Abstract Title Page

Title: Toward a Causally Informative Fit Index of Longitudinal Models: A Within-Study Design Approach

Authors and Affiliations:
Sirui Wan, University of California, Irvine
Daniela Alvarez-Vargas, University of California, Irvine
Drew Bailey, University of California, Irvine

Abstract

Background:

Structural equation modeling (SEM) is commonly used to relate theoretical assumptions to data and compare models that make substantively different predictions about the causal relations among these variables. Model fit indices are commonly used to evaluate whether a specific model is plausible or which of a set of alternative models is most plausible. However, it is well-known that a variety of models that make drastically different, mutually exclusive, theoretical predictions, can show similar or identical fit indices (Kline, 2011; Tomarken & Waller, 2003). The general problem that plausible competing theories can make the same set of predictions, referred to as "underdetermination" of scientific theory in philosophy of science (Stanford, 2017), can in some cases be addressed by generating stronger, more specific predictions on which theories can be compared (Mayo, 2018; Meehl, 1990).

Purpose:

We propose incorporating causal predictions to evaluate the fit of developmental models. We adapt the within-study design approach to generate long-run forecasts of intervention effects, conditional on the short-run effects and a longitudinal model fit to the control group relating end-of-treatment skills to later skills, and we derive an index of causal fit for these models fit to non-experimental data. We investigate whether some models are found to be consistently causally informative across the datasets. In addition, we test the usefulness and validity of the index of causal fit by asking the following questions:

1) Do traditional model fit indices and causally informative model fit indices yield similar patterns of results across conceptual replications?

2) Are discrepancies between traditional model fit indices and causally informative indices consistent across conceptual replications?

Data:

We analyze use data from three randomized control trials of mathematics interventions for children in classrooms spanning preschool through fifth grade: the Technology-enhanced, Research-based, Instruction, Assessment, and professional Development study (TRIAD; Clements & Sarama, 2013), an evaluation of a Number Knowledge Tutoring program (NKT; Fuchs et al., 2013), and an evaluation of a preschool mathematics curriculum called Pre-K Mathematics (PKM; Starkey & Klein, 2012). These were chosen because they share several design features (shown in Table 1) necessary to make the models we intended to run identified, and compare results. Details about the designs, participants, and methods are shown in Table 2.

Analyses:

Figure 1 illustrates the within-study design which contains four steps. First, we estimated the causal effect of each treatment on the same outcome measure at the posttest (the parameter $a_{experimental}$ in Figure 1) and all subsequent waves ($c_{experimental}$) by regressing the outcome measure on randomly assigned treatment status and all demographics and pretest covariates listed in Table 2. Second, we estimated the effects of earlier skills on later skills ($b_{non\text{-}experimental}$) using a variety of models (see Table 3) within the control group of each study. This parameter estimate can be interpreted as the estimated effect of a 1-unit boost to skill at time 1 (posttest at the end of treatment) to the same skill at time t (any subsequent wave following the end of treatment). Third, we projected the future impacts ($c_{non\text{-}experimental}$) by multiplying the end-of-treatment impact

($a_{experimental}$) by the estimates of $b_{nonexperimental}$ yielded by each model, such that $c_{\text{non-experimental}} = a_{\text{experimental}} * b_{\text{non-experimental}}$. That is, the estimate of the effect of an intervention on a later outcome equals to the product of the effect of an intervention on early outcome and the estimate of the effect of the early outcome on the later outcome. Finally, for each model used to generate estimates of $b_{nonexperimental}$, we computed an index of causal fit: the Causal Mean Squared Error (CMSE), based on the absolute deviations between the projected effects ($c_{non\text{-}experimental}$) and the observed causal effects ($c_{experimental}$). We calculated the mean of the squared differences between the observed impacts and the projected estimate at each wave, as shown in the formula: CMSE = $\sum(c_{experimental} - c_{non\text{-}experimental})^2$/Number of total waves.

Then we compared each model's CMSE to its performance on traditional fit indices based on the variance covariance matrix. In all we estimated $b_{non\text{-}experimental}$ using the following four models within each dataset: regression, AR (1), AR (2), RIAR (see Table 3).

Results:

Figure 2 presents the observed impacts, $c_{experimental}$, and projected impacts, $c_{non\text{-}experimental}$, on posttest and later math achievement in each dataset. The trajectories of the actual treatment impacts over time are displayed by the red lines in Figure 2. The projected impacts of different models are displayed using different colors.

There are several regularities in the lines plotted in Figure 2. The correlation-based estimates of the treatment effect by the models show consistent patterns of bias across datasets. Some models consistently outperform other models at forecasting later experimental impacts. For example, the RIAR model consistently outperform the AR(1) and AR(2) models. The regression, AR(1), AR(2) models consistently overestimate the long-run treatment impacts.

Table 4 summarizes the traditional SEM model fit indices and CMSE. The traditional fit indices and the index of causal fit are not perfectly related. Some of the models within the same dataset demonstrated inconsistent patterns with better (lower) CMSE but worse model fit indices as measured by (Kline, 2011).

The inconsistent pattern of statistical and causal fit within each dataset is consistent across all three datasets**.** Figure 3 highlights the similar pattern of discrepancy between the CMSE and RMSEA of the models for each dataset. The AR(2) models consistently show a very good RMSEA —often approaching zero— but they have the highest CMSE and so are bad at predicting the observe effect in all three datasets. The RIAR models have good model fit and lower CMSE scores than the AR(1) and AR(2) models. Therefore, they are better predictors of causal impacts among the models.

Conclusions:

Theoretical underdetermination is a critical problem for psychologists who work with observational data. Here we propose another tool for further addressing the underdetermination problem by incorporating causal benchmarks into the model selection process. Supporting the validity and potential usefulness of this approach, we present evidence that this approach can identify models that consistently perform better than others for a narrowly defined problem, and that our index of causal fit contains information not supplied by a commonly used index of statistical fit. We hope others will attempt to create and refine indices of causal model fit and apply this general approach to other important questions within psychology.

Appendices

Appendix A. References

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist, 73,* 81-94.

Clements, D. H., Sarama, J., Khasanova, E., & Van Dine, D. W. (2012). TEAM 3-5—Tools for elementary assessment in mathematics. *Denver, CO: University of Denver*.

Clements, D. H., Sarama, J., Layzer, C., Unlu, F., Wolfe, C. B., Spitler, M. E., & Weiss, D. (2016, March). *Effects of TRIAD on Mathematics Achievement: Long-Term Impacts.* Paper presented at the Spring 2016 Society for Research on Educational Effectiveness Conference, Washington, D.C. Abstract retrieved from https://www.sree.org/conferences/2016s/program/downloads/abstracts/1726.pdf

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., ... & Bryant, J. D. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *Journal of Educational Psychology*, *105*, 58-77.

Kline, R.B. (2011) *Principles and Practice of Structural Equation Modeling*. Guilford Press, New York

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244.

Stanford, K. (2017). Underdetermination of scientific theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of philosophy* (Winter 2017 Edition). Stanford. Retrieved from http://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/

Starkey, P., & Klein, A. (2012). Scaling up the implementation of a pre-kindergarten mathematics intervention in public preschool programs (Final Report: IES Grant R305K050004). Washington, DC: U.S. Department of Education.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with" well fitting" models. *Journal of Abnormal Psychology*, 112(4), 578.

Appendix B. Tables and Figures

Table 1

Design features shared across datasets and explanations

Design feature	Purpose
Randomization to groups	Strongly identified end-of-treatment impacts
Intervention is specifically designed to increase end of treatment math skills	Decreases potential biases from indirect effects of the treatment via unmeasured mediators
At least 3 waves of data on key outcome measure (including pretest)	Allows for the estimation of longitudinal models with latent intercepts
a) Participants in preschool or the early school years in the U.S. b) Math achievement measure at each wave c) Approximately 1-year time lags between waves	Reasonably comparable parameters being estimated across studies in longitudinal models
Non-trivial end-of-treatment impact	Analyses require multiplying estimated effects by end of treatment impact; null impacts will not allow to clearly differentiate among models
Pretest covariates	To statistically control for potential confounds influencing earlier and later math achievement

Table 2

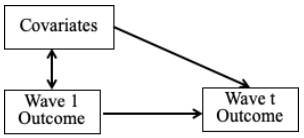
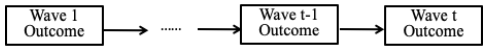
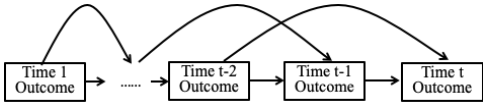
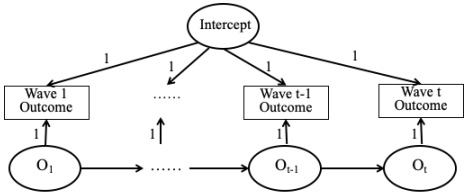
Datasets used in the study

	TRIAD	NKT	PKM
Key citation	Clements et al., (2013)	Fuchs et al. (2013)	Starkey & Klein (2012)
Level of randomization	Preschools (blocked)	Within classrooms	Preschool sites (blocked)
Sample size	42 schools, 1,375 students	591 students	63 preschool sites, 744 children
Race/Ethnicity, FRPL status	51% AA, 23% H, 85% FRPL	69% AA, 20% White, 84% FRPL	52% White, 18% H, 17% AA, 100% FRPL
Treatment group used in the study	Building Blocks-No Follow Through	Number Knowledge Tutoring (NKT) - Speeded Practice	Pre-K Mathematics
Waves (attrition rate from pretest)	F PK (pretest), S PK (posttest), S K, S G1, F G4, S G4, F G5	G1 (pretest and posttest), G2, G3	F PK (pretest), S PK (posttest), S K, S G1
Covariates			
Demographics	gender, ethnicity, age, FRPL, ELL, SE, mother's education	gender, ethnicity, FRPL	gender, ethnicity, ELL, Head Start Program
Pretests	REMA	FCR, KeyMath-Numeration, WRAT-Reading, Nonverbal Reasoning, Processing Speed, WM-Listening Recall, WM-Counting Recall, Listening Comprehension, Attentive Behavior	TEMA, WJ Letter-Word Identification, Spelling, and Understanding Directions subtests
Mathematics achievement outcomes measures	REMA (PK-1), TEAM 3-5 (G4-5)	FCR	TEMA (pre-k pre, post, K, grade 1)
End of Treatment Impacts	REMA (.56 SD)	FCR (.40 SD)	TEMA (.33 SD)

Notes. FRPL = Free or Reduced-Price Lunch eligibility, ELL = English Language Learner status, SE = Special Education status, AA=African American, H=Hispanic, F=Fall, W=Winter, S=Spring, G=Grade, PK=Pre-k, WJ = Woodcock Johnson, WM = working memory, FCR = Facts correctly retrieved, TEMA = Test of early mathematics ability , REMA = Research based early mathematics assessment.

Table 3

SEM models used in the study

Model	Picture	Advantages	Disadvantages
Regression		• Flexible, does not impose functional form on pattern of impacts across time	• Limited to linear relations • Possible omitted variable bias due to unmeasured confounds and measurement error in covariates
AR(1) model		• Only requires 2 waves of data • Does not require covariates	• Without covariates, time- and domain-general omitted variable bias are major concerns • Implies exponential decay of correlations across time, which is unrealistic
AR(2) model		• Does not require covariates • Accounts for stability in correlations across waves	• Requires at least 3 waves of data • Without covariates, time- and domain-general omitted variable bias are major concerns
RIAR model		• Does not require covariates • Accounts for stability in correlations across waves • Models time-general confounds	• Requires at least 3 waves of data

Note. AR = Autoregressive, RIAR = Random Intercept Autoregressive, in each figure, *c* indicates coefficients constrained to be equivalent within a model.

Table 4

Comparing traditional model fit indices and causal model indices

Dataset	Model	CFI	TLI	RMSEA	SRMR	CMSE overall	CMSE 1 year later	CMSE other waves
TRIAD	AR(1) model	0.971	0.912	0.210	0.034	0.045	0.026	0.050
	AR(2) model	0.999	0.998	0.019	0.012	0.059	0.009	0.072
	RIAR model	0.968	0.952	0.105	0.085	0.003	0.007	0.002
NKT	AR(1) model	0.997	0.993	0.027	0.029	0.006	0.001	0.010
	AR(2) model	1.000	1.006	0.000	0.010	0.007	0.000	0.014
	RIAR model	1.000	1.019	0.000	0.006	0.003	0.000	0.006
PKM	AR(1) model	0.968	0.936	0.161	0.047	0.032	0.035	0.029
	AR(2) model	1.000	1.004	0.000	0.000	0.023	0.018	0.029
	RIAR model	0.989	0.968	0.110	0.031	0.006	0.009	0.003

Notes. AR = Autoregressive, RIAR = Random Intercept Autoregressive, CMSE = causal mean squared error.

CMSE = $\frac{\sum(c_{experimental} - c_{non-experimental})^2}{Number\ total\ waves}$, RMSEA = $\sqrt{\{(x_t^2 - df_t)/(N-1)\}/(df_t/g)}$, where x_t^2 = model chi-square, N = sample size, and g= number of groups.

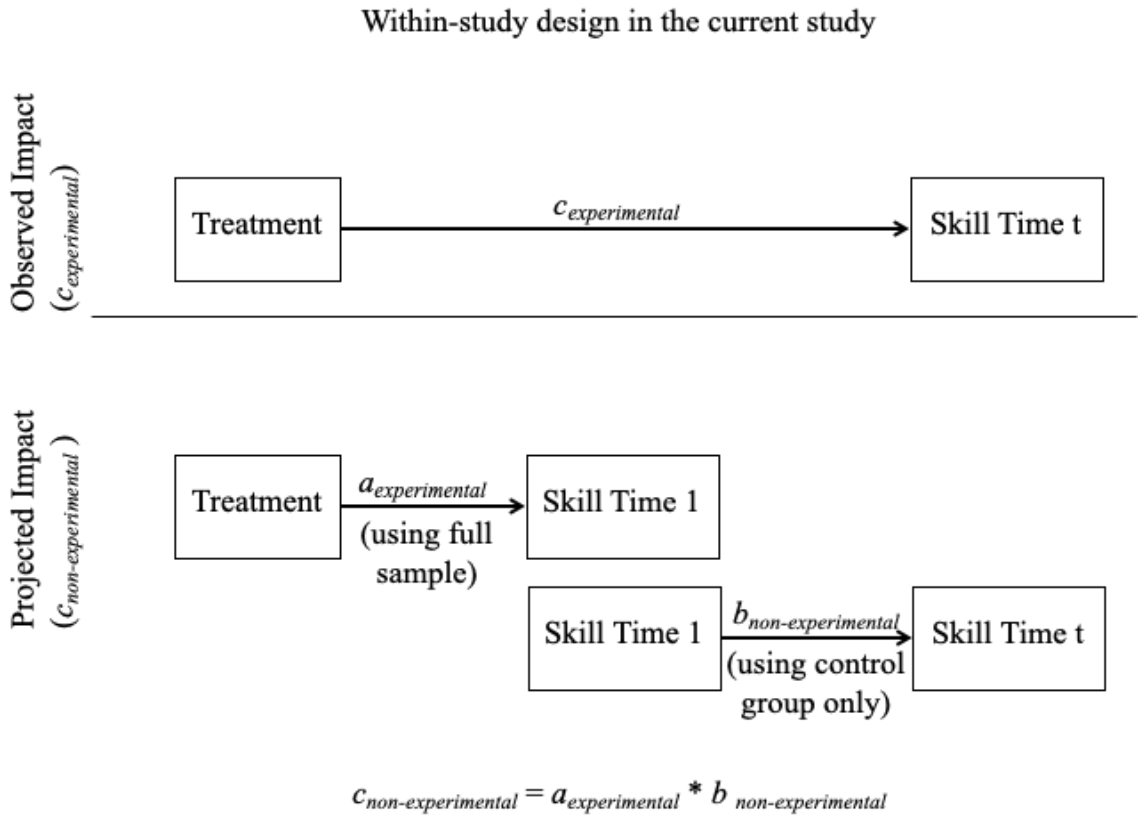


Figure 1. The within-study design used in the current study. The top panel presents $c_{experimental}$ which is the casual effect of treatment on skill time *t* drawn from an experiment with random assignment to a treatment or control group. The effects of the top panel are compared to $c_{non-experimental}$ in the bottom panel. The bottom panel demonstrates how we calculated the projected impact $c_{non-experimental.}$ The projected impact is the product of $a_{experimental}$ (the observed effect of the treatment on skill time 1), and $b_{non-experimental}$ (the estimated causal effect of skill at time 1 on skill at time *t* drawn from the control group). Therefore, $c_{non-experimental} = a_{experimental} * b_{non-experimental.}$

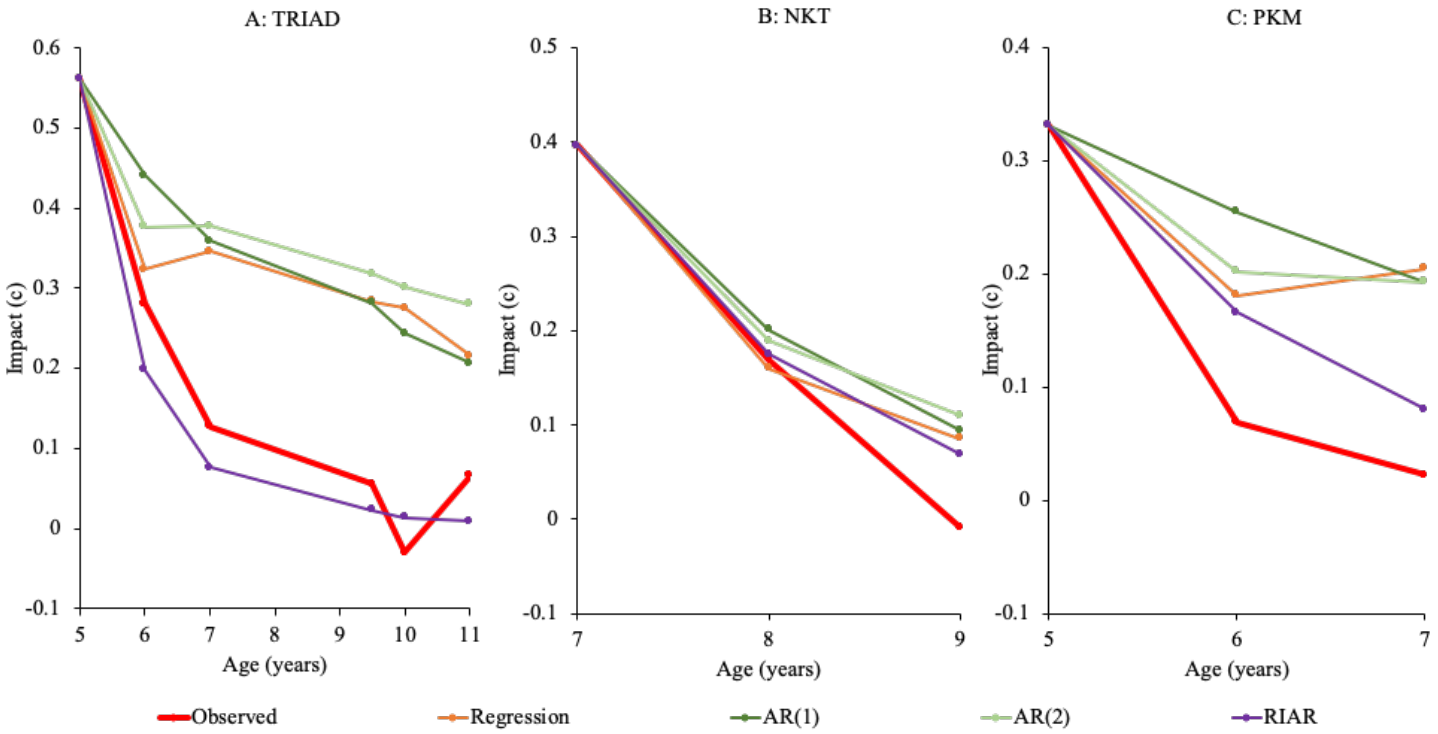


Figure 2. Observed impact ($c_{experimental}$) and projected impact ($c_{non\text{-}experimental}$) on later math achievement in each dataset. Note: The line Observed is the unbiased observed treatment effect ($c_{experimental}$). The lines Regression, AR(1), AR(2), and RIAR are the projected impacts $c_{non\text{-}experimental}$ which is the product of the observed effect of the treatment on end-of-treatment math achievement ($a_{experimental}$) and the estimated effects ($b_{non\text{-}experimental}$) of end-of-treatment math achievement on later achievement. AR = Autoregressive, RIAR = Random Intercept Autoregressive.

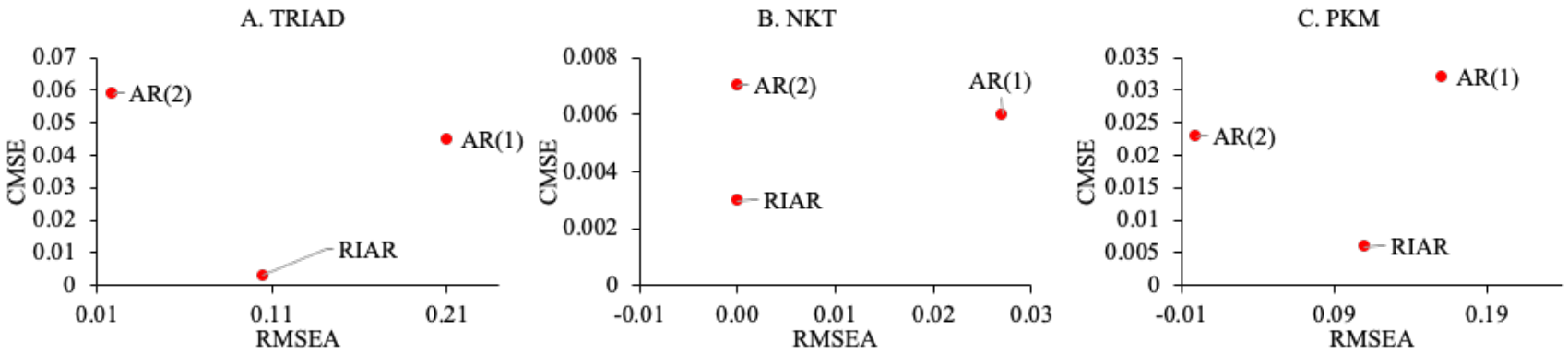


Figure 3. RMSEA and CMSE of models for each dataset. Note: AR = Autoregressive, RIAR = Random Intercept Autoregressive.