy among three different treatment variables. Most of the

impacts of the covariates dropped while controlling for self-efficacy. In other words, teacher self-efficacy shows as a good covariate that absorbs the impacts of other covariates on the association between teacher participation and job satisfaction. But, the magnitude of absorption of teacher self-efficacy did not show consistency across three treatment situation. For TP1, teacher self-efficacy absorbed better for EmplyStatus, EmplyTime, TBTRAD, TPACTIV, and TSRELAT. For TP2, it worked better for ExpTr, HighEdu, TPSTRUC, and TPSTUD. As for TP3, it absorbed more efficiently on variables AgeGrp, CCLIMATE, TBCONS, TCCOLLAB, TCEXCHAN, WorkASchl, and Gender.

The treatment effects and the standard errors of covariates with treatment variables TP1, TP2, and TP3 were identified, respectively, with unweighted least square regression, treatment effect for treated weighted by propensity score analysis, and with treatment effect for control weighted by EOTM. The separate results using different treatment variables are shown in Table 7, 8, and 9, and the covariate balance checks of these three across different weighting methods are presented in Figure 1, 2, and 3. Table 10 compares the treatment effect across TP1, 2, and 3. Applying different methods, the effect of TP1 can range from 0.208- 0.256, the effect of TP2 can range from 0.136 to 0.154, and that of TP3 can range from 0.078 to 0.09. Across different methods, TP1 effects were consistently larger than TP2 effects, and TP2 effects were consistently larger than TP3. The results show that the treatment variable with the cut point made between never and seldom showed the largest effect in this study. Across different weighting methods, the effect of teacher participation using EOTM weights showed the smallest standard error, and it were consistently the smallest across the three random variables as well.

Conclusions:

The findings of this study validate the inference that teacher participation can affect their job satisfaction. The results also use an index for the threshold for this inference statement, and for examining the robustness of the sample. The indexed threshold necessary for the impact and necessary to invalidate the inference, and the threshold for sample replacement provide power arguments for internal validity and external validity, respectively, and they also show a clear warrant for sensitivity analysis. Because this is an international survey, the large sample size also makes this argument stronger.

Propensity score weighting analysis has been greatly discussed recently, because it keeps every case in the sample during the analysis process, instead of disregarding unmatched cases, like the propensity score matching method does (Hirano & Imbens, 2001; Tan, 2006). EOTM weighting shows better results, with consistently small standard error of estimates, given the sample in this study.

Causal inference statements have been avoided in social science study, because confounding variables cannot be strictly controlled as they can in a lab research or in an experiment. Frank's approach gives a different perspective to identify the threshold of inference. Once a potential confounding variable is found, whether it can threaten the inference statements of the study can be identified based on its relationship with the treatment variable and the outcome variable. The approach can be extended to a multivariate situation by applying the information from the coefficient of the determinant in regression models. As issues regarding the policy making, this approach of sensitivity analysis is useful and helpful once applied to related studies and to reports to reexamine the robustness of researches, because causal statements and the validation of the inferences are more valuable. In addition, if the policy is conducted accordingly, there can be more evidence to support the arguments and to justify the internal and external validity of the inference of interest.

Different cut points in the treatment variable were applied to create three dichotomous treatment variables, and the results of these were compared in this study. The results were within expectation, since the lowest cut point made the size of treatment group larger, which allowed for a stronger treatment effect. Similar results, in contrast, can be found when identifying the third treatment variable, when the highest cut point was applied. The treatment effect of the last treatment variable was the smallest. At any rate, the three dichotomous variables were significantly different from zero, indicating their significant effect at any level. Multivalued treatments have not been broadly explored in the social science area, while the Likert scale with levels of four, five, or above has greatly been used in related studies. Knowing how to deal with multivalued treatments instead of losing information by dichotomizing a treatment effect

at a random cut point might be a valuable and warranted topic and a critical issue to investigate in future studies.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., & Li, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine, 31*(7, SI), 681–697. doi:10.1002/sim.4168

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research, 29*(2), 147–194. doi:10.1177/0049124100029002001

Frank, K., & Min, K.-S. (2007). Indices of robustness for sample representation. *Sociological Methodology, 37*, 349–XII. Retrieved from http://ezproxy.msu.edu/login?url=http://search.proquest.com/docview/216130365?accountid=12598

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*(3-4), 259–278. Retrieved from http://ezproxy.msu.edu/login?url=http://search.proquest.com/docview/228926421?accountid=12598

Hong, G. (2012). Marginal mean weighting through stratification: A generalized method for evaluating multivalued and multiple treatments With nonexperimental data. *Psychological Methods, 17*(1), 44–60. doi:10.1037/a0024918

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706–710. doi:10.2307/2673642

Landsman, V., & Pfeiffer, R. M. (2013). On estimating average effects for multiple treatment groups. *Statistics in Medicine, 32*(11), 1829–1841. doi:10.1002/sim.5690

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90*(429), 106–121. Retrieved from http://ezproxy.msu.edu/login?url=http://search.proquest.com/docview/38562598?accountid=12598

Shapiro, E. S., Kazdin, A. E., & McGonigle, J. J. (1982). Multiple-treatment interference in the simultaneous- or alternating-treatments design. *Behavioral Assessment, 4*, 105–115.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association, 101*(476), 1607–1618. doi:10.2307/27639776

Appendix B. Tables and Figures

Not included in page count.

TABLE 1
Frequency Table of Countries involved in this study

Country	*Frequency*	*Percent*
Australia	2275	3.29
Austria	4246	6.14
Belgium (Flemish Community)	3473	5.03
Brazil	5532	8
Bulgaria	3796	5.49
Denmark	1722	2.49
Estonia	3154	4.56
Hungary	2934	4.25
Italy	5213	7.54
Korea	2970	4.3
Lithuania	3535	5.12
Malaysia	4248	6.15
Malta	1142	1.65
Mexico	3368	4.87
Norway	2458	3.56
Poland	3184	4.61
Portugal	3046	4.41
Slovak Republic	3157	4.57
Slovenia	3069	4.44
Spain	3362	4.86
Turkey	3224	4.67

TABLE 2

ITCV Indices of Teacher Participation on Job Satisfaction in the Univariate Condition

Treatment level	*n*	*r#*	*observed t*	*r (x,y)*	*ITCV*	*r(x,cv)*	*r(y,cv)*	*Robustness to % bias*	*Replacement correlation*
TP1	66434	0.008	31.08	0.120	0.113	0.336	0.336	0.936	-0.105
TP2	66434	0.008	47.66	0.182	0.176	0.419	0.419	0.958	-0.167
TP3	66434	0.008	51.84	0.197	0.191	0.437	0.437	0.961	-0.182

Note. r# is the threshold of the inference.

TABLE 3

ITCV Indices of Teacher Participation on Job Satisfaction in the Multivariate Condition

Treatment Level	*Number of Covariates*	*r#*	R^2 *(x,z)*	R^2 *(y,z)*	*ITCV*	*r(x,cv)*	*r(y,cv)*	*Robustness to % bias*	*Replacement correlation*
TP1	404	0.008	0.090	0.303	0.090	0.321	0.281	0.936	-0.104
TP2	404	0.008	0.201	0.303	0.131	0.374	0.350	0.958	-0.167
TP3	404	0.008	0.282	0.303	0.135	0.370	0.365	0.961	-0.182

TABLE 4
A Simulation Result for the Sensitivity of ITCV Indices in Univariate Condition

		observed		
n	*r#*	*t*	*r (x,y)*	*ITCV*
100	0.195	2	0.198	0.004
200	0.138	2	0.141	0.003
300	0.113	2	0.115	0.002
400	0.098	2	0.100	0.002
500	0.088	2	0.089	0.002
600	0.080	2	0.082	0.002
700	0.074	2	0.075	0.002
800	0.069	2	0.071	0.001
900	0.065	2	0.067	0.001
1000	0.062	2	0.063	0.001
100	0.195	3	0.290	0.118
100	0.195	4	0.375	0.223
100	0.195	5	0.451	0.318
100	0.195	6	0.518	0.402
100	0.195	7	0.577	0.475
100	0.195	8	0.629	0.538
100	0.195	9	0.673	0.593

TABLE 5
The Impacts of All Covariates in Three Different Treatments Scenarios

Variables	*Impact _TP1*	*TP1*	*Impact _TP2*	*TP2*	*Impact _TP3*	*TP3*	*JobSat*
TSRELAT	0.049	0.160	0.086	0.280	0.102	0.333	0.308
TCEXCHAN	0.025	0.151	0.037	0.218	0.030	0.176	0.168
SELFEF	0.024	0.052	0.051	0.113	0.085	0.188	0.454
TCCOLLAB	0.015	0.130	0.021	0.182	0.016	0.142	0.116
CCLIMATE	0.010	0.047	0.017	0.083	0.017	0.084	0.206
TPSTUD	0.005	0.052	0.012	0.125	0.021	0.226	0.094
TPSTRUC	0.005	0.042	0.013	0.116	0.022	0.193	0.112
TPACTIV	0.003	0.028	0.010	0.092	0.021	0.201	0.105
HighEdu	0.000	-0.018	0.000	-0.019	0.001	-0.051	-0.022
COMPULPD	0.000	0.013	0.001	0.023	-0.001	-0.037	0.024
EmplyStatus	0.000	0.024	0.001	0.067	0.001	0.068	0.012
TBCONS	0.000	0.004	0.000	0.003	0.002	0.050	0.049
genderr	0.000	0.021	0.000	0.052	0.000	0.071	0.001
WorkASchl	0.000	-0.012	0.000	-0.035	-0.001	-0.070	0.008
ExpTr	0.000	-0.005	0.000	0.003	0.000	-0.009	0.021
EmplyTime	0.000	-0.004	0.001	0.022	0.001	0.062	0.023
AgeGrp	0.000	-0.010	0.000	0.010	0.000	-0.010	0.030
TBTRAD	-0.001	-0.012	-0.003	-0.028	-0.007	-0.064	0.113

TABLE 6
Absorb Index of Self-Efficacy in Different Treatment Scenarios

Variable	*TP1 (1 vs. 2 and above)*	*TP2 (1, 2 vs. 3, 4)*	*TP3 (3 and below vs. 4)*	*Jobsat_post*	*JobSat*
AgeGrp	0.997	0.965	1.010	0.001	0.030
CCLIMATE	0.660	0.680	0.735	0.100	0.206
COMPULPD	0.845	0.839	1.028	0.002	0.024
EmplyStatus	0.140	0.268	0.127	0.020	0.012
EmplyTime	0.674	0.174	0.217	0.025	0.023
ExpTr	0.959	1.746	0.952	-0.008	0.021
HighEdu	1.029	2.555	1.489	-0.041	-0.022
TBCONS	-0.056	1.548	1.804	-0.024	0.049
TBTRAD	1.113	1.120	1.073	0.022	0.113
TCCOLLAB	0.944	0.940	0.946	0.009	0.116
TCEXCHAN	0.855	0.863	0.884	0.038	0.168
TPACTIV	0.978	0.949	0.950	0.010	0.105
TPSTRUC	1.082	1.093	1.086	-0.014	0.112
TPSTUD	1.188	1.217	1.176	-0.026	0.094
TSRELAT	0.601	0.597	0.562	0.187	0.308
WorkASchl	0.547	0.533	0.707	0.003	0.008
Gender	-17.635	-17.665	-17.336	-0.010	0.001

TABLE 7
The Effects of Four Propensity Weighting Methods in TP1 Scenario

	unweighted	*se.unw*	*weight propensity (treatment effect for treated)*	*se.ttw*	*weight propensity (treatment effect for control)*	*se.tcw*	*weight EOTM*	*se.EOTM*
Intercept.	2.950	0.027	2.881	0.021	2.744	0.136	2.877	0.021
TP1	0.223	0.015	0.256	0.004	0.208	0.025	0.255	0.004
CCLIMATE	0.056	0.002	0.079	0.002	0.077	0.012	0.079	0.002
TSRELAT	0.096	0.003	0.084	0.002	0.083	0.012	0.084	0.002
SELFEF	0.220	0.003	0.199	0.002	0.217	0.012	0.199	0.002
TPSTRUC	-0.005	0.003	-0.007	0.002	-0.013	0.014	-0.007	0.002
TPSTUD	-0.041	0.004	-0.093	0.004	-0.084	0.025	-0.093	0.004
TPACTIV	0.035	0.004	0.065	0.003	0.079	0.022	0.064	0.003
TBTRAD	0.018	0.002	0.003	0.002	0.018	0.013	0.003	0.002
TBCONS	-0.015	0.003	0.012	0.002	-0.021	0.013	0.011	0.002
TCEXCHAN	0.012	0.003	-0.017	0.002	0.032	0.014	-0.015	0.002
TCCOLLAB	0.002	0.003	0.004	0.002	0.006	0.016	0.004	0.002
COMPULPD	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gender	0.023	0.005	0.074	0.004	0.035	0.026	0.073	0.004
AgeGrp	0.003	0.004	0.012	0.003	-0.013	0.019	0.011	0.003
EmplyTime	0.016	0.005	0.028	0.004	0.031	0.026	0.028	0.004
WorkASchl	-0.003	0.007	0.015	0.006	0.060	0.044	0.016	0.006
EmplyStatus	0.016	0.004	0.028	0.004	0.026	0.024	0.028	0.003
HighEdu	-0.019	0.003	-0.038	0.003	-0.007	0.017	-0.037	0.003
ExpTr	-0.006	0.002	-0.011	0.002	0.007	0.012	-0.011	0.002

TABLE 8
The Effects of Four Propensity Weighting Methods in TP2 Scenario

	unweighted	*se.unw*	*weight propensity (treatment effect for treated)*	*se.ttw*	*weight propensity (treatment effect for control)*	*se.tcw*	*weight EOTM*	*se.EOTM*
Intercept.	3.053	0.023	3.059	0.019	3.027	0.045	3.050	0.017
TP2	0.141	0.006	0.154	0.004	0.136	0.008	0.152	0.003
CCLIMATE	0.055	0.002	0.058	0.002	0.063	0.004	0.059	0.002
TSRELAT	0.089	0.003	0.075	0.002	0.081	0.005	0.076	0.002
SELFEF	0.220	0.003	0.216	0.002	0.225	0.005	0.217	0.002
TPSTRUC	-0.007	0.003	-0.015	0.002	-0.013	0.005	-0.014	0.002
TPSTUD	-0.046	0.004	-0.053	0.004	-0.061	0.009	-0.054	0.003
TPACTIV	0.038	0.004	0.049	0.003	0.049	0.008	0.048	0.003
TBTRAD	0.017	0.002	0.011	0.002	0.014	0.005	0.012	0.002
TBCONS	-0.014	0.003	-0.009	0.002	-0.019	0.005	-0.011	0.002
TCEXCHAN	0.010	0.003	-0.002	0.002	0.018	0.005	0.001	0.002
TCCOLLAB	-0.001	0.003	0.002	0.002	0.008	0.005	0.003	0.002
COMPULPD	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gender	0.021	0.005	0.032	0.004	0.033	0.009	0.032	0.004
AgeGrp	0.002	0.004	0.007	0.003	-0.001	0.007	0.006	0.003
EmplyTime	0.014	0.005	0.013	0.004	0.018	0.009	0.014	0.004
WorkASchl	0.003	0.007	-0.006	0.006	0.003	0.015	-0.005	0.006
EmplyStatus	0.015	0.004	0.007	0.003	0.015	0.008	0.009	0.003
HighEdu	-0.020	0.003	-0.022	0.003	-0.020	0.006	-0.022	0.002
ExpTr	-0.005	0.002	-0.008	0.002	-0.002	0.004	-0.007	0.002

TABLE 9
The Effects of Four Propensity Weighting Methods in TP3 Scenario

	unweighted	*se.unw*	*weight propensity (treatment effect for treated)*	*se.ttw*	*weight propensity (treatment effect for control)*	*se.tcw*	*weight EOTM*	*se.EOTM*
Intercept.	3.143	0.023	3.122	0.029	3.141	0.020	3.131	0.016
TP3	0.086	0.005	0.090	0.006	0.078	0.004	0.082	0.003
CCLIMATE	0.056	0.002	0.054	0.003	0.055	0.002	0.055	0.002
TSRELAT	0.090	0.003	0.087	0.003	0.095	0.002	0.091	0.002
SELFEF	0.216	0.003	0.206	0.003	0.230	0.002	0.222	0.002
TPSTRUC	-0.007	0.003	-0.004	0.004	-0.004	0.002	-0.004	0.002
TPSTUD	-0.043	0.004	-0.041	0.005	-0.050	0.004	-0.047	0.003
TPACTIV	0.033	0.004	0.035	0.005	0.033	0.003	0.033	0.003
TBTRAD	0.018	0.002	0.023	0.003	0.018	0.002	0.020	0.002
TBCONS	-0.018	0.003	-0.013	0.003	-0.015	0.002	-0.014	0.002
TCEXCHAN	0.013	0.003	0.010	0.004	0.011	0.002	0.010	0.002
TCCOLLAB	0.002	0.003	-0.005	0.004	0.003	0.002	0.001	0.002
COMPULPD	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gender	0.020	0.005	0.014	0.007	0.024	0.004	0.022	0.004
AgeGrp	0.003	0.004	0.008	0.005	0.001	0.003	0.004	0.003
EmplyTime	0.014	0.005	0.025	0.006	0.022	0.004	0.024	0.003
WorkASchl	0.000	0.007	0.002	0.009	-0.004	0.007	-0.002	0.005
EmplyStatus	0.016	0.004	0.017	0.005	0.018	0.004	0.018	0.003
HighEdu	-0.019	0.003	-0.020	0.004	-0.021	0.003	-0.021	0.002
ExpTr	-0.006	0.002	-0.008	0.003	-0.004	0.002	-0.005	0.002

TABLE 10
Comparison of Effects of TP1-TP3 across Different Propensity Score Weighting Methods

Propensity Score Weighting Methods	*TP1*	*TP2*	*TP3*
Unweighted	0.223 (0.015)	0.141 (0.006)	0.086 (0.005)
Weight propensity (treatment effect for treated)	0.256 (0.004)	0.154 (0.004)	0.09 (0.006)
Weight propensity (treatment effect for control)	0.208 (0.025)	0.136 (0.008)	0.078 (0.004)
Weight EOTM	0.255 (0.004)	0.152 (0.003)	0.082 (0.003)

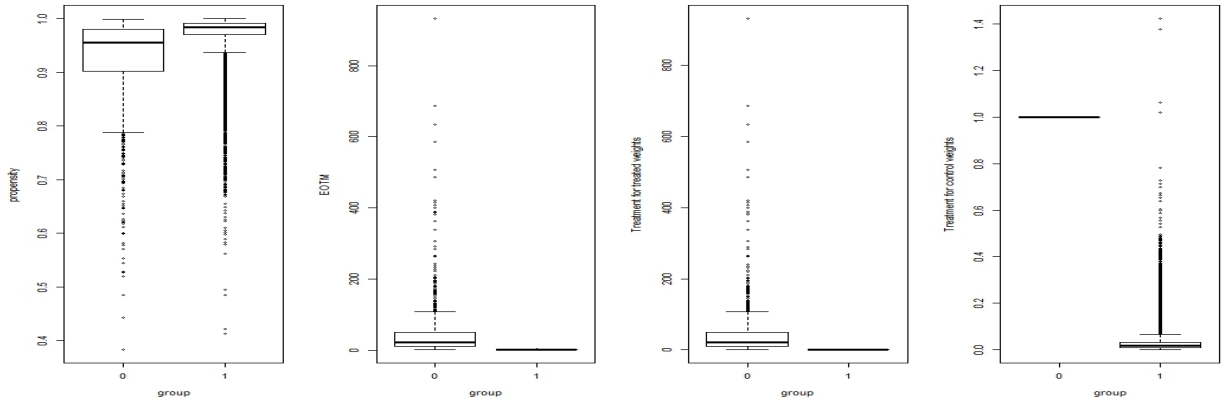

FIGURE 1 Covariate balance checking of four methods for treatment variable TP1.

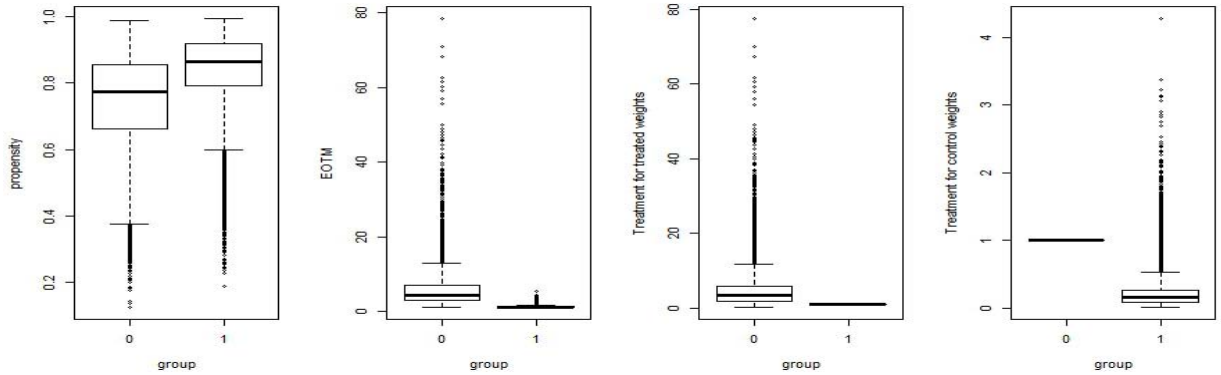

FIGURE 2 Covariate balance checking of four methods for treatment variable TP2.

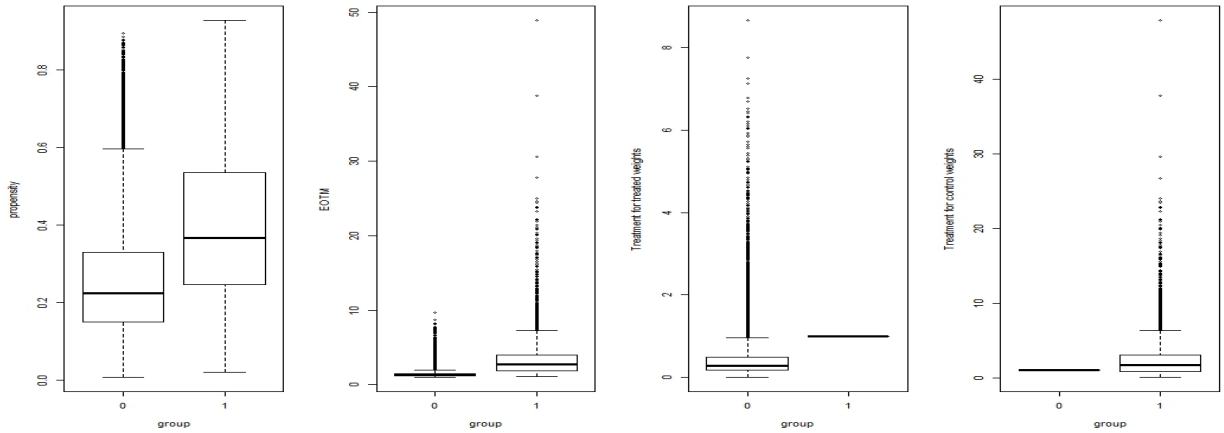

FIGURE 3 Covariate balance checking of four methods for treatment variable TP3.