**Title:**

# Are Gain Scores Really a “Treacherous Quicksand”? An Empirical Comparison Between Gain Score Methods and ANCOVA

**Authors and Affiliations:**

Yongnam Kim, University of Missouri—Columbia

## Background

Despite the heavy emphasis on the importance of pretest measures of the outcome (e.g., Cook & Steiner, 2010; Hallberg, Cook, Steiner, & Clark, 2018), a specific way of using the pretest has long been criticized in educational and psychological literature. While many researchers prefer to use the pretest as a control variable for ANCOVA, gain score methods that compute a gain score by subtracting the pretest from the posttest and use the gain score as an outcome variable have been widely avoided. Cronbach and Furby (1970) recommended researchers to "frame their questions in other ways [than gain score methods]" (p. 80), and Campbell and Erlebacher (1970) even claimed that "gain scores are in general such a *treacherous quicksand* [emphasis added]" (p. 197). Although subsequent studies have shown that, theoretically, under certain conditions, gain score methods can produce unbiased effect estimates while ANCOVA cannot (Allison, 1990; Maris, 1998), a strong negative view about gain score methods remains prevalent among researchers and practitioners (Smolkowski, 2019).

## Focus of Study

This article investigates how much bias can be eliminated by gain score methods and ANCOVA when the treatment is not randomized. Instead of presenting theoretical justifications (e.g., Kim & Steiner, 2019; Thomas & Zumbo, 2012), this article focuses on providing empirical evidence on bias-removing using the two methods in a real educational setting.

## Methods

To evaluate the bias-removing performance of different causal estimation methods in a real, not simulated, situation, a novel research design, which shall be referred to as a *part-whole design*, is proposed. The basic logic of the design is illustrated in Figure 1. The part-whole design rests on the fact that the functional relationships between parts and the whole are sometimes explicitly determined when defining the whole with the parts. For example, many academic or psychological test scores are the sum score of several individual item scores. From $S = X_1 + X_2 + \cdots + X_j$, where *S* denotes the sum score, *X* denotes the individual item score (1 = correct; 0 = incorrect), and *j* denotes item index, the functional relationship between $X_1$ and *S* is × 1. That is, holding all other things constant, if $X_1$ changes from 0 (incorrect) to 1 (correct), then *S* increases by 1. That means that the causal effect of the part ($X_1$) on the whole (*S*) is *one* (see Figure 1A). However, as is highlighted in Figure 1B, the total association between $X_1$ and *S* is not only due to the causal relation, colored in blue, but also non-causal relations, colored in red (see Pearl, 2000, for *d*-separation). The red-colored structure in Figure 1B can then be interpreted as a confounding structure with respect to the $X_1$-*S* relationship. Thus, in the part-whole design, researchers can have real confounding (i.e., real people make their decisions based on their real backgrounds in a real-world setting) while the true causal effect of the part on the whole is known.

## Findings

*Statewide math assessment data*. The common education dataset from the *Handbook of Quantitative Methods for Detecting Cheating on Tests* (Cizek & Wollack, 2017) was analyzed using the part-whole design. The data set includes fifth graders' (*n* = 69,806) item responses (correct/incorrect) on 53 math items. The results are presented in Figure 2. For the 1st item, the unadjusted estimate (U), obtained by regressing *S* on $X_1$, was 10.21. Since the true causal effect is 1.00, the difference between the unadjusted estimate and the true causal effect ($10.21 -$

$1.00 = 9.21$), is the initial confounding bias with respect to the $X_1$- $S$ relationship. To remove this bias, one can apply ANCOVA that regresses $S$ on $X_1$ and $P$, where $P$ denotes the pretest (the sum score of 58 math item scores from the prior year). The ANCOVA estimate (C) was 3.01 and thus controlling for the pretest decreases the bias but a substantial amount of bias (i.e., $3.01 - 1.00 = 2.01$) remains. Finally, the gain score estimate (G), obtained by regressing $G$ on $X_1$, where the gain score $G$ is computed by $S - P$, was .75. So, the bias in the gain score estimate is $.75 - 1.00 = -.25$. Thus, in the presence of real confounding between the first item score (part) and the sum score (whole), the use of gain score methods returned a least biased effect estimate. This relative superiority of gain score methods over ANCOVA in terms of bias-removing was replicated across all 53 items as illustrated in Figure 2.

*Tennessee STAR project data*. Figure 3 summarizes results that analyze the Tennessee STAR project data for eighth graders (*n* = 1,830) using the part-whole design. Unlike the previous math assessment data, the Tennessee STAR project data include many background variables (e.g., students' gender, race, etc.). As the open dataset does not include any individual items for composite scores, ten domain scores (e.g., science, social sciences, etc.) were used as parts and the whole consists of the sum score of all the domain scores. For the first domain, science score (sci), the unadjusted estimate (U) was 4.85, the ANCOVA estimate controlling for the pretest only (C colored in gray) was 1.92, and the ANCOVA estimate controlling for the pretest and 15 other baseline covariates (C colored in red) was 1.81. In contrast, the gain score estimate (G) was .80, which is closest to the true causal effect of 1.00. Again, this relative superiority of gain score methods over ANCOVA was replicated in all other domains except the spelling score (spe).

## Conclusion

Using the part-whole design that allows researchers to know the true causal effect in the presence of a real confounding bias, this article investigated how much bias can be eliminated using gain score methods and ANCOVA. Although gain score methods have long been criticized and discredited in educational and psychological research, the paper shows that gain score methods can outperform ANCOVA (even when other covariates are added in ANCOVA) and produce far less biased effect estimates in a real-world situation. Meeting necessary assumptions for gain score methods or ANCOVA is case-specific but the empirical evidence presented here shows that uninformed criticism on gain score methods should be discredited.

**References**

Allison, P. D. (1990). Change Scores as Dependent Variables in Regression Analysis. *Sociological Methodology 20*, 93-114.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education programs look harmful. In J. Hellmuth (Ed.), *The Disadvantaged Child: Vol. 3. Compensatory Education: A National Debate* (pp. 185-210). New York: Bruner/Mazel.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods, 15*(1), 56-68.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68-80.

Hallberg, K., Cook, T. D., Steiner, P. M., & Clark, M. H. (2016). pretest measures of the study outcome and the elimination of selection bias: Evidence from three within study comparisons. *Prevention Science, 15*, 1-10.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York, NY: Cambridge University Press.

Kim, Y., & Steiner, P. M. (2019). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research*. Advance online publication. doi:10.1177/0049124119826155

Maris, E. (1998). Covariance Adjustment versus Gain Scores—Revisited. *Psychological Methods 3*, 309-27.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Smolkowski, K. (2019, March 26). *Gain Score Analysis*. Retrieved August 23, 2019, from http://homes.ori.org/keiths/Tips/Stats_GainScores.html.

Thomas, D. R., & B. D. Zumbo. (2012). Difference Scores from the Point of View of Reliability and Repeated-measures ANOVA in Defense of Difference Scores for Data Analysis. *Educational and Psychological Measurement, 72*, 37-43.

Van Breukelen, G. J. (2006). ANCOVA versus Change from Baseline: More Power in Randomized Studies, More Bias in Nonrandomized Studies. *Journal of Clinical Epidemiology 59*, 920-25

Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness, 10*(1), 207-236.

FIGURE 1. *Logic of part-whole designs. Five item scores (1 = correct; 0 = incorrect), which are affected by some unmeasured common factors (e.g., ability, etc.), are summed and the total sum score is computed. The weight of each item score on the sum score is one (A). Viewing each item score (#1, #2, ..., #5) as a treatment and the sum score as an outcome, the causal effect of the item score (part) on the sum score (whole) is identical to the weight. For example, the causal effect of the fifth item correction #5 (treatment) on the sum score (outcome) must be one and the treatment-outcome relation is confounded by the structure in red (B). Trt = Treatment.*

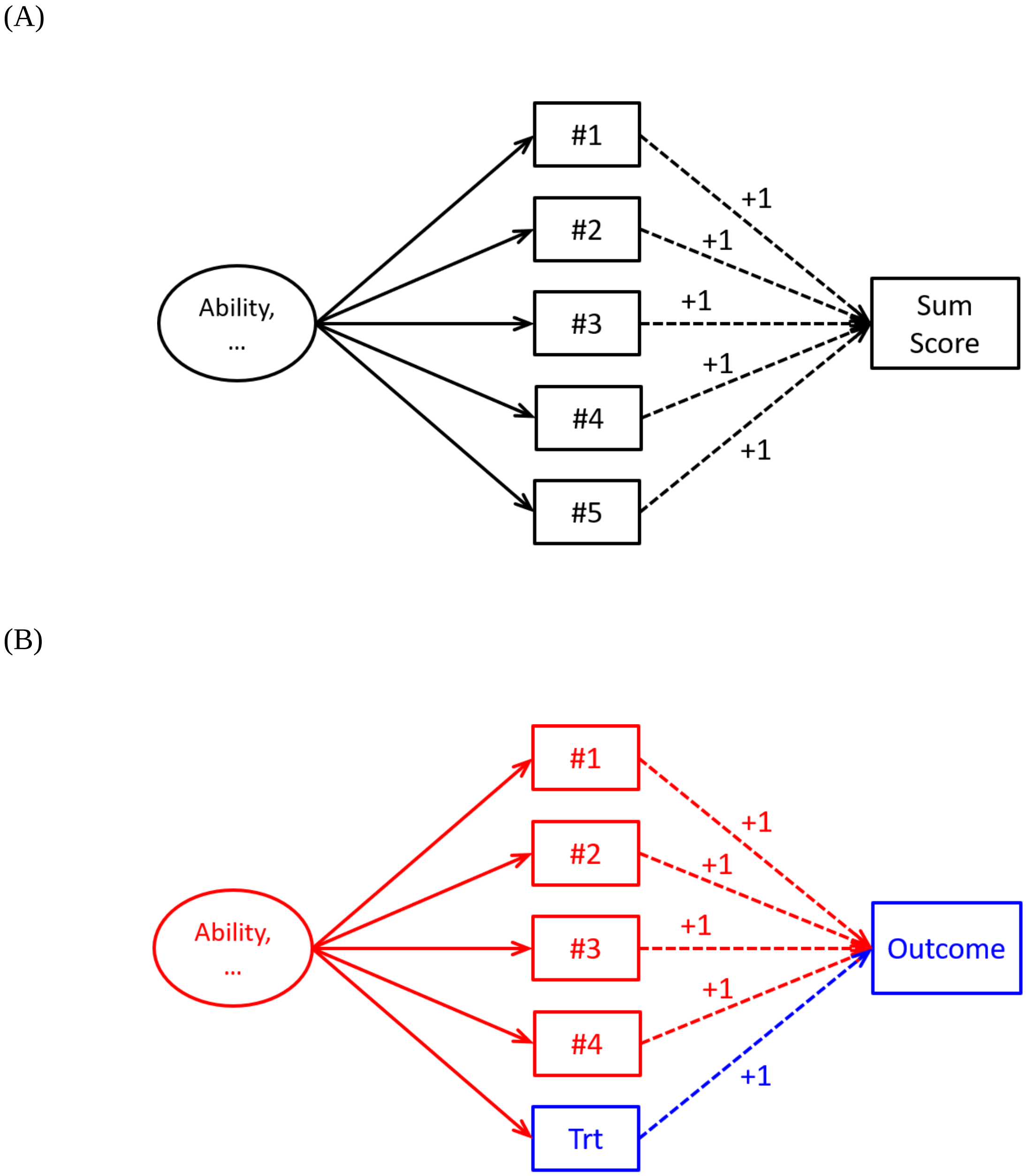


FIGURE 2. *Unadjusted (U), ANCOVA (C), and gain score (G) estimates across 53 math items from a statewide math assessment dataset from Cizek and Wollack (2017).*

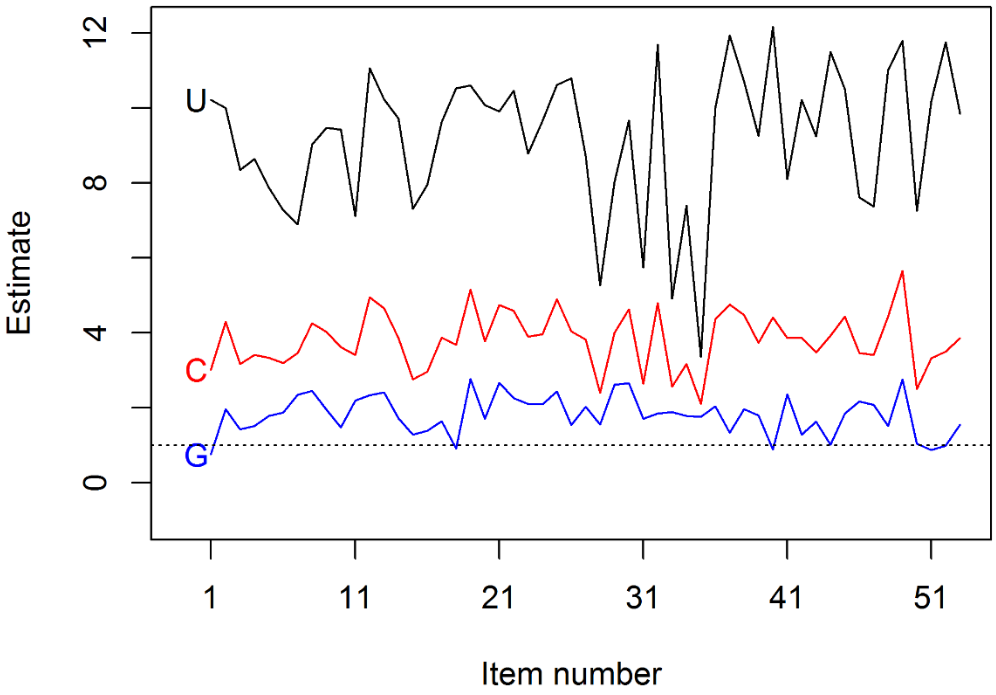
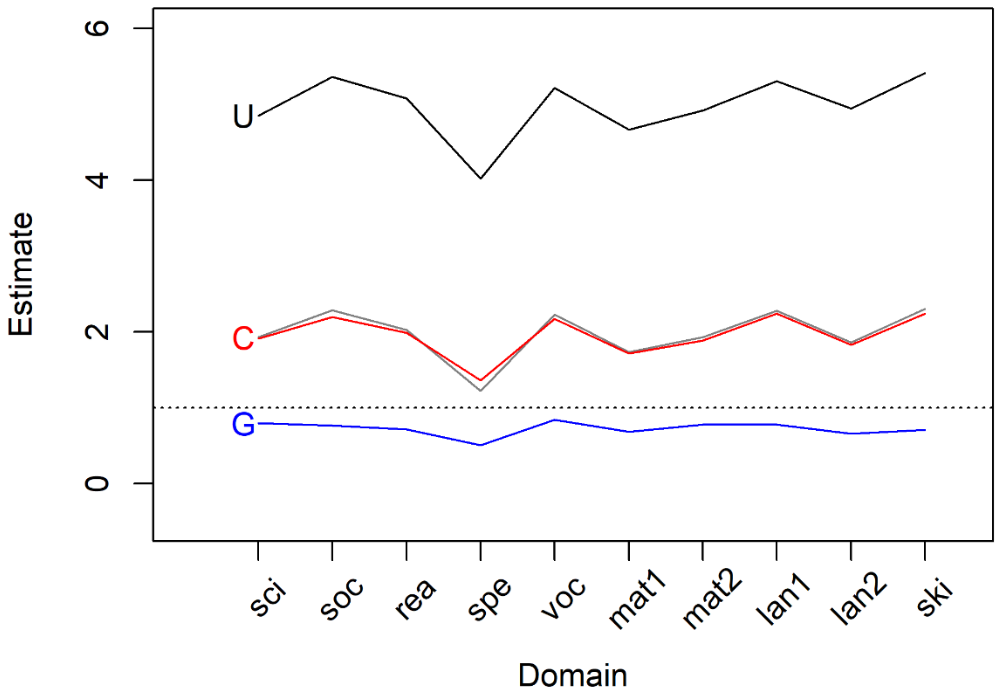


FIGURE 3. *Unadjusted (U), ANCOVA (C), and gain score (G) estimates across ten domains from the Tennessee STAR project dataset. For ANCOVA estimates, the gray line indicates the estimates controlling only for the pretest while the red line indicates the estimates controlling for the pretest and 15 other baseline covariates (e.g., gender, race). sci = science scale score; soc = social science scale score; rea = reading comprehension scale score; spe = spelling scale score; voc = vocabulary scale score; mat1 = math computation scale score; mat2 = math concepts and applications scale score; lan1 = language expression scale score; lan2 = language mechanics scale score; ski = study skills scale score.*


Estimate
U
C
G
sci
soc
rea
spe
voc
mat1
mat2
lan1
lan2
ski
Domain