Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Item-Level Heterogeneous Treatment Effects in Instrumental Variables Regression

Sanford R. Student,[*†] Joshua B. Gilbert,[‡] Jesse U. Eze,[†] William S. Young,[‡] and Benjamin W. Domingue[¶]

†School of Education, University of Delaware, Newark, Delaware, United States
‡Graduate School of Education, Harvard University, Cambridge, Massachusetts, United States
¶Graduate School of Education, Stanford University, Palo Alto, California, United States
*Corresponding author. Email: srstu@udel.edu

Abstract

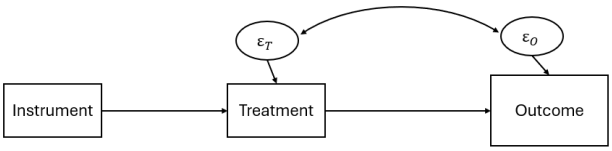
Recent research has highlighted that modeling heterogeneous treatment effects across the individual items on an outcome measure can yield additional insights into interventions' effects beyond what is available from comparisons of average scores. So far, this work has generally focused on randomized controlled trials. However, in social science research, randomization often does not hold perfectly, or may be impossible or impractical. This can lead to treatment endogeneity—correlation between the predictor and the error term of the outcome. Instrumental variables regression (IVR) is a popular method for correcting the resulting bias by introducing an exogenous variable that drives change in the predictor but is otherwise uncorrelated with the outcome. Though typically estimated via two-stage least squares, IVR can also be estimated in a latent variable framework. We develop an explanatory item response modeling approach for estimating the average of treatment effects that are heterogeneous across the items used to measure a latent outcome in an IVR, supporting valid causal inferences about impacts of nonrandom treatments/dosages on test items. We illustrate the use of the approach via an empirical example analyzing item-level variability in sensitivity to time spent on homework using data from a large-scale international assessment of advanced high school mathematics.

Keywords: heterogeneous treatment effects, instrumental variables, causal inference, psychometrics, latent variable modeling

Instrumental variables regression (IVR) is an approach to reducing bias in the estimate of a treatment effect when the treatment is endogenous—that is, not randomly assigned (see Angrist and Krueger, 2001; Murnane and Willett, 2010). Most treatments of IVR use the econometric two-stage least squares (2SLS) approach to estimation. However, IVR can also be represented within a structural equation modeling (SEM) framework by allowing for correlated error terms between the endogenous treatment variable and the outcome, effectively modeling it as a type of full mediation (Maydeu-Olivares et al., 2019).

Figure 1 shows a path diagram illustrating this IVR-as-SEM approach with one instrument. When the model assumptions (i.e., relevance, exogeneity, the exclusion restriction, and monotonicity) are met, IVR yields an unbiased estimate of the treatment effect and provides accurate standard errors, equivalent to those generated in 2SLS using a robust sandwich estimator (Maydeu-Olivares et al., 2019, p. 880).

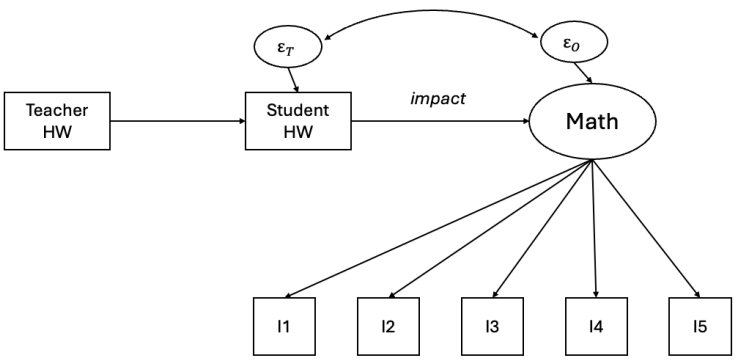
Although Maydeu-Olivares et al. (2019) limited their analysis to observed variables, one of the strengths of SEM is the ability to model latent variables and estimate relations between them that are corrected for measurement error. This involves embedding a confirmatory factor analysis (CFA) model within a larger structural model, where the CFA defines the latent variable as an unobserved



Notes: ε_0 is the error term for the outcome; ε_T is the error term for the endogenous treatment.

Figure 1. Path diagram for instrumental variables regression conceived as a structural equation model

common cause of observed variables known as its indicators. Figure 2 extends the SEM treatment of IVR to the case of a latent outcome variable. To make the example more concrete and preview the focal analysis of this paper, we have replaced "instrument," "treatment," and "outcome" with actual variables from an analysis of the influence of time spent on homework on math achievement that we expand upon below (Gustafsson, 2013). This model aims to estimate the effect of time spent on homework on math achievement. However, the original study was motivated by a concern about "reverse causation": time spent on homework may positively influence math achievement, while poor math achievement may also drive more time spent on homework. The presence of the instrument, teacher-assigned homework time, enables the separate estimation of the relation between exogenous homework time and math achievement in the presence of reverse causation if the model assumptions are met (Bollen, 2012). Furthermore, modeling math achievement as a latent variable corrects for attenuation bias in the main effect induced by outcome measurement error. Without correction, measurement error would affect standardization of the model coefficients and reduce power to detect statistically significant relations (Shear & Briggs, 2024).



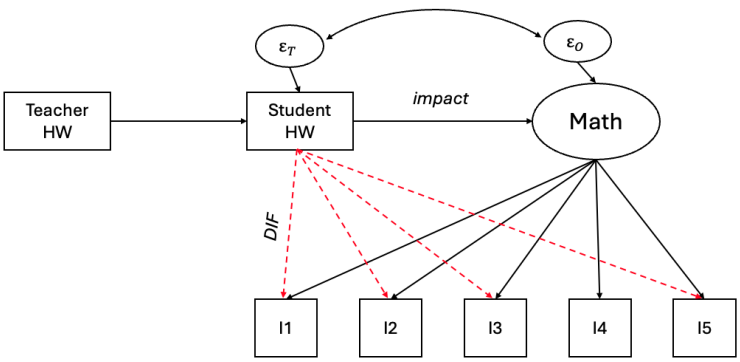
Notes: ε_0 is the error term for the outcome; ε_T is the error term for the endogenous treatment. I1–I5 are observed indicators of latent math achievement. Squares represent observed variables, circles represent latent variables.

Figure 2. Path diagram for instrumental variables regression conceived as a structural equation model with a latent outcome

IVR relies on several assumptions, most notably, that the instrument affects the outcome only through its influence on the treatment (i.e., the exclusion restriction), and that the instrument is strong (i.e., related to the treatment; the "relevance" assumption) and applied to a sufficiently large sample (Winship & Morgan, 1999). When the outcome is modeled as a latent variable, an additional

assumption comes into play: that the measurement model, specifically, the confirmatory factor analysis (CFA) portion, is invariant across levels of the endogenous treatment. In the context of Figure 2, this means that student homework time should affect performance on individual items (I1–I5) *only* via its influence on the underlying latent variable *Math*. In other words, the relationship between homework time and item responses must be fully mediated by the latent variable (Stoetzer et al., 2025). Item-specific effects of the treatment above and beyond this main effect are examples of differential item functioning, or DIF (Angoff, 1981; Angoff & Ford, 1973), with respect to the treatment variable.

In the present context, treatment-induced DIF implies that test items vary in their sensitivity to time spent on homework, beyond the main mediated effect through the latent variable of math achievement. This situation is illustrated in Figure 3. When DIF is considered with respect to a treatment rather than demographic background variables, it is not an issue of fairness but rather an issue of accurately characterizing treatment effects and uncertainty around them (Gilbert et al., 2023). That is, when item-specific effects exist and are ignored, standard errors of the main effect are underestimated (Gilbert et al., 2023). In line with this view, this paper introduces a modeling approach for handling item-to-item variation in sensitivity to treatment within an IVR model with a continuous treatment variable.



Notes: ε_0 is the error term for the outcome; ε_T is the error term for the endogenous treatment. I1-I5 are observed indicators of latent math achievement.

Figure 3. Path diagram for instrumental variables regression conceived as a structural equation model with potential DIF

1. Background

We approach the general problem of treatment-based DIF in IVR with latent variables via the framework of Item-Level Heterogeneous Treatment Effects (IL-HTE; Gilbert, Himmelsbach, et al., 2025; Gilbert et al., 2023) using the Explanatory Item Response Model (EIRM; Wilson and De Boeck, 2004). Throughout this paper, we assume that all indicators of latent variables are binary, such as items scored as correct or incorrect and coded as 1 or 0.

1.1 Item-Level Heterogeneous Treatment Effects

IL-HTE frames DIF as a phenomenon in which items' sensitivity to a given treatment varies randomly and symmetrically about an overall treatment effect. The basic unconditional IL-HTE model for a

randomly assigned treatment, estimated as an EIRM, is given as:

$$\log\left(\frac{P(X_{ij})}{1-P(X_{ij})}\right) = \beta_0 + \beta_1 \text{treat}_i + \theta_i + \zeta_{0j} + \zeta_{1j}\text{treat}_i \tag{1}$$

$$\theta_i \sim N(0, \sigma^2_\theta), \tag{2}$$

$$\begin{bmatrix}\zeta_{0j}\\ \zeta_{1j}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2_{\zeta_0} & \rho_{10}\\ \rho_{10} & \sigma^2_{\zeta_1}\end{bmatrix}\right), \tag{3}$$

where

- X_{ij} is the observed response of person i to dichotomous item j, taking on values of either 0 or 1.
- β_0 is the logit of a correct response, from a respondent of mean ability at a value of 0 on the treatment, to an item of average difficulty.
- β_1 is the extent to which average item easiness changes with the treatment, equivalently framed as the average treatment effect on θ_i (these two interpretations are statistically equivalent in single-timepoint, single-population analyses; San Martín, 2016).
- θ_i is a random effect representing person ability not explained by the treatment variable treat_i, which has an expectation of 0 and (residual) variance σ^2_θ.
- ζ_{0j} is a random intercept representing variance in item easiness (i.e., the inverse of the item difficulty captured by the typically-parameterized Rasch model), with variance $\sigma^2_{\zeta_0}$.
- ζ_{1j} is a random slope representing item-specific deviations in their sensitivity to the treatment, relative to the average treatment effect β_1, with variance $\sigma^2_{\zeta_1}$.
- ζ_{0j} and ζ_{1j} freely covary with covariance ρ_{10}.

Equations 1-3 constitute an extended version of a Rasch model with random effects for items (De Boeck, 2008). That is, every item's slope/discrimination—the rate at which the log-odds of a correct item response changes as θ_i changes—is fixed to 1 (though this assumption can be relaxed in a 2PL EIRM, see Gilbert, Zhang, et al., 2025; Petscher et al., 2020). Building on this basic model, additional terms allow for variation in item easiness both at the baseline level of treatment (e.g., control) and in how easiness shifts in response to different treatment levels. This approach assumes that items differ in both their baseline difficulty and in their responsiveness to treatment, and that these variations follow a multivariate normal distribution. Note as well that the link function used here is the logit, connecting the linear predictor to binary outcomes, though a probit link could also be used. We return to this distinction below.

2. Application to Instrumental Variables Regression

We extend the IL-HTE model from randomized controlled trials to instrumental variables regression, which entails modeling the influence of the instrument on the treatment and the residual covariance between treatment and latent outcome.

We implement IL-HTE in the context of IVR using Mplus's general latent variable modeling framework (Asparouhov & Muthén, 2016; B. Muthén & Muthén, 2017; B. O. Muthén, 2002) and Markov Chain Monte Carlo estimation. Specifically, one can extend the basic IL-HTE model for an

RCT as follows:

$$\text{treat}_i = \eta_0 + \boldsymbol{\eta}'\boldsymbol{inst_i} + \varepsilon_i \tag{4}$$

$$\Phi(X_{ij}) = \beta_0 + \beta_1\text{treat}_i + \theta_i + \zeta_{0j} + \zeta_{1j}\text{treat}_i \tag{5}$$

$$\begin{bmatrix}\theta_i \\ \varepsilon_i\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\sigma^2_\theta & \sigma_{10} \\ \sigma_{10} & \sigma^2_\varepsilon\end{bmatrix}\right) \tag{6}$$

$$\begin{bmatrix}\zeta_{0j} \\ \zeta_{1j}\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\sigma^2_{\zeta_0} & \rho_{10} \\ \rho_{10} & \sigma^2_{\zeta_1}\end{bmatrix}\right) \tag{7}$$

This extension incorporates prediction of treat_i by one or more exogenous instruments via the vector of regression slopes $\boldsymbol{\eta}'$ and the vector of instruments **inst**. Each instrument has its own slope, such that, e.g., η_1 would refer to the slope for instrument inst_1. The residual of this regression, ε_i, is assumed to have an expectation of 0 and to be multivariate normally distributed with θ_i with covariance σ_{10}. In terms of Figure 3, this model estimates all possible red lines capturing DIF via a distribution of random effects, and then operationalizes *impact* as the mean of those random effects. The use of the probit link (i.e., expressing the function for a correct response via Φ, the normal cumulative density) is a requirement of Mplus's Bayesian estimation framework, but does not fundamentally change the relations expressed by the model compared to the logit link; it rescales them by a relation that is almost, but not exactly, a linear constant (Camilli, 2017; Savalei, 2006). We note as well that we take a "computational frequentist" perspective on Bayesian estimation in this case, meaning that Bayesian estimation is solely employed as means to produce parameter estimates for a model that cannot easily be estimated using maximum likelihood (Levy & McNeish, 2023). We thus will estimate the model below using Mplus's default, diffuse priors (Asparouhov & Muthén, 2021a, 2021b). Opportunities for a more truly Bayesian perspective on this model, including the use of informative priors, are left as a future direction (see also Gilbert, Domingue, and Kim, 2025; Gilbert, Himmelsbach, et al., 2025 for more detail on Bayesian perspectives in this context).

Finally, we note that IVR does not support valid inference on the proportion of variance in the outcome explained or the magnitude of the residual covariance between outcome and treatment; it only allows for valid inference on the regression coefficient(s) of interest (Maydeu-Olivares et al., 2019). Thus, we set the scale for the latent outcome using the Rasch specification of fixing all item slopes to 1, as the alternative identification strategy of fixing the residual variance of θ to 1 would involve fixing the value of a meaningless parameter.

3. Application Demonstration

To demonstrate the application of IL-HTE to IVR with a latent outcome variable, we revisit an IVR analysis of the effect of time spent on homework on math achievement (Gustafsson, 2013) cited in a recent review of causal inference papers using large-scale international assessment data (Cordero et al., 2018). This is intended solely as a proof of concept.

3.1 The original study

The original study (Gustafsson, 2013) investigated the effect of time spent on homework on academic achievement, measured by performance on the Trends in International Math and Science Study (TIMSS) math test, in several countries. The analysis considered the 2003, 2007, and 2008 TIMSS waves, and used three methods in analyzing the data: multilevel regression nesting students within classrooms, IVR using teacher-reported homework assigned as an instrument for time spent on homework, and difference-in-differences. For the present paper, the most relevant approach is IVR. Our empirical example revisits this general design but uses the models described above to investigate the effect of homework time on math achievement modeled as a latent variable subject to

potential DIF. The original analysis was a set of observed-variable IVRs, run separately within each country using sampling weights, to estimate the impact of homework time on math performance with teacher-assigned homework time as an instrument.

3.2 Our dataset

We consider the same phenomenon—the effect of time spent on homework on math achievement. Our analysis focuses on the 2008 TIMSS Advanced dataset, specifically the mathematics test items, and the student population in the Philippines. The dataset consists of item responses from 3972 students nested in 116 classrooms. Students answered randomly-assigned subsets of a total of 71 math items, all scored dichotomously after collapsing categories for a small number of polytomous items. For our instrument, we used the teacher-assigned homework variable thw_i (originally named MT2MHWKM in the dataset), which was coded as 1 = 30 minutes or less, 2 = 31-60 minutes, 3 = 61-90 minutes, 4 = >90 minutes. For the endogenous cause variable, we transformed student-reported time spent on homework (originally named MS2MHTIM and reported in minutes spent per week) to a standard normal variable using the *bestNormalize* R package (Peterson, 2021). Note that we did not employ sampling weights or account for classroom clustering for this proof-of-concept analysis.

3.3 Model specification and estimation

We estimate an IL-HTE IVR model via equations 8-11, which employ a single instrument. As noted above, this model is estimated using Mplus' "Bayes" estimator, which implements a Markov Chain Monte Carlo approach to estimating the parameters' distributions (Asparouhov & Muthén, 2021a, 2021b). A maximum-likelihood solution may be possible as well but was found to be unacceptably slow in computation time during initial testing. We estimate the model using five chains and a convergence criterion of a maximum potential scale reduction factor, or $\widehat{R}$ (Gelman & Rubin, 1992), of 1.002. We thin 99 of every 100 draws (i.e., remove the draws from the construction of the posteriors) to reduce autocorrelation, which initial testing indicated was substantial without thinning, and specify a minimum of 10,000 draws per chain post-thinning. The full model specification is:

$$\text{shw}_i = \eta_0 + \eta_1 \,\text{thw}_i + \varepsilon_i \tag{8}$$

$$\Phi(X_{ij}) = \beta_0 + \beta_1 \,\text{shw}_i + \theta_i + \zeta_{0j} + \zeta_{1j} \,\text{shw}_i \tag{9}$$

$$\begin{bmatrix} \theta_i \\ \varepsilon_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_\theta & \sigma_{10} \\ \sigma_{10} & \sigma^2_\varepsilon \end{bmatrix} \right) \tag{10}$$

$$\begin{bmatrix} \zeta_{0j} \\ \zeta_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\zeta_0} & \rho_{10} \\ \rho_{10} & \sigma^2_{\zeta_1} \end{bmatrix} \right) \tag{11}$$

where shw_i is normalized student homework time, recoded from weekly time spent in minutes, and thw_i is untransformed teacher-assigned homework time.

3.4 Comparison to constant-effects model

We also specify and report a constant-effects model that does not attempt to capture IL-HTE. This model is identical to the main model except for the removal of the random effect for item-level treatment effects (i.e., $\sigma^2_{\zeta_1}$ is fixed to zero), and is given by:

$$\text{shw}_i = \eta_0 + \eta_1 \, \text{thw}_i + \varepsilon_i \tag{12}$$

$$\Phi(X_{ij}) = \beta_0 + \beta_1 \, \text{shw}_i + \theta_i + \zeta_{0j} \tag{13}$$

$$\begin{bmatrix} \theta_i \\ \varepsilon_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_\theta & \sigma_{10} \\ \sigma_{10} & \sigma^2_\varepsilon \end{bmatrix} \right) \tag{14}$$

$$\zeta_{0j} \sim N\left(0, \sigma^2_{\zeta_0}\right) \tag{15}$$

4. Results

Convergence as measured by $\widehat{R} < 1.002$ was reached at the minimum number of draws. Convergence was further confirmed by inspecting the effective sample size for each parameter; the minimum ESS was 2359. Table 1 summarizes the interpretable parameter estimates from this model in terms of the posterior median and a 95% credible interval ranging capturing the 2.5th and 97.5th percentiles of the posterior. As in the original paper (Gustafsson, 2013), we find that, under IVR assumptions, increased time spent on homework leads to higher math achievement. Specifically, we estimate a 0.717-probit increase in correct responding for each one-standard-deviation increase in normalized homework time. The residual correlation between teacher-assigned homework and student performance remains negative, consistent with findings on reverse causation from the original study (Gustafsson, 2013, p. 288).

All posterior 95% credible intervals exclude 0 except for ρ_{10}, the covariance between item difficulty and item treatment sensitivity. That is, we do not find any evidence that harder or easier items were systematically more sensitive to time spent on homework. The credible interval on $\sigma^2_{\zeta_{1j}}$ indicates that there is non-zero heterogeneity in the sensitivity of the individual items to time spent on homework. Assuming that the assumptions of IL-HTE hold, ignoring this variation would lead to under-estimated credible intervals for the main treatment effect (Gilbert, Himmelsbach, et al., 2025; Gilbert et al., 2023). We can see this in Table 1, alongside the estimates from the IL-HTE model. While differences are modest, we do observe a slightly narrower credible interval on β_1 in line with the previously documented effects of ignoring IL-HTE (Gilbert et al., 2023): the credible interval width is .107 when IL-HTE is not modeled, versus .126 when it is.

5. Conclusion

This study extended IL-HTE analysis (Gilbert, Miratrix, et al., 2025; Gilbert et al., 2023) from RCTs to IVR. A proof of concept demonstrates the use of the model, but many avenues for further research remain. These include simulation studies to establish sample size requirements under different data-generating conditions and extension to include item-level covariates. In particular, the latter is straightforward in RCTs but less simple in IVR where at least one instrument per endogenous cause is needed. Extension to binary causes (e.g., in experiments with non-compliance) is another important area for further work, and more generally, the use of IV methods with latent outcomes is greatly under-studied.

Funding Statement No funding was received for this research.

Competing Interests The authors report no competing interests for this work.

Table 1. Parameter estimates from unconditional IL-HTE model and Explanatory Item Response Model (EIRM) with an instrumental variable

Parameter	IL-HTE Median (95% CI)	EIRM Median (95% CI)	Substantive interpretation
β_0	-0.531 (-0.639 – -0.428)	-0.489 (-0.592 – -0.394)	Mean item easiness at avg amount of HW time
β_1	0.717 (0.651 – 0.777)	0.773 (0.714 – 0.821)	Mean of item-level impacts on item correctness probit
σ^2_θ	–	–	Residual θ variance (not interpretable or reported by Mplus)
σ^2_ε	0.967 (0.959 – 0.976)	0.967 (0.959 – 0.976)	Residual of HW on teacher assigned
σ_{10}	-0.693 (-0.742 – -0.634)	-0.744 (-0.791 – -0.685)	Covariance of above (not interpretable)
$\sigma^2_{\zeta_0}$	0.172 (0.122 – 0.250)	0.141 (0.096 – 0.213)	Variance of item difficulty
$\sigma^2_{\zeta_1}$	0.016 (0.012 – 0.022)	–	Variance of item-level sensitivity to treatment
ρ_{10}	0.004 (-0.009 – 0.018)	–	Covariance of above
η_0	-0.133 (-0.147 – -0.118)	-0.126 (-0.141 – -0.110)	Intercept of regression of HW time on teacher
η_1	0.096 (0.087 – 0.105)	0.091 (0.082 – 0.101)	Slope of student time spent on teacher time assigned

References

Angoff, W. H. (1981). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). The Johns Hopkins Press.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95–105. https://doi.org/10.1111/j.1745-3984.1973.tb00787.x

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69–85. https://doi.org/10.1257/jep.15.4.69

Asparouhov, T., & Muthén, B. O. (2021a). Bayesian analysis of latent variable models using Mplus (version 5). https://www.statmodel.com/download/BayesAdvantages18.pdf

Asparouhov, T., & Muthén, B. O. (2021b). Bayesian analysis using Mplus: Technical implementation (version 3). https://statmodel.com/download/Bayes3.pdf

Asparouhov, T., & Muthén, B. (2016). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In J. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 163–192). Information Age Publishing.

Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences [Publisher: Annual Reviews]. *Annual Review of Sociology*, *38*(1), 37–72. https://doi.org/10.1146/annurev-soc-081309-150141

Camilli, G. (2017). The scaling constant D in item response theory. *Open Journal of Statistics*, 7(5), 780–785. https://doi.org/10.4236/ojs.2017.75055

Cordero, J. M., Cristóbal, V., & Santín, D. (2018). Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys*, *32*(3), 878–915. https://doi.org/10.1111/joes.12217

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gilbert, J. B., Domingue, B. W., & Kim, J. S. (2025). Estimating causal effects on psychological networks using item response theory. *Psychological Methods*. https://doi.org/10.1037/met0000764

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2025). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *Journal of Policy Analysis and Management*, *44*(4), 1417–1449. https://doi.org/10.1002/pam.70025

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions. *Journal of Educational and Behavioral Statistics*, *48*(6), 889–913. https://doi.org/10.3102/10769986231171710

Gilbert, J. B., Miratrix, L. W., Joshi, M., & Domingue, B. W. (2025). Disentangling person-dependent and item-dependent causal effects: Applications of item response theory to the estimation of treatment effect heterogeneity. *Journal of Educational and Behavioral Statistics*, *50*(1), 72–101. https://doi.org/10.3102/10769986241240085

Gilbert, J. B., Zhang, L., Ulitzsch, E., & Domingue, B. W. (2025). Polytomous explanatory item response models for item discrimination: Assessing negative-framing effects in social-emotional learning surveys. *Behavior Research Methods*, *57*(4), 1–21. https://doi.org/10.3758/s13428-025-02625-2

Gustafsson, J.-E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, *24*(3), 275–295. https://doi.org/10.1080/09243453.2013.806334

Levy, R., & McNeish, D. (2023). Perspectives on Bayesian inference and their implications for data analysis. *Psychological Methods*, *28*(3), 719–739. https://doi.org/10.1037/met0000443

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2019). Instrumental variables two-stage least squares (2SLS) vs. maximum likelihood structural equation modeling of causal effects in linear regression models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 876–892. https://doi.org/10.1080/10705511.2019.1607740

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Muthén, B., & Muthén, L. (2017). Mplus. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 507–518). Chapman & Hall/CRC.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81–117. https://doi.org/10.2333/bhmk.29.81

Peterson, R. A. (2021). Finding optimal normalizing transformations via bestNormalize. *The R Journal*, *13*(1), 310–329. https://doi.org/10.32614/RJ-2021-041

Petscher, Y., Compton, D. L., Steacy, L., & Kinnon, H. (2020). Past perspectives and new opportunities for the explanatory item response model. *Annals of Dyslexia*, *70*, 160–179. https://doi.org/10.1007/s11881-020-00204-y

San Martín, E. (2016). Identification of item response theory models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 127–150, Vol. 2). Chapman & Hall/CRC.

Savalei, V. (2006). Logistic approximation to the normal: The KL rationale. *Psychometrika*, *71*(4), 763–767. https://doi.org/10.1007/s11336-004-1237-y

Shear, B. R., & Briggs, D. C. (2024). Measurement issues in causal inference. *Asia Pacific Education Review*, *25*, 719–731. https://doi.org/10.1007/s12564-024-09942-9

Stoetzer, L., Zhou, X., & Steenbergen, M. (2025). Causal inference with latent outcomes. *American Journal of Political Science*, *69*(2), 624–640. https://doi.org/10.1111/ajps.12871

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43–74). Springer.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*(1), 659–706. https://doi.org/10.1146/annurev.soc.25.1.659

Appendix 1. Mplus script to run example model

```
DATA: FILE = respdata_long.dat;
  VARIABLE:
  NAMES = student teacher tch_hwmins stu_hw item resp;
  MISSING=.;


  USEVARIABLES = stu_hw tch_hwmins resp;

  CATEGORICAL = resp;
  CLUSTER =  item;
  WITHIN = stu_hw tch_hwmins;

  ANALYSIS:
      type = twolevel random;
      estimator = bayes;
      processors = 5;
      chains = 5;
      bconvergence = .002;
      biterations = (10000);
      THIN=100;
```

```
MODEL:
    %WITHIN%

   hte | resp ON stu_hw;
   stu_hw ON tch_hwmins;
   resp WITH stu_hw;

   %BETWEEN%
   hte WITH resp;
   [ hte ];
   hte;
   [ resp$1 ];
   resp;
```