Causal Inference with Latent Outcomes[*]

LUKAS F. STOETZER[†], XIANG ZHOU[‡], AND MARCO STEENBERGEN[§]

Abstract

While causal inference has become front and center in empirical political science, we know little about how to analyze causality with latent outcomes, such as political values, beliefs, and attitudes. In this article, we develop a framework for defining, identifying, and estimating the causal effect of an observed treatment on a latent outcome, which we call the latent treatment effect (LTE). We describe a set of assumptions that allow us to identify the LTE and propose a hierarchical item response model to estimate it. We highlight an often overlooked exclusion restriction assumption, which states that treatment status should not affect the observed indicators other than through the latent outcome. A simulation study shows that the hierarchical approach offers unbiased estimates of the LTE under the identification and modeling assumptions, whereas conventional two-step approaches are biased. We illustrate our proposed methodology using data from two published experimental studies.[1]

Keywords: Causal Inference; Latent variables; Hierarchical IRT Model; Measurement

[Word Count: 9944[2]]

[*] We thank Dominik Hangartner, Neal Beck, Naoki Egami, Koskue Imai for valuable comments on previous versions of this paper and Yuteng Li for excellent research assistance. Versions of this Paper have been presented at the 2019 Meeting of DAGStat, the 2019 Annual Meeting of the Midwest Political Science Association, the 2021 Annual Meeting of PolMeth Europe, and the 2021 Annual Meeting of PolMeth XXXVIII.

[†] Corresponding author. Professor of Quantative Methods, Witten/Herdecke University. E-Mail: `lukas.stoetzer@uni-wh.de`

[‡] Associate Professor, Harvard University. E-Mail: `xiang_zhou@fas.harvard.edu`

[§] Professor, University of Zurich. E-mail: `steenbergen@ipz.uzh.ch`

[1] The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: `https://doi.org/10.7910/DVN/UZFWT7`.

[2] Includes the main body of text, notes, parenthetical references, and the headers and notes of tables and figures.

1. Introduction

Causal inference is one of the most far-reaching developments in social science methodology over the past two decades. As more scholars engage in causal inference, the range of social and political outcomes that are being investigated expands. Increasingly, the outcomes of interest include values, beliefs, and attitudes. Such outcomes are often construed as latent variables because they cannot be directly observed (Bollen 2002); as such, they are typically inferred from a set of manifest indicators through measurement models. Researchers who are interested in studying causal effects on latent outcomes, therefore, face both causal identification and measurement issues. The goal of this study is to develop a framework for drawing causal inferences with latent outcomes that accounts for both of these issues.

To illustrate our framework, we revisit two recent studies that employ field and survey experiments to draw causal inferences with latent outcomes. The first study uses a survey experiment to investigate how women's descriptive representation in decision-making bodies affects citizens' perceptions of democratic legitimacy (Clayton et al. 2019). The second study uses field experiments to evaluate the argument that non-judgmental exchange of narratives is persuasive and can reduce exclusionary attitudes (Kalla and Broockman 2020). The outcomes of the two studies, perceptions and attitudes, respectively, are both latent constructs that the researchers measure using a set of survey items. These studies are merely two examples of a large body of political science research involving latent outcomes. Of all quantitative empirical studies published in three leading journals of the discipline (APSR, AJPS, and JOP) over the past five years, roughly one third have analyzed latent outcomes. The share is particularly high for experimental studies, almost half of which investigate latent outcomes (49%).[3] In these studies, latent outcomes have been variously referred to as "attitudes," "preferences," "public opinion," or "support" (See SM A).

Despite their prevalence in political science research, we know surprisingly little about how

[3] These numbers are based on a hand-classification of 1,630 abstracts of articles published in APSR, AJPS, and JOP between 2015 and mid 2021. For more details, see Supplementary Material(SM) A.

to analyze causality with latent outcomes. While there is an increasing interest in how to use imperfect proxies in causal analysis (Knox et al. 2022; Fong and Grimmer 2021; Mayer 2019; Egami et al. 2022), ad hoc solutions still abound. From our perspective, research interest often lies in evaluating general arguments about causal effects on latent outcomes, independent of their measurement proxies. This suggests a reflective approach to studying latent variables (Borsboom et al. 2003), in which the latent outcome is defined independent of the measurement device and causally precedes its manifest indicators. In this approach, different sets of indicators serve as manifestations of the latent outcome. Yet, researchers still lack clear guidelines on how to define causal estimands involving latent outcomes, how to estimate them consistently and efficiently using manifest indicators, what identification and modeling assumptions are required, and how to assess the credibility of these assumptions.

In this article, we develop a framework for defining, identifying, and estimating the causal effect of a treatment on a latent outcome. We first use the potential outcomes notation to define the latent treatment effect (LTE), discuss how it differs from conventional causal estimands, and describe the assumptions required for identifying LTEs in randomized experiments. We highlight an exclusion restriction assumption in addition to standard identification assumptions. To identify the LTE, the treatment should affect the observed indicators only through the latent variable. Otherwise, we cannot distinguish the effect of the treatment on the latent outcome from its effect on the manifest indicators. Potential violations of this assumption may stem from demand effects (Mummolo and Peterson 2019) or social desirability bias (Grimm 2010). We show that the exclusion restriction assumption is akin to the concept of measurement equivalence, and that it is partially testable with standard psychometric tools.

Second, we provide a statistical framework, in the form of hierarchical item response theory (hIRT), that allows us to consistently estimate the LTE. The prevailing practice for estimating causal effects on latent outcomes is to use a two-step procedure—by first deriving a proxy measure of the latent outcome from a set of observed indicators and then estimating the effect of the treatment on the proxy measure. The proxy measure is often constructed using a simple average or a principal component of the observed indicators. Sometimes, researchers also turn to factor analysis or item response theory (IRT) models. These two strategies are akin to the formative indicators approach (e.g. principal component analysis) and the reflective indicators approach (e.g. factor analysis) to tackling latent variables, respectively (Borsboom et al. 2003). In contrast to these two-step approaches, the hIRT approach integrates measurement and causal analysis in a single step and unequivocally aligns with the reflective perspective. This integration eliminates the biases induced by measurement error associated with all two-step approaches. Moreover, by maximizing the marginal likelihood via the EM algorithm, the hIRT approach is statistically efficient, computationally fast, and offers valid asymptotic inference for the LTE. Importantly, we extend the hIRT approach described in Zhou (2019) to develop tests for differential item functioning (DIF) and additional violations of the identification assumptions, producing a flexible toolkit for estimating causal effects on latent outcomes. An R-package, *hIRT*, is available for implementing these methods.

We present a simulation study highlighting the advantages of the hIRT approach. Compared with two-step approaches, the hIRT estimates of the LTE are unbiased. Two-step approaches, using either reflective or formative measurement models, yield biased estimates of the LTE. The reason for this is that the standardization of the LTE in two-step approaches does not account for measurement uncertainty, while the hIRT approach does. By taking measurement uncertainty into account, the hIRT approach reports larger standard errors with accurate coverage of the true effect. Furthermore, we examine the consequences of potential violations of the exclusion restriction assumption by inducing differential item functioning (DIF) between treatment and control units. If treatment has direct effects on the indicators not through the latent variable, they will be conflated with the LTE regardless of the estimation method. This type of bias increases with the degree of DIF. Our simulation study also demonstrates the validity of our proposed extension of the hIRT approach to detect DIF.

A replication of the aforementioned studies illustrates the proposed methodology. The first study (Clayton et al. 2019) uses survey experiments to identify treatment effects of women's descriptive representation on perceptions of legitimacy. Our LTE estimates from the hIRT approach are broadly similar to the two-step estimates from the original study; yet, by accounting for measurement uncertainty, our estimates exhibit larger and more accurate standard errors. The analysis further reveals a potential violation of the exclusion restriction, casting doubt on the validity of the corresponding estimates. The application of our hIRT approach to the second study (Kalla and Broockman 2020) yields similar conclusions to those of the original article. For this study, however, we find no evidence of a violation of the exclusion restriction, providing robustness to the original findings.

The contributions of this paper are twofold. Conceptually, it is one of the first studies to integrate measurement and causal inference by defining the latent treatment effect and discussing its key identification assumptions. Second, compared with existing discussions (see Mayer 2019; VanderWeele and Vansteelandt 2022; Knox et al. 2022), it provides a more comprehensive and user-friendly suite of methods for estimating causal effects on latent outcomes and testing the exclusion restriction assumption. Taken together, this paper highlights that social science research often faces two research design challenges: identification and measurement. These two challenges should not be considered in isolation, as they are inherently intertwined.

On a more practical note, our work points to a set of considerations that applied researchers should bear in mind when designing experiments. For example, to avoid potential violations of the exclusion restriction, researchers should use reliable measures of the latent outcome that are unlikely to be affected by treatment directly. We discuss our suggestions for applied researchers in more detail in the concluding section.

2. Causal Inference, Measurement Error, and Latent Variables

2.1. Existing Work

Despite the central role of causal inference in empirical political science and the ubiquity of latent variables in theoretical models of social and political processes, the intersection between these two areas of inquiry remains underexplored. To make credible causal claims with latent variables, one needs reliable measurements of the latter. Depending on the role a latent variable plays in a particular application, the measurement problem may manifest itself in treatment, confounding variables, or the outcome of interest (e.g., Millimet 2010). So far, most scholarly attention has been devoted to the problem of measurement error in the treatment variable (Banerjee and Basu 2021; Dafoe et al. 2018; Fong and Grimmer 2021; Imai et al. 2010; Millimet 2010; VanderWeele 2022) or pretreatment confounders (Battistin and Chesher 2014; Millimet 2010; Pearl 2010), rather than the outcome variable.

Discussions of the outcome variable focus mostly on reduced statistical power for detecting effects due to measurement error (e.g., Aaby and Siddique 2021). There is little discussion of bias, let alone inferences for treatment effects on latent outcomes. This is not altogether surprising given the canonical result that in a linear model, classic measurement error in the dependent variable does not induce bias in the regression coefficients. The conditions underlying this canonical result, however, are often violated in practice. For example, it is well known that when the outcome is binary, misclassified outcomes can result in biased estimates of regression slopes (Carroll et al. 2006). In a recent review article, Knox et al. (2022) discuss several other ways contaminated outcome measures can bias estimates of causal effects, and how signed causal diagrams can be used to draw qualitative inferences about the existence and direction of theorized effects (see also Masyn 2017; Vermunt and Magidson 2021).

The measurement problem is especially salient when the outcome of interest is latent. A latent outcome is a theoretical construct that cannot be directly measured.[4] This requires researchers to consider different measurement strategies. To date, few studies have systematically discussed the definition, identification, and estimation of causal effects on latent outcomes. One exception is VanderWeele and Vansteelandt (2022), who propose a test for the structural interpretation of a latent factor model. According to the structural interpretation, only the univariate factor representation, not the indicators themselves, are causally efficacious. It means that in a randomized control trial, the treatment only affects the univariate factor representation of the observed indicators but not the indicators themselves. In a similar vein, Masyn (2017) and Vermunt and Magidson (2021) propose techniques for estimating causal effects on latent classes. These ideas build on a long tradition in structural equation modelling (SEM) and latent class analysis (LCA) that integrates inference about effects with the measurement of latent variables (Bollen 1989). Many SEM and LCA specifications, however, implicitly assume that the causal structure is valid, without stating the required identification assumptions or proposing a test for them (for exceptions, see Masyn 2017; Vermunt and Magidson 2021).[5] Additional research has focused on text data as outcomes. Egami et al. (2022) discuss how to make causal inferences with text data using a structural topic model, in which topic proportions are of key interest. Their work on text data underscores the importance of a no-interference assumption regarding the measurement device (in this case, topic models) and causal inference.

Our point of departure is similar to VanderWeele and Vansteelandt (2022) in that we highlight an exclusion restriction assumption stating that the treatment should affect the measurement indicators only through the latent construct. According to VanderWeele and Vansteelandt (2022), this causal interpretation holds under the structural interpretation of a latent factor model. These authors propose a test for the structural interpretation under an additional assumption that the latent factor model is linear, i.e., the relationships between

[4] For a discussion of different definitions of latent variables, see Bollen (2002).

[5] While some studies have applied the SEM approach to estimate causal parameters (see e.g. Rabbitt 2018; Mayer 2019), they do not provide conditions under which latent causal effects can be identified.

the latent variable and all its manifest indicators are linear. The latter assumption, however, is unrealistic when some or all of the measurement indicators are on a binary or ordinal scale, as is the case with most survey items in political science research. In this study, we do not presume a linear factor model; instead, we use an item response approach where the indicators depend on the latent variable via an item characteristic function, which can be linear, logit, or any user-specified form. Within this more general framework, we propose a hierarchical model for estimation and inference as well as a test for the exclusion restriction.

2.2. Two Perspectives on Latent Outcomes

At this point, it is worth asking whether it is sensible to draw causal inferences at the level of latent outcomes. Borsboom et al. (2003) discuss several perspectives on the status of latent variables, two of which are particularly relevant here: the formative approach and the reflective approach.

The formative approach views the nature of latent variables as dependent on specific indicators. It presupposes that a change in indicators causally precedes a change in the latent variable. The characteristics of indicators define the construct and can thereby affect its validity. This view applies to concepts such as socioeconomic status, which are conceptualized as a summary of measurable characteristics such as education, income, and occupation.

By contrast, in a reflective approach, the latent construct is independent of its measures, enabling general arguments about the latent outcome. A change in the construct is assumed to causally precede a change in indicators, which also means the indicators chosen in a particular study do not affect the validity of the construct. The measures are also deemed imperfect; the explicit accommodation of measurement error in the indicators also sets the reflective model apart from its formative counterpart.

Given these theoretical considerations, the reflective model is our preferred approach for studying attitudes, beliefs, intentions, as it enables hypotheses about these latent outcomes without referring to how they are measured. The reflective model assumes latent outcomes are real entities manifested through observed indicators. In this approach, different studies may use different sets of indicators, but these differences do not alter the underlying concept because the causal arrows flow from the latent variable to its observed indicators and not vice versa. Therefore, one can gauge the same latent variable in different ways, seek to improve statistical efficiency using different measures, and adjust measures to accommodate cross-cultural sensitivities—all without challenging the premise that the observed indicators are merely a window onto the underlying concept.

However, if one takes the reflective indicators approach, it is no longer justified to merely estimate causal effects on observed indicators. Rather, the causal effect of interest is at the level of the latent outcome. It can manifest itself in myriad ways including but not restricted to those captured by the indicators used in a particular study. In many applications, scientific interest lies in the latent outcome rather than in its myriad manifestations, and this principle should be reflected in the definition, identification, and estimation of the causal estimands. To illustrate the prevalence of this perspective in political science, we provide additional examples of latent outcomes in SM A.2.

3. Causal Inference with Latent Outcomes

3.1. The Latent Treatment Effect (LTE)

We first define the causal estimand pertaining to latent outcomes such as values, beliefs, and attitudes. Our discussion builds on the potential outcomes framework (Rubin 1974). For expositional simplicity, we focus on a binary treatment, D, that takes 1 if a unit is treated and 0 otherwise.[6] When the outcome of interest Y is observed, the treatment effect is defined as the difference between two potential outcomes, i.e., $Y(1) - Y(0)$.[7] For each unit, we observe either $Y(1)$ (if the unit is treated) or $Y(0)$ (if the unit is not treated), but not both, which is

[6] With proper modifications, our framework can be easily generalized to continuous or multidimensional treatments.

[7] We omit the unit index i for notational conciseness, with the implicit assumption that the potential outcomes refer to unit-level counterfactuals.

known as the fundamental problem of causal inference. The impossibility of observing both potential outcomes for the same unit means that we can never directly compute treatment effects at the individual level (Holland 1986). With appropriate identification strategies, however, the average treatment effect, i.e., ATE $= \mathbb{E}\left[Y(1) - Y(0)\right]$, can still be identified.

Our estimand deviates from the standard ATE because we focus on the causal effect of the treatment D on a latent outcome, which we denote by Θ.[8] Following the potential outcomes notation, we use $\Theta(1)$ and $\Theta(0)$ to represent the potential values of the latent outcome under treatment and control, respectively. Analogous to the ATE on a manifest outcome, we define the (average) latent treatment effect (LTE) as:

$$\text{LTE} = \mathbb{E}[\Theta(1) - \Theta(0)]. \tag{1}$$

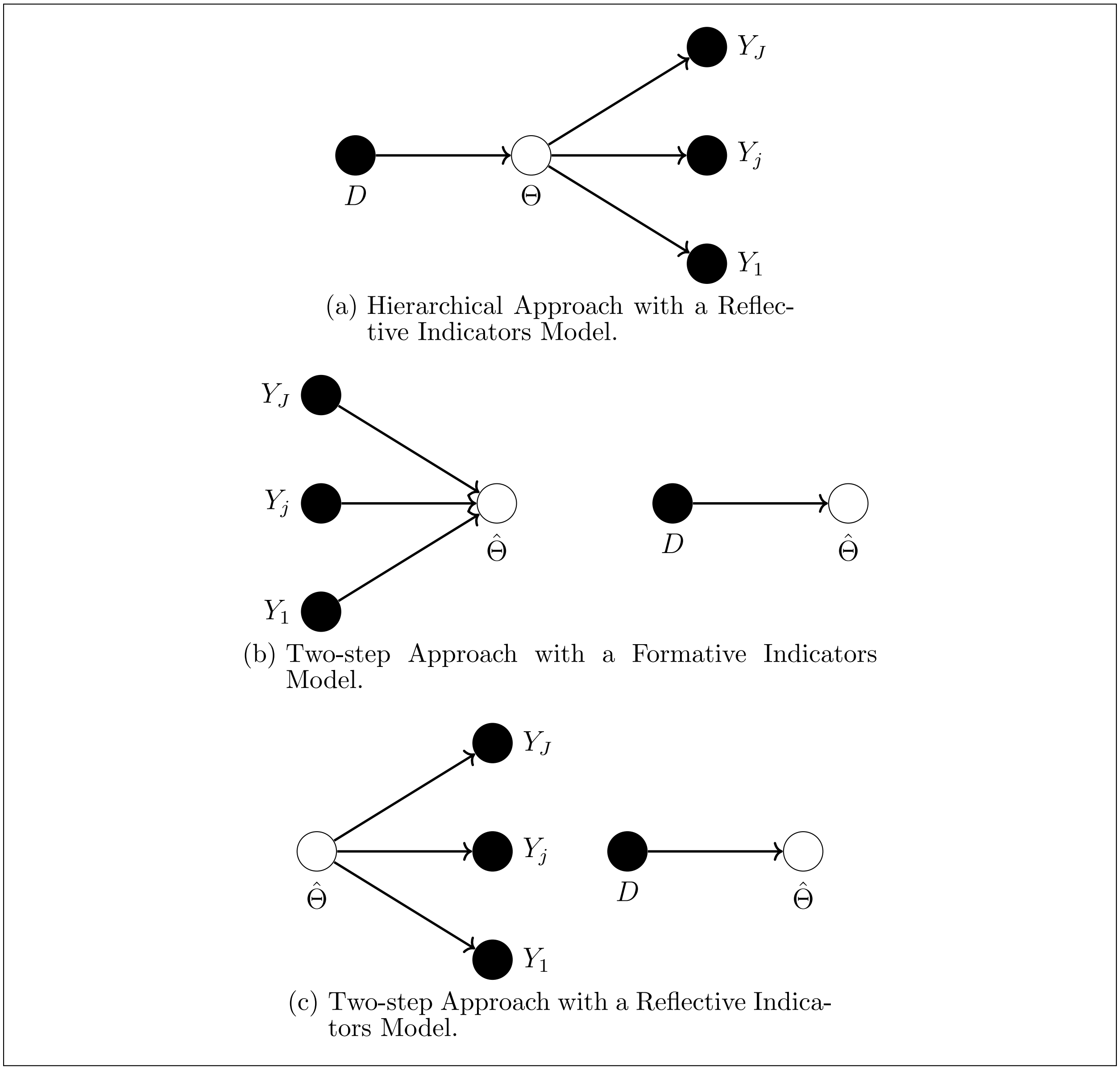
Defined in terms of a latent variable, the LTE is a theoretically motivated estimand. But it is also practically significant. In our first motivating example, the LTE corresponds to the average treatment effect of women's representation in decision-making on democratic legitimacy, a quantity that interests scholars from both a theoretical and a policy perspective. By contrast, the items used in the survey are merely a window onto the concept of legitimacy. As such, there is no inherent interest in treatment-induced differences in the observed indicators per se; a researcher could have used a different set of indicators to gauge legitimacy.

This brings us to the second fundamental problem of causal inference with latent outcomes: none of the counterfactual latent outcomes are directly observable. Unlike the standard case where one of the potential outcomes is observed, latent outcomes require measurement. With our interest in the LTE, we must use a *set* of observed indicators to measure the latent outcome Θ. The measurement device, i.e., the mapping between the latent outcome and the observed indicators, is almost always imperfect. Henceforth, we use $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_J\}$ to

[8] We use uppercase letters for random variables (e.g. Θ), uppercase letters with parentheses for the corresponding potential outcomes (e.g. $\Theta(0)$,$\Theta(1)$), and lowercase letters for fixed/realized values (e.g. d, θ).

denote a set of J observed indicators of the latent outcome Θ.

As noted previously, empirical studies have often used a two-step approach to tackle latent outcomes, where the first step involves the construction of a proxy measure for the latent variable of interest, using either a formative indicators model (e.g., principle component analysis) or a reflective indicators model (e.g., factor analysis).



(a) Hierarchical Approach with a Reflective Indicators Model.

(b) Two-step Approach with a Formative Indicators Model.

(c) Two-step Approach with a Reflective Indicators Model.

Figure 1: Different Approaches to Analyzing Latent Outcomes. D denotes treatment, Θ denotes the latent outcome of interest, and $Y_1, \ldots Y_J$ denote J observed indicators of Θ. Solid nodes are manifest and hollow nodes latent.

In Figure 1, panels (b) and (c) represent the two-step approach with a formative indicators model and with a reflective indicators model, whereas panel (a) represents the underlying causal relationships between our treatment D, latent outcome Θ, and its observed indicators $\mathbf{Y}$. We can see that a formative indicators model implies a set of causal arrows flowing from the observed indicators $\mathbf{Y}$ to the latent outcome $\boldsymbol{\Theta}$, which is at odds with their posited relationships (panel a). By contrast, the two-step approach with a reflective indicators model preserves the underlying causal structure. However, both approaches split measurement and analysis into two steps, which, as noted earlier, can complicate the inference of treatment effect estimates. In section 3.3, we introduce a hierarchical approach that directly models the underlying causal structure in panel (a), circumventing the limitations of the two-step approaches. Below, we outline a set of assumptions for identifying the causal effect of D on Θ.

3.2. Identification Assumptions

Given the DAG $D \rightarrow \Theta \rightarrow \mathbf{Y}$ presented in panel (a) of Figure 1, we now specify the assumptions needed for identifying the LTE:

1. Stable Unit Treatment Value Assumption (SUTVA): $\Theta_i = \Theta_i(D_i)$ and $\mathbf{Y}_i = \mathbf{Y}_i(D_i, \Theta_i)$, where D_i and Θ_i denote treatment status and the latent outcome for unit i.

2. Unconfounded Treatment: $\big(\Theta(d), \mathbf{Y}(d, \theta)\big) \perp\!\!\!\perp D$ for any d and θ.

3. Unconfounded Measurement: $\mathbf{Y}(d, \theta) \perp\!\!\!\perp \Theta | D$ for any d and θ

4. Exclusion Restriction: $\mathbf{Y}(d, \theta) = \mathbf{Y}(d', \theta)$ for any d, d' and θ.

Assumption 1 (SUTVA) is a standard assumption in the potential outcomes approach to causal inference. It requires that there must not be any interference between individuals or multiple versions of the treatment so that the realized outcomes Θ and $\mathbf{Y}$ for each unit are equal to their potential outcomes under the realized values of their antecedent variables.

In our case, since the observed indicators $\mathbf{Y}$ may depend on both treatment status and the latent outcome, we use $\mathbf{Y}(d, \theta)$ to denote their potential values when D is set to d and Θ set to θ. Experimental designs address SUTVA by randomizing the same type of treatment to respondents that are isolated from each other. In our running examples, the same types of treatments (in the form of protocol or information treatments) are administered to respondents in isolation from other participants.

Assumption 2 (Unconfounded Treatment) is met by design in experiments where treatment is randomly assigned. Research often relies on balance checks to show that the random assignment of treatment status created groups with similar observed characteristics. The unconfounded treatment assumption, however, can be extended for observational studies to conditional unconfoundedness, i.e., $\big(\Theta(d), \mathbf{Y}(d, \theta)\big) \perp\!\!\!\perp D|\mathbf{X}$, where $\mathbf{X}$ denotes a set of pretreatment confounders of the treatment-outcome relationship. In section 6, we discuss this point in more detail and outline potential applications of our approach beyond randomized experiments.

Assumption 3 (Unconfounded Measurement) assumes that no unobserved confounders are allowed to affect both the latent variable Θ and the outcomes $\mathbf{Y}$. This assumption ensures that the measurement model is unconfounded, and it is implicit in most item response and factor-analytic models. The assumption accentuates the importance of reliable and valid measurement models in the analysis of causal effects on latent outcomes.

There are different ways in which the unconfounded measurement assumption can be violated. One possible violation is when researchers assume a single underlying construct, but several latent dimensions influence the responses and the latent constructs have complex interactions with each other. This might occur, for example, if the two dimensions of our first running example, substantial and procedural legitimacy, were not separated, but procedural legitimacy affects substantial legitimacy. Another possibility is that certain confounders influence the measurement indicators above and beyond the latent variable, which is sometimes referred to as method effects (Eid et al. 2016). In section 6, we discuss how our framework can be extended to accommodate such violations of unconfounded measurement.

Assumption 4 (Exclusion Restriction), in our context, means that the treatment D affects the observed indicators $\mathbf{Y}$ only via the latent outcome Θ. This is a central assumption for identifying the LTE that has not been fully acknowledged and formalized in the literature. If receiving the treatment induces changes in the measurement indicators above and beyond changes in the latent variable, it will be difficult to attribute the changes in the indicators to a treatment effect on Θ. The assumption is closely related to the exclusion restriction assumption in instrumental variable (IV) models, except that in our case, the intermediate variable is latent and there can be multiple indicators as measured outcomes.

Potential violations of the exclusion restriction assumption can arise from induced demand effects (Mummolo and Peterson 2019) or social desirability bias (Grimm 2010), in which respondents feel compelled to answer in a certain way even if their latent attitude is unchanged. When the treatment in an experiment highlights a social norm and makes the purpose of the experiment clear to the participants, it may increase the likelihood of social desirability bias. This, for example, can occur in our first study where the treatment involves presenting respondents with a picture of a decision panel with men and asking about the procedural legitimacy of an anti-feminist decision, as it reinforces a norm of gender equality. This effect may be particularly strong if the treatment is administered just before the measurement of the central outcomes. In the second example, a door-to-door canvasing treatment may also lead to social desirability bias in the conversation. However, if the outcome measures are collected in an independent survey weeks after the treatment, the effects of this bias may be reduced.

Overall, our discussion reveals previously overlooked assumptions for the identification of causal effects on latent outcomes. Compared to causal inference with manifest outcomes, the exclusion restriction plays a central role in the identification and estimation of the LTE. This assumption makes it possible to attribute treatment-induced variation in the indicators to treatment effects on the latent outcome. Without this assumption, the observed data can be consistent with an observationally equivalent model where changes in the indicators occur without any changes in the latent variable of interest (see Masyn 2017; VanderWeele and Vansteelandt 2022; Vermunt and Magidson 2021). We propose a method for testing this assumption in Section 3.4.

3.3. Estimation

The LTE is defined as the difference in expectation between the two potential outcomes $\Theta(1)$ and $\Theta(0)$. Thus, the key to estimating the LTE is to evaluate the mean potential outcome $\mathbb{E}[\Theta(d)]$ for both $d = 0$ and $d = 1$. Given the exogeneity of the treatment (Assumption 2), we can rewrite $\mathbb{E}[\Theta(d)]$ as

$$\mathbb{E}[\Theta(d)] = \mathbb{E}[\Theta \mid D = d]. \tag{2}$$

In general, researchers interested in estimating the LTE needs to evaluate this expression. There are potentially several approaches that can be used for this purpose. Below, we discuss an apparent two-step approach, which involves first measuring the latent variable and then performing inference on it, and a hierarchical item response model, which combines measurement and causal analysis into a single step.

3.3.1. Two-step approaches

Equation (2) suggests a two-step approach to estimating the LTE, which involves first obtaining proxy measures of the latent variable Θ and then taking a difference in means of these measures between the treated and control units. As noted earlier, such a two-step approach, whether the first step is a formative indicators model or a reflective indicators model, has been widely used to evaluate treatment effects on latent constructs in political science. Below, we show that the two-step approach can result in biased estimates of the LTE.

To see why the two-step approach is biased, let us first, using the law of total expectation, rewrite the right-hand side of equation (2) as

$$\mathbb{E}[\Theta \mid D = d] = \mathbb{E}\big[\mathbb{E}[\Theta|\mathbf{Y}, D] \mid D = d\big] \tag{3}$$

This equation suggests that we could, for example, use a factor-analytic model to first estimate $\mathbb{E}[\Theta|\mathbf{Y}, D]$ and then evaluate how estimates of these conditional means differ between treated and control units. The standard two-step approaches, however, do not condition on treatment status D in obtaining the proxy measures of Θ. Instead, they approximate $\mathbb{E}[\Theta|\mathbf{Y}, D]$ with estimates of $\mathbb{E}[\Theta|\mathbf{Y}]$. In other words, researchers often first calculate factor scores or principal component scores without considering treatment status, and then evaluate how these scores differ by treatment status. However, from the DAG (Figure 1a), we can see that the latent outcome is still affected by treatment status even after we condition on $\mathbf{Y}$, which means $\mathbb{E}[\Theta|\mathbf{Y}, D] \neq \mathbb{E}[\Theta|\mathbf{Y}]$. Therefore, the two-step procedure described above will likely result in biased estimates of the LTE.

Given the above discussion, it might be supposed that we could avoid the bias by conditioning on D in the two-step approach. For example, we could imagine constructing proxy measures of the latent outcome separately for the treated and control units. Unfortunately, this is not a workable solution because the latent outcome, by definition, has no intrinsic scale. If we used two separate measurement models for the treated and control units, the latent outcome would be standardized, for example, to have a mean of zero and a standard deviation of one, for each group. Consequently, no meaningful comparisons can be made between the treated and control units.

In fact, even if we could condition on treatment status while constructing a proxy measure of the latent outcome that lies on the same scale for treated and control units, measurement uncertainty would still lead to biased estimates of the LTE. To see this point, we first note that because the latent outcome Θ has no intrinsic scale, the magnitude of the LTE must be evaluated on a relative basis. Suppose, for example, we evaluate the LTE in terms of the standard deviation of Θ in the population, then our estimand can be written as

$$\text{LTE} = \frac{\mathbb{E}[\Theta | D = 1] - \mathbb{E}[\Theta | D = 0]}{\text{sd}[\Theta]} \tag{4}$$

Now, consider a plug-in estimate of equation (4) where Θ is replaced by an imperfect proxy, say $\hat{\Theta}$. Without loss of generality, suppose $\hat{\Theta} = a + b\Theta + e$ where a and b are constants and e is an error term independent of Θ. Then the plug-in estimate of equation (4) will equal

$$\widehat{\text{LTE}} = \frac{b \cdot \text{sd}(\Theta)}{\sqrt{b^2 \text{sd}^2(\Theta) + \sigma_e^2}} \text{LTE},$$

where σ_e^2 denotes the variance of e. This expression makes it clear that the plug-in estimate will be biased unless $\sigma_e^2 = 0$, i.e., unless the proxy measure $\hat{\Theta}$ is perfect (equivalent to the true latent outcome up to a linear transformation). In other words, even if the proxy measure is "unbiased" in the sense that $a = 0$ and $b = 1$, the plug-in estimate of the LTE will still be biased. This result contrasts sharply with the case of manifest outcomes, in which classic measurement error does not induce bias in estimated regression coefficients. When the outcome is latent and thus has no intrinsic scale, measurement error *will* translate into a bias in the estimated treatment effects.

3.3.2. Hierarchical item response theory model

To circumvent the problems associated with the two-step approaches, we propose the use of hierarchical item response theory (hIRT), a model-based approach, to directly estimate the effect of D on Θ—without constructing intermediate estimates of $\mathbb{E}[\Theta \mid \mathbf{Y}, D]$. The model consists of two components, one on the dependence of $\mathbf{Y}$ on Θ and the other on the dependence of Θ on D:

$$\begin{aligned} \mathbb{P}[Y_j = h \mid \Theta = \theta] &= P_{jh}(\theta; \alpha_{jh}, \beta_j) && (5) \\ \Theta &= \gamma_0 + \gamma_1 D + \epsilon. && (6) \end{aligned}$$

(for $h = 1, \cdots, H$). In the above equations, the parameter γ_1 denotes the average treatment effect of D on Θ, i.e., the LTE. The function $P_{jh}(\cdot)$ is the item characteristic function linking the observed indicator Y_j to the latent outcome Θ, and α_{jh} and β_j are the item difficulty and item discrimination parameters for item j and answering category h, respectively. The item characteristic function can be applied to binary indicators using the logit or probit link, or to ordinal responses using the graded response model.[9] The latent outcome Θ is modelled hierarchically as a function of the treatment plus a random noise ϵ. In our applications, we assume that ϵ is normally distributed with a constant variance σ^2.

The hIRT model makes a set of parametric assumptions about how the latent variable relates to the indicators. Our model assumes (a) a uni-dimensional latent variable, (b) local independence of items conditional on the latent variable, and (c) a particular item characteristic function. It bears noting that the aforementioned two-step approaches rely on similar parametric assumptions to obtain a proxy measure of the latent outcome before making inferences about the effects. For example, a factor-analytic approach rests on the same assumptions of unidimensionality, local independence, a functional relationship between the latent trait and the responses. Moreover, given that treatment is binary, no parametric assumptions about the $D - \Theta$ relationship is needed, as the γ_1 parameter in equation (6) is equivalent to a difference in means between treated and control units. In sum, the hIRT approach imposes no more additional parametric assumptions than conventional two-step approaches, although it makes the assumptions required more transparent.

To identify the hIRT model, several identification constraints have to be imposed. Specif-

[9] For more details on the IRT specification, see SM C.

ically, a location constraint must be imposed on either γ_0 or the item difficulty parameters α_{jh}, and a scale constraint must be imposed on either $\{\gamma_1, \sigma^2\}$ or the item discrimination parameters β_j. For example, we can set $\gamma_0 = 0$ so that the prior mean of the latent outcome among control units equals zero, and set $\sigma^2 = 1$ so that the prior conditional variance of the latent outcome given D equals one. Alternatively, we could impose location and scale constraints on the item parameters, such as $\sum_{j,h} \alpha_{jh} = 0$ and $\prod_j \beta_j = 1$.

After the identification constraints are imposed, the hierarchical IRT model can be estimated via an Expectation-Maximization (EM) algorithm described in Zhou (2019). Basically, the EM algorithm treats the latent outcome Θ_i as missing data and maximizes the marginal likelihood for γ_0, γ_1, as well as the item parameters α_{jh} and β_j. We can see that unlike conventional two-step approaches (using either a formative indicators model or a reflective indicators model), the latent outcomes are not explicitly estimated in this approach. However, $\hat{\gamma}_1$ gives a consistent and asymptotically normal estimate of the LTE. Moreover, consistent standard errors of all the parameters can be derived from either the Hessian matrix or the outer product of the gradients of the log marginal likelihood. We describe the estimation method in more detail in SM C.2.

3.4. Testing the exclusion restriction

One appealing feature of our framework is that the exclusion restriction assumption is partially testable in our model. Unlike in instrumental variable settings, our unconfounded measurement assumption (Assumption 3) allows us to test if D has a direct effect on $\mathbf{Y}$ above and beyond the pathway $D \to \Theta \to \mathbf{Y}$. This relates closely to the concept of measurement equivalence in psychometrics (see Meredith 1993). Because $\mathbf{Y}(d, \theta) \perp\!\!\!\perp \Theta | D$ and

$\mathbf{Y}(d, \theta) \perp\!\!\!\perp D$ implies $\mathbf{Y}(d, \theta) \perp\!\!\!\perp D|\Theta$, we have

$$
\begin{aligned}
f(\mathbf{Y}|\Theta = \theta, D = d) &= f(\mathbf{Y}(d, \theta)|\Theta = \theta, D = d) && (7)\\
&= f(\mathbf{Y}(d, \theta)|\Theta = \theta) && (8)\\
&= f(\mathbf{Y}|\Theta = \theta) && (9)
\end{aligned}
$$

which implies measurement equivalence between treatment and control units. In IRT models, measurement equivalence is violated in the presence of differential item functioning (DIF) by treatment status (Osterlind and Everson 2009). If the exclusion restriction holds, there should be no systematic difference between treatment and control groups in their response patterns conditional on the latent variable.

Two conceptually different types of DIF should be distinguished when discussing violations of the exclusion restriction in our context: uniform DIF and non-uniform DIF. Uniform DIF implies a shift in the difficulty parameter of an item. For example, comparing a person in the control group with a person in the treatment group, uniform DIF can lead one of the respondents to be more likely to answer that "they should allow many undocumented immigrants to become U.S. citizens" even if both have the same latent immigration attitude. One reason for this could be that treatments pressure respondents to answer in a particular way without changing their actual immigration attitude. Non-uniform DIF affects the discrimination parameter of an item. With higher item discrimination, latent attitudes more strongly influence answers to the indicator questions. In this case, treatment could strengthen or weaken the discriminatory power of certain indicators. For example, an intervention to reduce exclusionary attitudes that focuses heavily on undocumented immigrants could strengthen the relationship between latent immigration attitudes and the aforementioned indicator. Both types of DIF constitutes a violation of the exclusion restriction.

Based on these insights, we extend the hIRT model to develop DIF tests for violations of equation (9). Specifically, to test for uniform DIF for a particular item or a set of items, we fit an "augmented" hIRT model where treatment D is allowed to affect the item responses directly (not only through its effect on Θ) and then conduct a likelihood-ratio test comparing the generalized hIRT model with the baseline model. To test for non-uniform DIF, we fit another augmented hIRT model where the treatment variable enters the $P_{jh}(\cdot)$ model not only through its main effect on Y_j but also through its interaction effect with the latent variable Θ. Then, a likelihood-ratio test comparing the two augmented models can be used to detect non-uniform DIF.[10]

4. Simulation Results

In this section, we use a simulation study to evaluate the properties of different estimators of the LTE and their biases in the presence of exclusion restriction violations. We also demonstrate the ability of our DIF test to reveal exclusion restriction violations.

We consider a simple data-generating model with a constant treatment effect on the latent variable: $\theta_i \sim \mathrm{N}(\gamma_0 + \gamma_1 d_i, 1)$, where units are randomized into treatment and control conditions. We draw the answering patterns on J manifest items from a graded response model. We sample the item discrimination parameter from a log-uniform distribution over the interval (-1, 1) ($\log \beta_j \sim \mathrm{Unif}(-1, 1)$). The item difficulty parameters for item j $(\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jH-1})$ are taken from independent draws from a uniform distribution over the interval $(-H+1, H-1)$, where H is the number of response categories. We then simulate item response y_{ij} from a graded response model. Apart from this baseline setup, we also impose both uniform and non-uniform DIF as violations of the exclusion restriction on Q items. For treated units, uniform DIF shifts the latent response propensity for item q by λ_q, whereas non-uniform DIF increases the item discrimination parameter of item q by a positive constant ϕ_q.[11]

[10] In SM D we, furthermore, describe a common alternative DIF test based on logistic regression models.

[11] The latent response propensity for the item q becomes $y^*_{iq} = \lambda_q d_i + (\beta_q + \phi_q d_i)\theta_i$.

In all simulations, we fix the sample size to 1,000 and the number of items to $J = 5$ with an equal number of categories ($H = 5$).[12] We vary the latent treatment effect from small $\gamma_1 = 0.1$, medium $\gamma_1 = 0.25$, large $\gamma_1 = 0.5$, to very large $\gamma_1 = 0.75$ (with $\gamma_0 = 0$). DIF is then imposed for zero, one, and two items ($Q \in \{0, 1, 2\}$). We vary the strength of the uniform DIF from 0 to 0.75 ($\lambda_q \in \{0, 0.25, 0.5, 0.75\}$). In addition, we consider two versions with and without non-uniform DIF: $\phi_q \in \{0, 0.5\}$.

We generate 1,000 random samples in each setting. For each sample, we estimate the LTE using our proposed hierarchical graded response Model (hgrm) and four two-step approaches with different measurement methods: a) taking the average of all scores (two-stepe AVE), b) using the first principal component scores as the outcome as a formative measurement model (two-step PCA), c) using factor scores as a reflective measurement model (two-step FA), and d) using the graded response model in a two-step procedure (two-step hgrm). As in most applications, the two-step approaches do not condition on the treatment in the measurement stage.[13] The theoretical discussion of this approach predicts biased estimates of the LTE. The simulation allows us to investigate the extent and direction of this bias. To make the estimates comparable, we re-scale the outcome variables, such that the effects can be interpreted in terms of the total standard deviation in the latent outcome. After estimation, we run the DIF tests to detect potential violations of the exclusion restriction.

Figure 2 shows that when the identification assumptions hold, only the hIRT approach provides unbiased estimates of the LTE. All two-step procedures exhibit a bias (around 22% on average) towards zero. We can see that the size of the bias increases as the LTE increases. We can also see a large variability of $\widehat{\text{LTE}} - \text{LTE}$ across simulations. Hence, in a particular application the two approaches might not show a marked difference in their estimates, but on average the two-step approaches suffer a clear bias. Overall, the bias does not depend much on whether the first step estimates come from an index, principal component analysis,

[12] SM E.5 presents additional results with different numbers of items and categories.

[13] An alternative two-step approach is fitting the measurement model using only the control unit and then using the fitted model to obtain latent scores for both treated and control units. Results from this approach are similar and detailed in Appendix E.6.

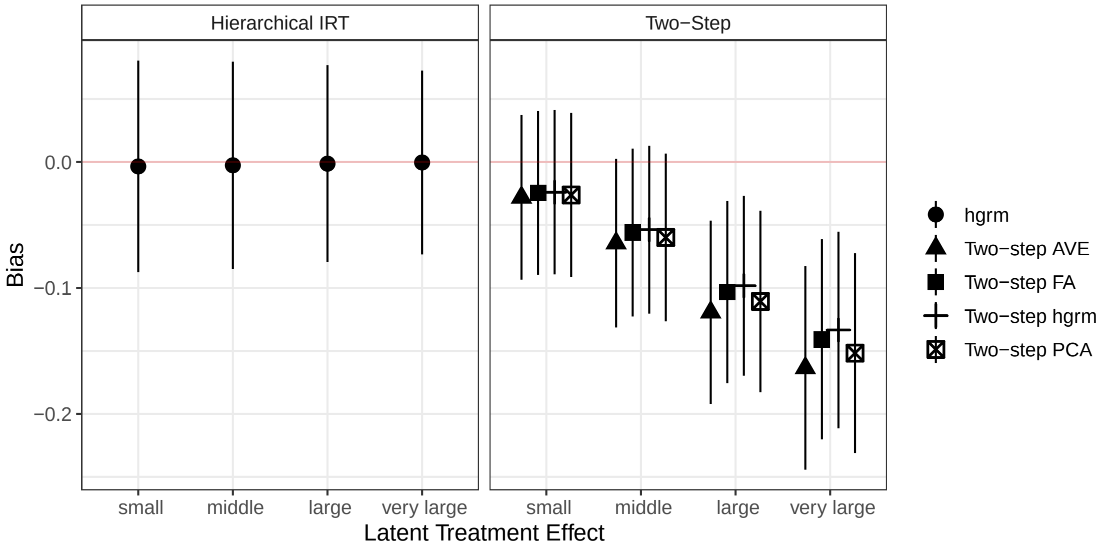


Figure 2: Simulation results for the bias of different estimation methods without differential item functioning. The figure shows the estimated bias as well as the variability of $\widehat{\text{LTE}} - \text{LTE}$ measured in terms of standard deviations of Θ.

factor scores, or an IRT model. This means that the bias is not a result of misspecification of the measurement model, as even the correctly specified two-step IRT model exhibits a bias. Instead, the bias stems from measurement error in the first-step estimates. Hence, to obtain unbiased estimates, the hierarchical IRT approach should always be preferred.

In addition, the hierarchical approach leads to larger sampling variation. Its correct standard error estimates guarantee nominal coverage of the confidence intervals (See SM E.2). By contrast, the coverage of the two-step procedures for large and very large effects is often below 80%. The larger standard errors make the RMSEs of the hIRT estimates comparable to those from the two-step procedures when the effect size is small (See SM E.1). When the effect size is large, the hierarchical approach is superior in terms of both bias and the RMSE.

What happens if the exclusion restriction is violated? Figure 3 shows that the bias for the hierarchical IRT estimator increases linearly with the degree of uniform DIF.[14] This comes as no surprise, as part of the induced effect on the outcome indicators is mistaken as changes in the latent variable. The biases are of modest magnitude: with uniform DIF half the size

[14] SM E.3 shows additional results for other values of γ_1.

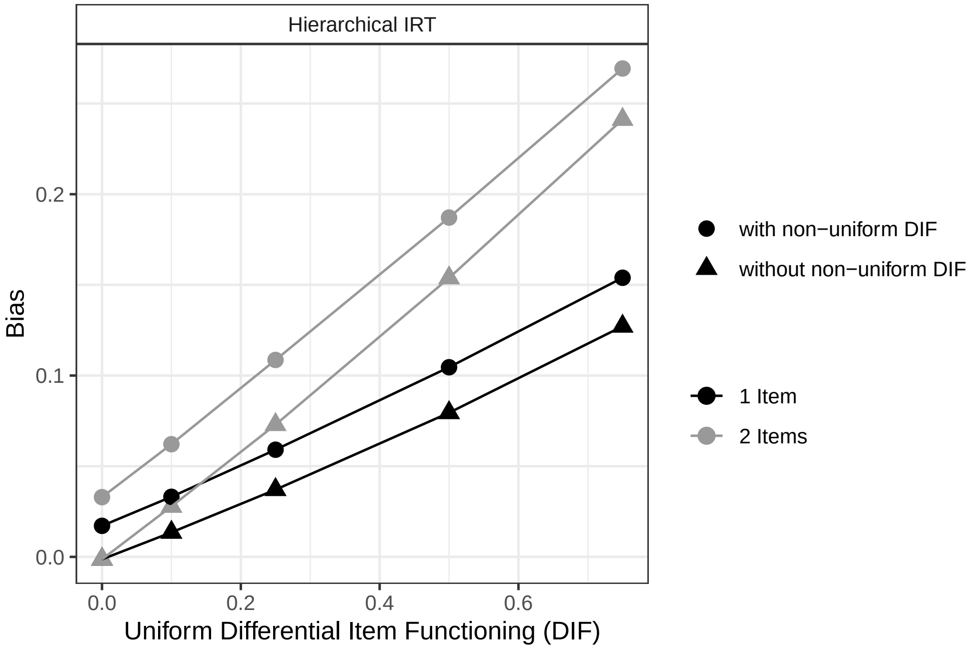


Figure 3: Simulation results for the bias of the hierarchical IRT estimator under exclusion restriction violation, where $\gamma_1 = 0.5$.

of the treatment effect (.25 standard deviations), the LTEs are overestimated by 0.04 (for one item) and 0.08 (for two items). In the case where the uniform DIF is comparable to the LTE, the bias is 0.08 for one item and 0.16 for two items. Non-uniform DIF also induces bias. All simulations that have non-uniform DIF reveal higher bias. SM E.3 shows that the increased bias due to non-uniform DIF is higher when the LTE is stronger.

Our simulation results also reveal that our proposed DIF test works to detect violations of the exclusion restriction. SM E.4 shows that our test has a high sensitivity for strong violations of the exclusion restriction. With a uniform DIF of 0.75, the sensitivity is above 90%. For minor violations that only induce small bias, the test has a lower sensitivity. For a uniform DIF of 0.25, the sensitivity is around 28% for one item and 36% for two items. The sensitivity increases in the presence of non-uniform DIF. Our test is also well-calibrated: in the case of no DIF, the test indicates the expected 5% frequency of violations.[15]

For practitioners, our simulation results lead to two key takeaways. First, use the hierar-

[15] A test based on nested logistic regressions (se SM D) performs similarly to our routine, but has slightly lower sensitivity rates.

chical IRT model instead of two-step procedures to reduce bias. Second, use our DIF tests to detect potential violations of the exclusion restriction that would lead to biased estimates.

5. Empirical Illustrations

5.1. Descriptive Representation and Democratic Legitimacy

In this section, we illustrate our procedure by reanalyzing data from a survey experiment that investigates if representatives from underrepresented groups legitimize outcomes and decision-making procedures in the eyes of the public. Clayton et al. (2019, p.114) "provide the first causal test examining how women's descriptive representation affects citizens' perceptions of democratic legitimacy." They employ a survey experimental design in which respondents are exposed to a newspaper article describing a committee's decision about penalties for sexual harassment in the workplace. The experiment varies the gender composition of the eight-member committee using a picture and a headline reference either to an all-male panel or a gender-balanced panel. The experiment further varies the decision reached by the committee to either increase ("feminist decision") or decrease ("anti-feminist decision") penalties for sexual harassment in the workplace.

The survey experiment uses multiple indicators to measure the perceived legitimacy of the decisions. In particular, they argue that descriptive representation can impact immediate reactions to a decision's content (substantive legitimacy). Women's presence can also affect perceptions of fairness in decision-making, including assessments of the process, acquiescence to decisions, and trust in representative institutions (procedural legitimacy). For both dimensions of legitimacy, they use a set of four-point Likert scale indicators, three questions for substantive legitimacy, and four questions for procedural legitimacy (See SM B.1). After establishing that the items map onto two separate latent constructs, they form two factor scores, which are used as the main outcome of the experiment. As discussed above, this represents a two-step approach to estimate the causal effects on the latent legitimacy

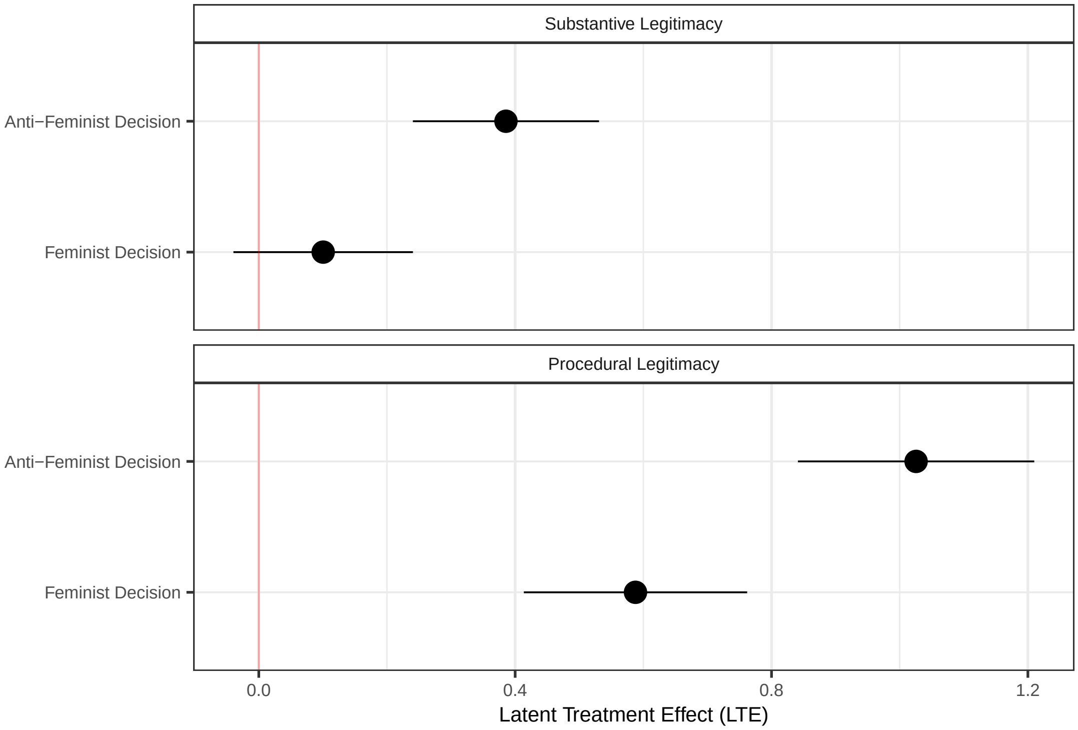


Figure 4: Estimates of latent treatment effects of an all-male versus a gender-balanced panel for feminist and anti-feminist decisions based on the hierarchical item response model (hIRT).

perceptions.

We reanalyze the authors' data using the hierarchical item response model (hIRT) to estimate the LTE.[16] Figure 4 shows the LTE estimates based on the hIRT model, comparing a gender-balanced panel to an all-male panel. We report the standardized effect sizes and find substantial effects for both substantive and procedural legitimacy. The effect size is larger for the anti-feminist decision. If a gender-balanced panel reaches an anti-feminist decision, respondents see the decision to have 0.39 standard deviations more substantive legitimacy and 1.03 standard deviations more procedural legitimacy compared to when the decision is reached by an all-male panel. This constitutes a sizable effect and clearly confirms that representation matters, especially for procedural legitimacy. Procedural legitimacy also increases by 0.59 standard deviations for feminist decisions, which implies that the process

[16] We analyze the data from the authors' Amazon Mechanical Turk sample. SM B provides the underlying DAG for the latent outcome of substantive legitimacy within our framework.

	Label	Uniform	Non-uniform	Overall
Item 1	FairProcess	86.21^{***}	4.58	90.79^{***}
Item 2	TrustCommittee	21.68^{***}	5.23	26.9^{***}
Item 3	OverturnR	33.82^{***}	11.51^{**}	45.33^{***}
Item 4	TrustLegislature	67.78^{***}	3.08	70.85^{***}

(a) Procedural Legitimacy

	Label	Uniform	Non-uniform	Overall
Item 1	RightDecision	60.31^{***}	0.93	61.24^{***}
Item 2	RightDecisionGroup	2.49	0.96	3.45
Item 3	FairDecision	20.79^{***}	1.99	22.77^{***}

(b) Substantive Legitimacy

Table 1: Testing exclusion restriction using the differential item functioning test. Table reports the χ^2-values of nested Likelihood-ratio tests. Stars indicate statistical significance ($***$ p-val < 0.001; $**$ p-val < 0.01; $*$ p-val < 0.05)

reached with a gender-balanced panel is generally perceived to be more legitimate. Similar to the original article, we find a small and statistically insignificant effect of gender composition on substantive legitimacy with a feminist decision.[17]

The effects based on the hIRT model are comparable to the effects of the original article (see SM F.1.1). We observe a small difference in the effect estimates for procedural legitimacy. Compared with the two-step estimate, the hIRT estimate is larger for the feminist decision (0.59 versus 0.51), but slightly smaller for the anti-feminist decision (1.03 versus 1.09). In light of the simulation results, it is difficult to tell where the deviations come from, as the bias of the two-step approaches can be offset by an bias that results from potential violations of the exclusion restriction. The standard errors are slightly larger for the hIRT, as the model takes estimation uncertainty into account when estimating the LTE.

The large effect estimates should be interpreted with caution. The DIF test reveals that for both scales, the exclusion restriction is likely violated. Table 1 reports results from the DIF tests, which show that for the procedural legitimacy scale, all four items exhibit uniform DIF, which implies that descriptive representation has a direct effect on the observed indicators

[17] We also estimated the conditional latent treatment effects for men and women (See SM F.1.2).

that does not operate through a change in the latent variable. This result calls into question the estimates we report and those reported in the original article. Our simulation study suggests that this violation can substantially bias the estimated latent treatment effects. It is likely that the reported estimates overstate the true LTE. One reason for the violation could be a social desirability bias induced by priming people with an all-male versus gender-balanced panel when asking about the procedural legitimacy of decisions about gender-related policies. This could work in a way similar to gender interviewer effects in questions about women's movement, women's issues, and gender equality (Huddy et al. 1997). The authors discussed this point (Clayton et al. 2019, p.126), although they did not formalize the underlying assumptions and put it to a test.

5.2. Reducing exclusionary attitudes

While latent outcomes are prevalent in the recent surge of survey experiments, other experimental work also studies latent treatment effects. We now apply our model to revisit a field experiment that evaluated different conversation strategies to reduce exclusionary attitudes. Kalla and Broockman (2020) present results from three field experiments testing the theoretical argument that non-judgmental exchange of narratives is persuasive and can reduce exclusionary attitudes. In these experiments, they randomly "varied the presence of the non-judgmental exchange of narratives strategy"(Kalla and Broockman 2020, p.2) to study if this strategy is more successful than standard protocols. The latent outcomes in these field experiments are different aspects of exclusionary attitudes towards unauthorized immigrants and transgender people.

We reanalyze data from the first field experiment about reducing exclusionary attitudes towards unauthorized immigrants. In this study, the authors conclude that "door-to-door canvassing conversations that employed [non-judgmental exchange of narratives] strategy reduced exclusionary attitudes towards unauthorized immigrants for at least four months, whereas otherwise identical conversations that omitted this strategy had no detectable ef-

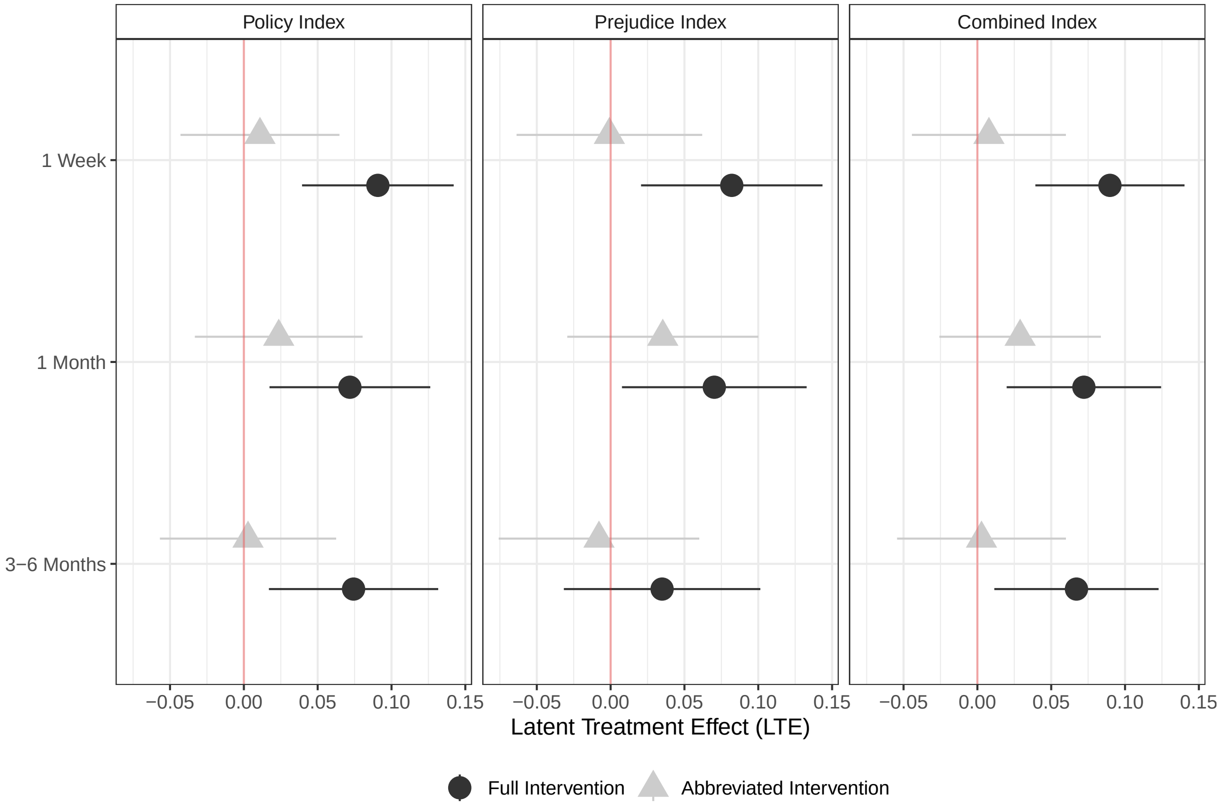


Figure 5: Estimates of latent treatment effects of full intervention and abbreviated intervention on policy index, prejudice index, and the combined index using the hierarchical IRT approach.

fects." The authors use six items to measure support for policies related to immigrants and seven items to measure anti-immigrant prejudice. Using factor scores they combine the items into an "anti-immigration policy index," an "anti-immigration prejudice index," and a "combined index." They report intent-to-treat effects calculated from a linear regression model with pre-treatment variables as controls.

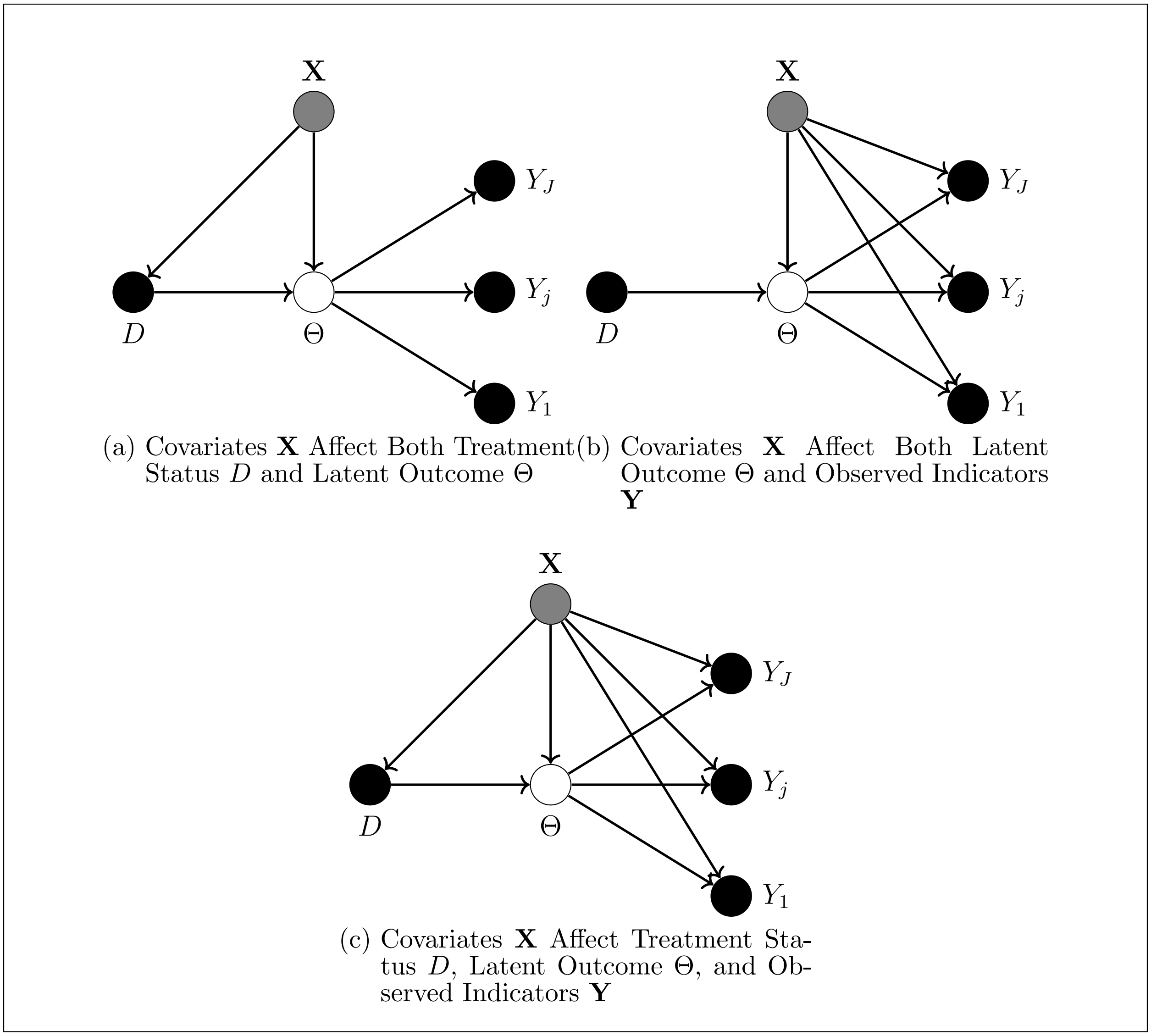
Our reanalyses using the hIRT Model confirm the conclusions from the original study. Figure 5 reports our LTE estimates. The results show the lasting effects of the full intervention on immigration policy attitudes, as well as the decaying effects on prejudice attitudes. The results are comparable to those from the two-step approach of the original study, and differ mostly in the uncertainty they convey (See SM F.2.2). With a small effect size (below 0.1 standard deviations), our simulation results suggest that we should not expect a large difference between the two approaches. But the hIRT standard errors are slightly larger as they account for measurement uncertainty. The estimated LTEs also show that the abbreviated canvassing protocol had no statistically significant effects on any of the three outcomes.

In contrast to the first application, we find DIF only for some of the outcomes. The tests indicate no uniform DIF for any of the items, but some indication of non-uniform DIF for prejudice and policy items (see SM F.2.1). Our simulations indicate that non-uniform DIF alone does not necessarily result in biased estimates. Thus, these small deviations might not have impaired the results. However, to be cautious, researchers can always exclude the affected items from the scale and reanalyze the data. In this application, excluding the items does not substantially affect the estimated LTE of the protocols. This strengthens the conclusion that the new protocol significantly altered exclusionary attitudes towards immigrants and not just their manifest indicators.

6. Extensions To Accommodate Confounding

The identification assumptions outlined in Section 3.2 stipulate that both the treatment-outcome relationship and the measurement device are unconfounded. Both of these assumptions can be relaxed in our framework to incorporate observed confounders.

First, in observational studies where treatment is not randomly assigned or experimental studies where the probability of receiving treatment is a function of observed covariates (e.g., in stratified randomization designs), we can modify our Assumption 2 to $\big(\Theta(d), \mathbf{Y}(d, \theta)\big) \perp\!\!\!\perp D|\mathbf{X}$, which allows for observed confounding of treatment assignment by a set of covariates $\mathbf{X}$ (see the DAG in Figure 6a). In such cases, we can include the covariates in equation (6) when fitting the hIRT model. When the functional form for $\mathbb{E}[\Theta|D, \mathbf{X}]$ is correctly specified, the hIRT model will still provide unbiased estiamtes of the LTE. To mitigate potential biases due to model misspecification, we can also augment the hIRT model with matching or inverse probability weighting. For example, we can restrict our analysis to a matched sample where



(a) Covariates $\mathbf{X}$ Affect Both Treatment Status D and Latent Outcome Θ

(b) Covariates $\mathbf{X}$ Affect Both Latent Outcome Θ and Observed Indicators $\mathbf{Y}$

(c) Covariates $\mathbf{X}$ Affect Treatment Status D, Latent Outcome Θ, and Observed Indicators $\mathbf{Y}$

Figure 6: Extensions of the Unconfounded Treatment and Unconfounded Measurement Assumptions. D denotes treatment, Θ denotes the latent outcome of interest, and $Y_1, \ldots Y_J$ denote J observed indicators of Θ. Solid nodes are manifest and hollow nodes latent.

the treated and control units are sufficiently balanced in their values of observed covariates (Ho et al. 2007).

Second, our framework can be extended to account for observed confounders of the measurement device, i.e., the $\Theta - \mathbf{Y}$ relationship. Consider our first running example, where male and female respondents might answer the questions about procedural legitimacy differently.

Then measurement equivalence for male and female respondents would be violated. At the same time, it is likely that male and female respondents hold different latent procedural legitimacy beliefs about the process. This possibility is visualized in the DAG in Figure 6b, where we observe these pretreatment confounders of the measurement device. In this case, we can modify our Assumption 3 to $\mathbf{Y}(\theta) \perp\!\!\!\perp \Theta | D, \mathbf{X}$, which allows for DIF with respect to $\mathbf{X}$. Then, the item characteristic function in our hIRT model can be adapted to include $\mathbf{X}$ as predictors. We illustrate this extension in SM F.1.3 using data from the first study. In this particular case, we found no signs of violation of the unconfounded measurement assumption with respect to gender.

Finally, our framework can also be adapted to accommodate settings where observed covariates $\mathbf{X}$ may confound both the treatment-outcome relationship and the measurement device, as shown by the DAG in Figure 6c. This might occur, for example, in observational studies where item response patterns vary systematically by gender and race, which might also confound treatment assignment. In such cases, we can modify both Assumption 2 and Assumption 3 by conditioning on these covariates. Accordingly, these covariates need to be included as predictors in both equations (5) and (6) when fitting the hIRT model.

7. Concluding Remarks

Latent outcomes abound in the social sciences. Experimental social scientists often put considerable effort into designing treatments, and make sure that the experiment is implemented under optimal conditions. Less attention is devoted to the identification and estimation of causal effects on latent outcomes. In this article, we have developed a new framework for conducting causal inference on latent outcomes. We clarify the identification assumptions and propose a model-based estimator of the latent treatment effect. Two applications highlight the usefulness of our approach in political science. While we have focused on experimental settings, the proposed methodology can also find applications in observational studies, where we can adjust for confounders either within the hIRT model or using matching/weighting methods.

To be sure, our proposed framework is not without limitations. First, the hIRT model presumes that the indicators are reflective of the latent variable. While we have discussed the benefits of this perspective for studying attitudes, beliefs, and other latent outcomes such as political knowledge, a formative approach to measurement may be more appropriate for studying outcomes that defy a consensual conceptualization. For example, when studying political participation, different researchers may use different sets of activities to form an index of participation of their own choosing. In such cases, a two-step approach with a formative indicator model may still be preferred.

Second, our proposed estimation strategy rests on a set of parametric assumptions. For example, for ordinal items, the default choice of the item characteristic function (equation 5) will be a graded response model, which imposes a proportional odds assumption for cumulative logits. Moreover, the hIRT model assumes that the latent outcome follows a normal distribution conditional on its predictors. When these assumptions are violated, our estimates of the LTE can be biased. In future work, we plan to relax these assumptions by integrating our framework with recent advances in nonparametric IRT estimation (Duck-Mayr et al. 2020). In future research, it would be valuable to also explore the development of measurement models that enable separate estimation of latent outcomes on the same scale for treatment and control groups. This approach could allow for a certain level of measurement inequivalence and potentially provide unbiased estimates of the LTE even in cases where the outlined measurement assumptions are violated.

To conclude, we provide a few suggestions for practitioners. First, to realize the full potential of the reflective approach, it is best to use multiple indicators of the latent construct. With a single indicator, it is still possible to envision a DAG with an arrow from the treatment to the latent outcome, and an arrow from the latent outcome to the indicator. However, in this case, it is hard to distinguish between the latent outcome and the observed indicator, and impossible to test potential violations of the exclusion restriction. Moreover, the estimated coefficients of hIRT models will be more precise with a larger number of items (see SM E.5).

Second, the unconfounded measurement assumption highlights the importance of valid measurement device. Well-tested items make sure that the responses are governed only by the latent variable of interest and do not reflect multiple latent dimensions or influences of confounding variables. Increasing the number of reliable indicators can also increase the efficiency of the LTE estimation (see SM E.5). In this regard, scholars should invest more resources in the development of subject-specific item pools. Finally, it is crucial to prevent treatment from intervening in the measurement device. Considering this should be part of the research design. Whenever possible, pre-tests should be used to check for potential DIF by treatment status under different treatment protocols so that only those protocols that do not induce DIF are to be used in the main study.

References

Aaby, David, and Juned Siddique. 2021. “Effects of Differential Measurement Error in Self-Reported Diet in Longitudinal Lifestyle Intervention Studies.” *International Journal of Behavioral Nutrition and Physical Activity* 18(125).

Banerjee, Souvik, and Anirban Basu. 2021. “Estimating Endogenous Treatment Effects Using Latent Factor Models with and without Instrumental Variables.” *Econometrics* 9(14).

Battistin, Erich, and Andrew Chesher. 2014. “Treatment Effect Estimation with Covariate Measurement Error.” *Journal of Econometrics* 178(2): 707–715.

Bollen, Kenneth A. 1989. *Structural Equations with latent Variables.* Wiley.

Bollen, Kenneth A. 2002. “Latent variables in psychology and the social sciences.” *Annual review of psychology* 53(1): 605–634.

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap van Heerden. 2003. “The Theoretical Status of Latent Variables.” *Psychological Review* 110(2): 203–219.

Carroll, Raymond J, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall/CRC.

Clayton, Amanda, Diana Z O’Brien, and Jennifer M Piscopo. 2019. “All male panels? Representation and democratic legitimacy.” *American Journal of Political Science* 63(1): 113–129.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. “Information Equivalence in Survey Experiments.” *Political Analysis* 26: 399–416.

Duck-Mayr, JBrandon, Roman Garnett, and Jacob Montgomery. 2020. Gpirt: A Gaussian Process Model for Item Response Theory. In *Conference on uncertainty in artificial intelligence.* PMLR pp. 520–529.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. “How to make causal inferences using texts.” *Science Advances* 8(42): eabg2652.

Eid, Michael, Christian Geiser, and Tobias Koch. 2016. “Measuring Method Effects: From Traditional to Design-Oriented Approaches.” *Current Directions in Psychological Science* 25(4): 275–280.

Fong, Christian, and Justin Grimmer. 2021. “Causal inference in latent treatments.”.

Grimm, Pamela. 2010. “Social desirability bias.” *Wiley international encyclopedia of marketing* .

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(dec): 945–960.

Huddy, Leonie, Joshua Billig, John Bracciodieta, Lois Hoeffler, Patrick J Moynihan, and Patricia Pugliani. 1997. "The effect of interviewer gender on the survey response." *Political Behavior* 19(3): 197–220.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological Methods* 15(4): 309.

Kalla, Joshua L, and David E Broockman. 2020. "Reducing exclusionary attitudes through interpersonal conversation: evidence from three field experiments." *American Political Science Review* pp. 1–16.

Knox, Dean, Christopher Lucas, and Wendy K Tam Cho. 2022. "Testing Causal Theories with Learned Proxies." *Annual Review of Political Science* 25.

Masyn, Katherine E. 2017. "Measurement Invariance and Differential Item Functioning in Latent Class Analysis With Stepwise Multiple Indicator Multiple Cause Modeling." *Structural Equation Modeling* 24(2): 180–197.

Mayer, Axel. 2019. "Causal Effects Based on Latent Variable Models." *Methodology* 15(Suppl.): 15–28.

Meredith, William. 1993. "Measurement Invariance, Factor Analysis and Factorial Invariance." *Psychometrika* 58(dec): 525–543.

Millimet, Daniel L. 2010. The Elephant in the Corner: A Cautionary Tale about Measurement Error in Treatment Effects Models. Technical Report IZA Discussion Paper 5140 Institute for the Study of Labor Bonn: .

Mummolo, Jonathan, and Erik Peterson. 2019. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review* 113(2): 517–529.

Osterlind, Steven J, and Howard T Everson. 2009. *Differential item functioning.* Sage Publications.

Pearl, Judea. 2010. On Measurement Bias in Causal Inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intellifence*, ed. Peter Grunwald, and Peter Spirtes. Catalina Island, CA: .

Rabbitt, Matthew P. 2018. "Causal inference with latent variables from the Rasch model as outcomes." *Measurement* 120: 193–205.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66(5): 688–701.

VanderWeele, Tyler J. 2022. “Constructed Measures and Causal Inference: Towards a New Model of Measurement for Psychosocial Constructs.” *Epidemiology* 33(1): 141–151.

VanderWeele, Tyler J, and Stijn Vansteelandt. 2022. “A statistical test to reject the structural interpretation of a latent factor model.” *Journal of the Royal Statistical Society Series B* 84(5): 2032–2054.

Vermunt, Jeroen K., and Jay Magidson. 2021. “How to Perform Three-Step Latent Class Analysis in the Presence of Measurement Non-Invariance or Differential Item Functioning.” *Structural Equation Modeling* 28(3): 356–364.

Zhou, Xiang. 2019. “Hierarchical Item Response Models for Analyzing Public Opinion.” *Political Analysis* pp. 1–22.

Supplementary Material

Causal Inference with Latent Outcomes

Lukas F. Stoetzer, Xiang Zhou, and Marco Steenbergen

Table of Contents

A. Latent Outcomes in Political Science

A.1. Hand-coding of Latent Outcomes

We hand-coded 1,630 abstracts of articles published in APSR, AJPS, and JOP from 2015 to mid 2021 to obtain an overview of articles that focus on latent outcomes in political science. The articles are classified into different categories based on the abstracts: Does the article contain a quantitative analysis? An experiment? Observational data? Does it contain at least one latent outcome? We also saved the description of latent outcome from abstract.

We trained a Master's student in coding abstracts by first explaining our definition of latent outcomes. Our informal definition considers latent variables as theoretical constructs that cannot be directly observed and are therefore often inferred from a set of observable indicators. To help the student apply this definition, we asked the student to read our working paper and a review article (Bollen 2002) and then we discussed the relevance of our definition to political science research. During the training, we used a set of abstracts from published works to illustrate the definition of latent variables and their relationships to various concepts such as attitudes, opinions, preferences, and values, political behavior, policy outcomes, voting decisions, and institutions. The training also helped the student to recognize the primary outcome and the research method of a study from its abstract. After a training period of 50 abstracts, we discussed cases that the student considered uncertain.

Figure SM1 shows the proportion of quantitative articles that analyze latent outcomes for experimental and non-experimental research designs. Around one third of quantitative research uses latent variables. The share of latent outcomes is higher among experimental research designs.

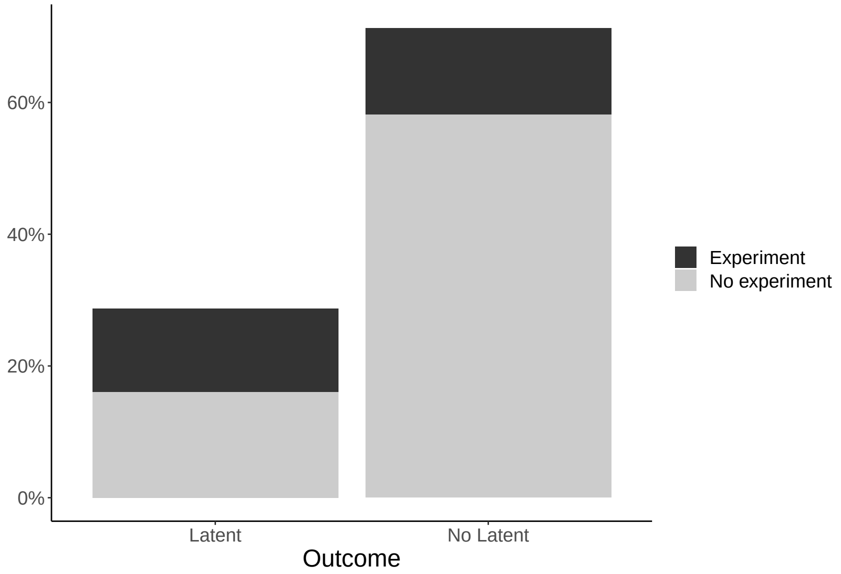
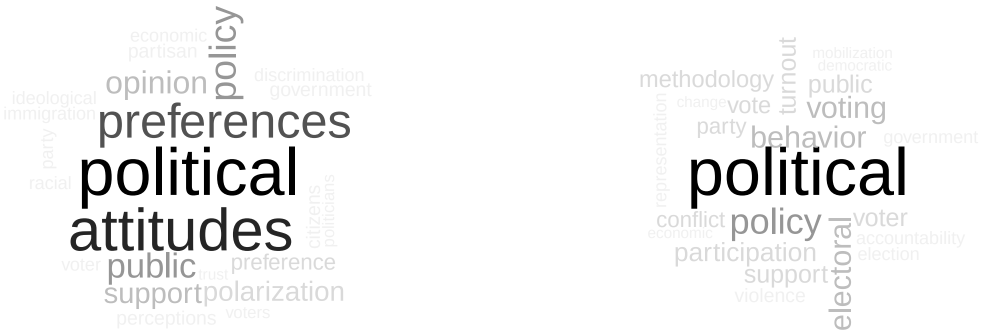


Figure SM1: Proportion of articles that analyze latent outcomes in APSR, AJPS and JOP from 2015 to mid 2021

When we compare the terminology used to describe the outcome variable using a word-cloud, we observe references to ”attitudes”, ”preferences”, ”support”, and so on in those that analyze latent outcomes. This follows our definition, as these constructs are not directly observable and require observable indicators. Outcomes that are classified as not latent tend to be behavioral, like ”voting”, ”turnout”, and ”participation”, or outputs of policymaking (e.g. ”policy”). This again aligns with our definition, as these concepts can in principle be observed. See Figure SM2.



(a) With Latent outcomes

(b) Without Latent Outcomes

Figure SM2: Words used in abstracts to describe the outcome/dependent variable in quantitative research articles in APSR, AJPS, and JOP that we hand-coded as having latent and not having latent outcomes.

A.2. Some Examples of Latent Outcomes

Table SM1 provides some examples that illustrate the prevalence of our perspective on latent outcomes in political science research. Apart from our two running examples, it includes six additional examples that focus on immigration attitudes, support for redistribution, and political knowledge, all of which are common latent outcomes studied in the political science.

All of these studies are concerned with a theoretical outcome that is not directly observable. For instance, the two studies on support for redistribution examine the idea that rich people in more unequal regions in Western Europe are more supportive of redistribution due to their concern with crime (Rueda and Stegmueller 2016), and the idea that low trust in government depresses support for redistribution (Peyton 2020). These are general arguments about the concept of ”support for redistribution” that exists irrespective of the measurement indicators. However, to test the arguments we need manifest indicators of this concept, which

Article	Latent Outcome	Indicators	Research Design
Examples Main Text			
Clayton et al. (2019)	Perceived democratic legitimacy	Multiple items	Survey exp.
Kalla and Broockman (2020)	Exclusionary attitude	Multiple items	Field exp.
Additional Examples			
Cavaille and Marshall (2019)	Immigration attitude	Multiple item	Natural exp.
Williamson et al. (2021)	Immigration attitude	Single item	Survey exp.
Peyton (2020)	Support for redistribution	Multiple items	Survey exp.
Rueda and Stegmueller (2016)	Support for redistribution	Single item	Survey anal.
Pereira (2019)	Political knowledge	Multiple items	Survey anal. & exp.
Campbell and Niemi (2016)	Political knowledge	Multiple items	Survey anal.

Table SM1: Examples of studies of latent outcomes

are obtained via survey items.[18] Yet, the arguments do not depend on the specific survey items, making them good examples of the reflective approach (Borsboom et al. 2003).

Furthermore, different studies can use different indicators without changing the knowledge accumulation about the determinants of the latent outcome. For example, Williamson et al. (2021) use a single indicator to show that immigrant histories can lead to more favorable views of immigration,[19] while Cavaille and Marshall (2019) use a different set of survey items as indicators to show that education decreases anti-immigration attitudes later in life.[20] Both studies choose indicators that they consider reflective of the underlying attitude toward immigration. In principle, the reflective approach allows the effects of different studies to be compared in terms of the same latent outcome.

The same holds for the two studies that test hypothesis about the effects on political knowledge. Both articles use two set of indicators to measure political knowledge. Campbell

[18] Rueda and Stegmueller (2016) use the respondent's response to the statement "the government should take measures to reduce differences in income levels" measured on a 5 point scale to gauge support for redistribution. Peyton (2020) employs four survey items about support for spending on redistributive policies, e.g. food stamps, welfare, programs that assist minorities, and assistance to the homeless.

[19] The indicator is "Do you agree or disagree that the United States should limit the number of immigrants entering the country?" Scale ranges from 1 to 7, with 7 indicating support for more open immigration.

[20] They include classic survey items such as "To what extent do you think [country] should allow people of the same race or ethnic group as most [country] people to come and live here?"

and Niemi (2016) test two hypotheses about the effect of civic education requirements on political knowledge, using the National Assessment of Educational Progress (NAEP) test and a set of additional indicators in survey data to measure the latent outcome. Pereira (2019) assesses the argument that female political representation affects the levels of political knowledge. This paper draws on two studies, a study that employs cross-section public opinion surveys with varying knowledge items and a survey experiment that relies on six open-ended political knowledge questions. The reflective perspective makes it possible that the authors can use different set of indicators to evaluate their arguments about the same latent outcome (political knowledge).

It is further noteworthy that our conceptualization of latent outcomes applies to different common research designs for the identification of causal effects. For instance, Cavaille and Marshall (2019) use a natural experiment in the form of a regression discontinuity design to study the effects of compulsory schooling reforms, while Rueda and Stegmueller (2016) analyze data from the European Social Survey. While the bulk of our discussion focuses on experimental interventions, we highlight extensions of our framework to alternative research designs toward the end of the article.

Finally, it is important to note that not all studies fit perfectly in our conceptualization of latent outcomes. For example, behavioral outcomes such as voting decisions, participation in protests, and donations to charity are directly observable and do not exist independently of the behavior itself. The same holds for support for a specific policy, which, although not directly observable, is often operationalized as a decision in a referendum or as the answer to one particular survey item. For these types of outcomes, it is unnecessary to consider different indicators that are reflective of the behavior or the latent support, making the single measure equivalent to the outcome of interest. Standard tools for causal inference are more suitable for these types of behavioral outcomes.

Additionally, some latent outcomes are better captured with a formative perspective. For example, studies of political participation often define the degree of a citizen's participation as the sum of different activities (Van Deth 2016). Depending on the definition, varying survey items about activities constitute a political participation index, making it a good example of a formative measurement model. Alternative frameworks should be used to study the determinants of formative latent outcomes.

B. Directed Acyclic Graphs for Empirical Illustrations

B.1. Descriptive Representation and Democratic legitimacy

The measurement indicators for substantive legitimacy are:

- The committee made the right decision for all the state's citizens.
- The committee made the right decision for women.
- How fair was this decision to women?

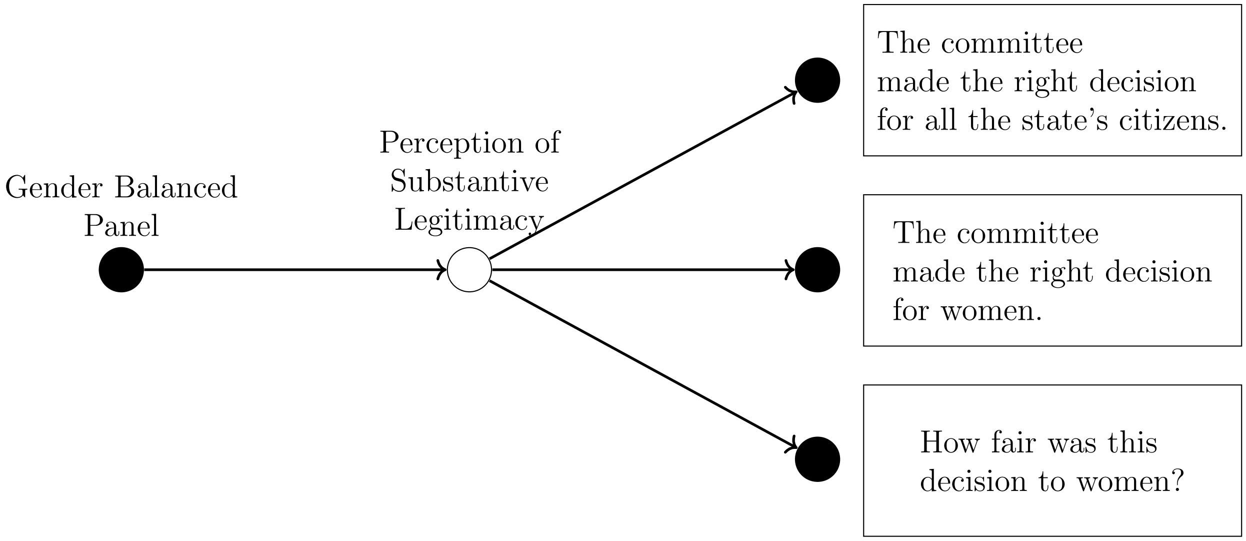

Gender Balanced Panel
Perception of Substantive Legitimacy
The committee made the right decision for all the state's citizens.
The committee made the right decision for women.
How fair was this decision to women?

Figure SM3: Illustrative DAG for the Empirical Example from Clayton et al. (2019)

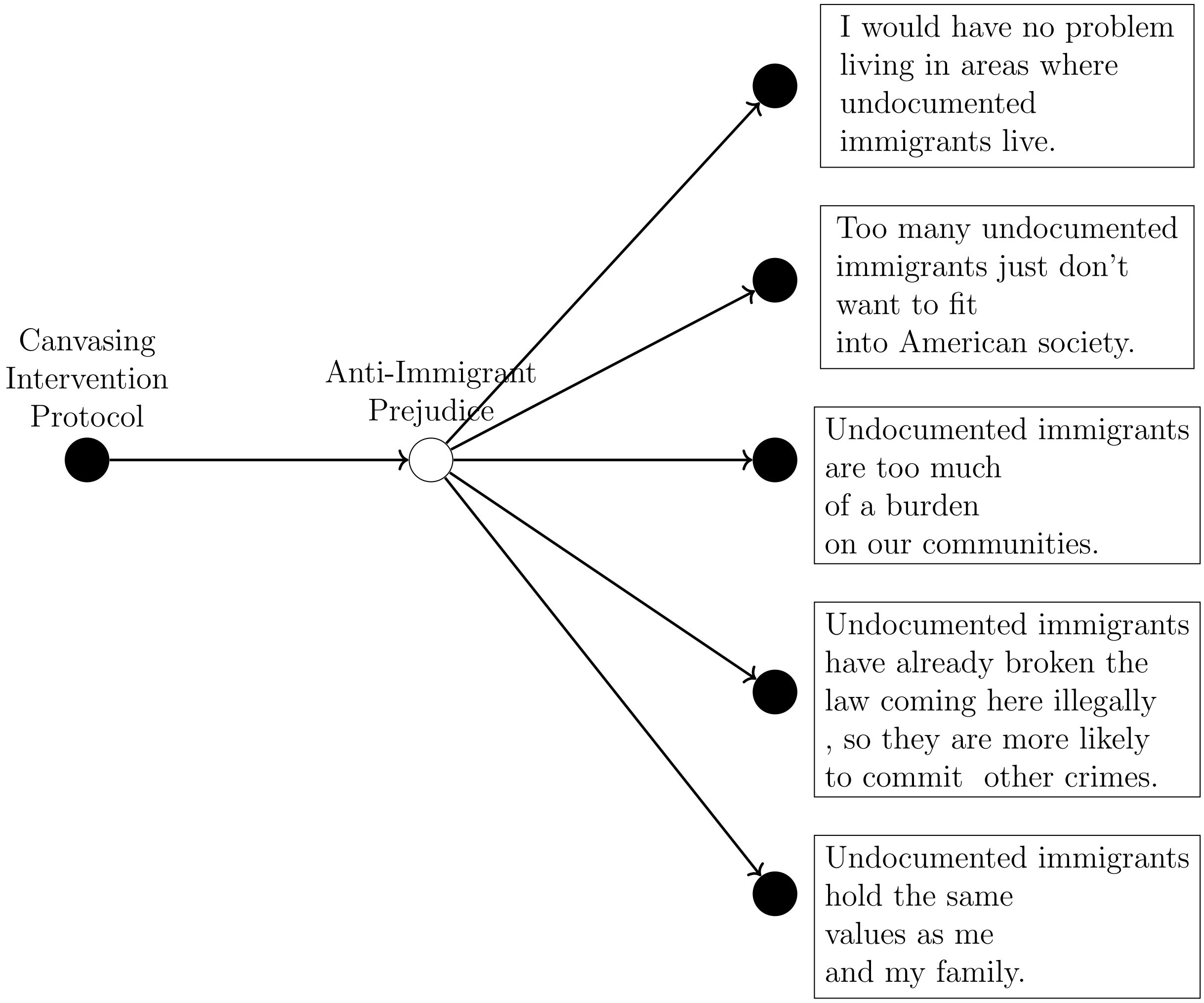
Which leads to the DAG in Figure SM3 for substantive legitimacy.

The measurement indicators for procedural legitimacy are:

- Thinking for a moment about the gender composition of the committee, how fair was the decision-making process?
- Thinking about the gender composition of the committee, the committee's decision should be overturned
- Thinking about the gender composition of the committee, the committee can be trusted to make decisions that are right for the state's citizens
- The State Legislature can be trusted to make decisions that are right for the state.

The DAG for procedural legitimacy is comparable to Figure SM3, exchanging the indicators.

B.2. Reducing exclusionary attitudes



Canvasing Intervention Protocol
Anti-Immigrant Prejudice
I would have no problem living in areas where undocumented immigrants live.
Too many undocumented immigrants just don't want to fit into American society.
Undocumented immigrants are too much of a burden on our communities.
Undocumented immigrants have already broken the law coming here illegally , so they are more likely to commit other crimes.
Undocumented immigrants hold the same values as me and my family.

Figure SM4: Illustrative DAG for the Empirical Example from Kalla and Broockman (2020)

C. The Hierarchical IRT Model

This section provides more details on the hierarchical IRT model and some illustrative code for the R implementation of the model.

C.1. Item characteristic function

The main text gives a generic form of the item characteristic function $P_{jh}(\theta, \alpha_{jh}, \beta_j)$ for item j and category h. This generic form encompasses item characteristic functions for binary indicators (Lord and Novick 1968) and ordinal indicators (Samejima 1969). H_j indicates the number of response categories for question j. In the case of binary indicators, the formulation reduces to a 2pl item characteristic function:

$$Pr(y_j \geq 1) = \frac{exp(\alpha_j + \beta_j \theta)}{1 + exp(\alpha_j + \beta_j \theta)},$$

With ordinal responses, this formulation leads to the graded response model (Samejima 1969):

$$P_{jh}(\theta) = Pr(y_j \geq h) - Pr(y_j \geq h+1)$$

$$= \frac{exp(\alpha_{jh} + \beta_j\theta)}{1 + exp(\alpha_{jh} + \beta_j\theta)} - \frac{exp(\alpha_{jh+1} + \beta_j\theta)}{1 + exp(\alpha_{jh+1} + \beta_j\theta)}, \qquad h = 0, 1, 2, ..., H_j - 1,$$

where $\infty = \alpha_{j0} > \alpha_{j1} ... > \alpha_{jH_j - 1} > \alpha_{jH_j} = -\infty$.

C.2. Estimation

The hierarchical model can be estimated using marginal maximum Likelihood. Zhou (2019) describes the procedure using the EM algorithm in detail and the R-package *hIRT* implements it. The following is a restatement of the EM estimation algorithm for the model.

- Define the following shorthand:

$$\begin{aligned}
&\boldsymbol{\alpha} = \{\alpha_{jh}; 1 \leq j \leq J, 0 \leq h \leq H_j - 1\}, \quad \boldsymbol{\alpha_j} = \{\alpha_{jh}; 0 \leq h \leq H_j - 1\}, \\
&\boldsymbol{\beta} = \{\beta_j; 1 \leq j \leq J\} \\
&\boldsymbol{\gamma} = \{\gamma_0, \gamma_1\}, \\
&\boldsymbol{\theta} = \{\theta_i; 1 \leq i \leq N\}, \quad \boldsymbol{x} = \{x_i; 1 \leq i \leq N\}, \\
&\boldsymbol{y} = \{y_{ij}; 1 \leq i \leq N, 1 \leq j \leq J\}.
\end{aligned}$$

- The complete data likelihood for the model is:

$$\begin{aligned}
p(\boldsymbol{y}, \boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) \\
&= \prod_{i=1}^{N} \left\{ \prod_{j=1}^{J} p\left(y_{ij} \mid \theta_i, \boldsymbol{\alpha}_j, \beta_j\right) \right\} p\left(\theta_i \mid \boldsymbol{\gamma}\right).
\end{aligned}$$

- The EM-algorithm treats θ as missing data and $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*$ as a set of existing parameter estimates.

- The Q-function of the EM algorithm, i.e., the conditional expectation of the log complete data likelihood, is

$$\begin{aligned}
Q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \mathbb{E}\left[\log p(\boldsymbol{y}, \boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \mid \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{y}\right] \\
&= \int_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N} \left[\sum_{j=1}^{J} \log p\left(y_{ij} \mid \theta_i, \boldsymbol{\alpha}_j, \beta_j\right) + \log p\left(\theta_i \mid \boldsymbol{\gamma}\right) \right] \right\} p\left(\boldsymbol{\theta} \mid \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{y}\right) d\boldsymbol{\theta} \\
&= \sum_{i=1}^{N} \int_{\theta_i} \left[\sum_{j=1}^{J} \log p\left(y_{ij} \mid \theta_i, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right) + \log p\left(\theta_i \mid \boldsymbol{\gamma}\right) \right] p\left(\theta_i \mid \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{y}_i\right) d\theta_i
\end{aligned}$$

The integral can then be evaluated using quadrature methods, which gives an approximation of:

$$\begin{aligned}
Q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &\approx \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \left[\sum_{j=1}^{J} \log p\left(y_{ij} \mid \theta^k, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right) + \log p\left(\theta^k \mid \boldsymbol{\gamma}, \boldsymbol{x}_i\right) \right] \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{h=0}^{H_j-1} w_{ik} \mathbf{1}\left(y_{ij} = h\right) \log P_{jh}\left(\theta^k\right) + \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \log p\left(\theta^k \mid \boldsymbol{\gamma}, \boldsymbol{x}_i\right) \\
&= \sum_{j=1}^{J} \left[\sum_{k=1}^{K} \sum_{h=0}^{H_j-1} f_k^{jh} \log P_{jh}\left(\theta^k\right) \right] + \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \log p\left(\theta^k \mid \boldsymbol{\gamma}, \boldsymbol{x}_i\right)
\end{aligned}$$

where $f_k^{jh} = \sum_{i=1}^{N} w_{ik} \mathbf{1}\left(y_{ij} = h\right)$. and the weights:

$$w_{ik} = \frac{w_k \left[\prod_{j=1}^{J} p\left(y_{ij} \mid \boldsymbol{\alpha}_j^*, \boldsymbol{\beta}_j^*, \theta^k\right)\right] p\left(\theta^k \mid \boldsymbol{\gamma}^*, \boldsymbol{x}_i\right)}{\sum_{k=1}^{K} w_k \left[\prod_{j=1}^{J} p\left(y_{ij} \mid \boldsymbol{\alpha}_j^*, \boldsymbol{\beta}_j^*, \theta^k\right)\right] p\left(\theta^k \mid \gamma^*, \boldsymbol{x}_i\right)}$$

- The M-step of the EM algorithm are the following optimization problems:

  $$\operatorname{argmax}_{\alpha_j}, \beta_j \sum_{k=1}^{K} \sum_{h=0}^{H_j-1} f_k^{jh} \log P_{jh}\left(\theta^k\right) \text{ for all } j, \quad \text{and} \quad \operatorname{argmax}_{\gamma} \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \log p\left(\theta^k \mid \gamma, \boldsymbol{x}_i\right)$$

  The first optimization problem is equivalent to fitting J separate generalized linear models (either logit/probit or proportional odds models) — one for each item to the "pseudo data" f_k^{jh}. The second optimization problem is a linear regression model with the weights w_{ik}.

- Upon convergence, the EM algorithm gives final estimates $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$.

- The latent outcomes θ_i can be estimated using empirical Bayes inference. For example, we can directly use the final posterior means, giving the expected a posterior (EAP) estimates

  $$\hat{\theta}_i = \mathbb{E}\left(\theta_i \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \boldsymbol{y}\right) = \sum_{k=1}^{K} w_{ik} \theta^k$$

- Inference for the key parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ can be conducted using the asymptotic variance-covariance matrix $\hat{I}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, which can be estimated using either the Hessian matrix or the outer product of the gradients of the log marginal likelihood.

C.3. Implementation in R

This section provides some example code on how the model can be estimated in R. The hIRT model can be estimated using the function *hgrm* of the R-package *hIRT* (Zhou 2019).

The code also contains a description how the estimates from the *hgrm* function can be standardized. For standardization, the total standard deviation of the latent traits should account for estimation uncertainty (see code line 22).

```
 1 # load libraries
 2   library(hIRT)
 3   library(tidyverse)
 4 
 5 # Read in data
 6   data <- readRDS(file=”data.Rdata”)
 7 
 8 # Select items from data
 9   items <-  data %>% dplyr::select(starts_with(”item”))
10 
11 # Prepare Model Matrix: y = gamma_0 + gamma_1 * treatment
12   mdl <- model.matrix(~ treatment,data = data)
13 
14 
15 # Estimation HGRM
16   res <- hgrm(y = items,x=mdl)
17 
18 # Summary
19   summary(res)
20 
21 # Standardized LTE
22   total_sd <- sqrt(var(res$scores$prior_mean) + mean(res$scores$prior_
       sd^2))
23 
24   coef_hIRT <- coef_mean(res) %>%
25        mutate(est = ‘Estimate‘/total_sd,
26                se = ‘Std_Error‘/total_sd) %>%
27        select(est, se)
```

The package also includes the function hgrmDIF to allow for DIF. Comparing model fit using likelihood-ratio tests permits us to test for violations of the exclusion restriction with respect to individual items.

```
# Test for DIF (for all items separately)

# Estimation Uniform models
resDIFu <- map(1:ncol(items), ~ hgrmDIF(items, x,x0=x, form_dif = ”
    uniform”, items_dif = .x)

# Estimation Non uniform models
resDIFn <- map(1:ncol(items), ~ hgrmDIF(items, x,x0=x, form_dif = ”non-
    uniform”, items_dif = .x)

# Uniform DIF test
DIFu_test <- map(resDIFu, ~ lrtest2(res, .x))

# Nonuniform DIF test
```

```
DIFn_test <- map2(resDIFu, resDIFn,  ~ lrtest2(.x, .y))

# Overall DIF test
DIFo_test <- map(resDIFn, ~ lrtest2(res, .x))
```

The above code requires lrtest2 function to conduct a log-likelihood ratio test. For completeness, here is our implementation of it:

```
# Function log-likelihood ratio test
lrtest2 <- function(m0, m1){

  df <- nrow(m1$coefficients) - nrow(m0$coefficients)
  llr <- m1$log_Lik - m0$log_log_Lik

  rval <- matrix(rep(NA, 10), ncol = 5)
  colnames(rval) <- c(”#Df”, ”LogLik”, ”Df”,
                      ”Chisq”, ”Pr(>Chisq)”)
  rownames(rval) <- 1:2

  rval[, 1] <- c(nrow(m0$coefficients), nrow(m1$coefficients))
  rval[, 2] <- c(m0$log_Lik, m1$log_Lik)
  rval[2, 3] <- rval[2, 1] - rval[1, 1]
  rval[2, 4] <- 2 * abs(rval[2, 2] - rval[1, 2])
  rval[, 5] <- pchisq(rval[, 4], round(abs(rval[, 3])), lower.tail =
    FALSE)

  title <- ”Likelihood ratio test\n”
  topnote <- paste(”Model ”, format(1:2), ”: ”,
                   c(m0$call, m1$call), sep = ””, collapse = ”\n”)
  structure(as.data.frame(rval), heading = c(title, topnote),
            class = c(”anova”, ”data.frame”))
}
```

D. Alternative DIF Tests using Nested Logistic Regressions

An alternative approach to testing DIF is to use nested (binary/ordered) logistic regressions, where we fit a logistic model for each of the manifest indicators as a function of our estimated latent outcomes and see if adding treatment status D as an additional predictor improves the fit (Swaminathan and Rogers 1990; French and Miller 1996; Miller and Spray 1993). Specifically, following Choi et al. (2011), we can evaluate different forms of DIF by estimating three (binary/ordered) logistic regression models for each item: Model 1) includes the estimated Θ as the only predictor, Model 2) includes treatment status D in addition to the estimated Θ, and Model 3) includes an additional interaction effect between D and the estimated Θ. Likelihood ratio tests can then be used to evaluate different types of violations. Specifically, a uniform DIF is likely to exist if model 2 fits better than model 1, and a non-uniform DIF is likely to exist if model 3 fits better than model 2. We can also conduct an omnibus test by comparing model 3 with model 1 directly.

E. Additional Results for Simulations

E.1. Root Mean Square Error for Different Methods

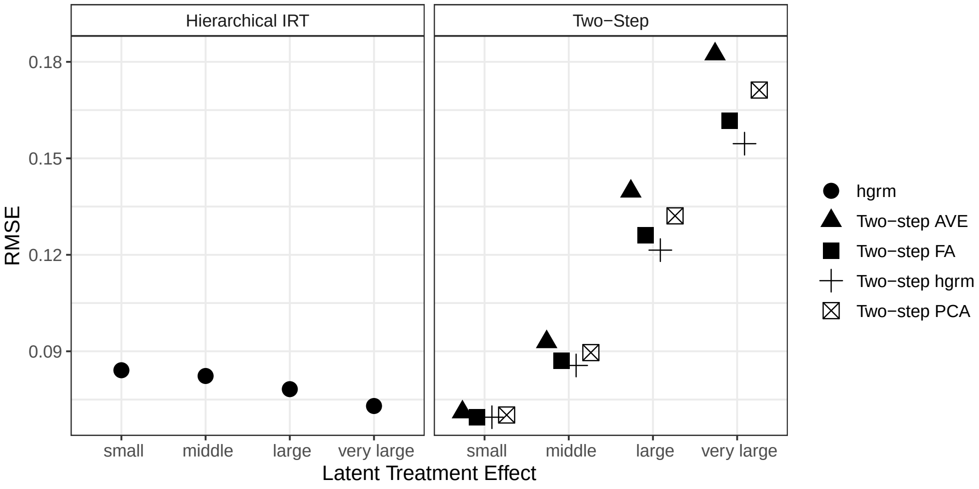


Figure SM5: Simulation results for the RMSE of different estimation methods, without differential item functioning.

E.2. Coverage and Standard Errors for Different Estimation Methods

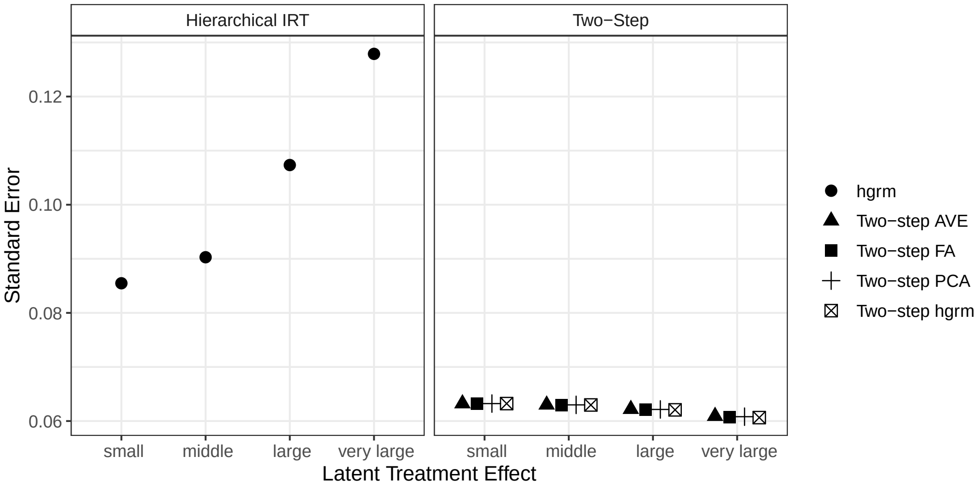


Figure SM6: Simulation results for the average standard errors for different estimation methods.

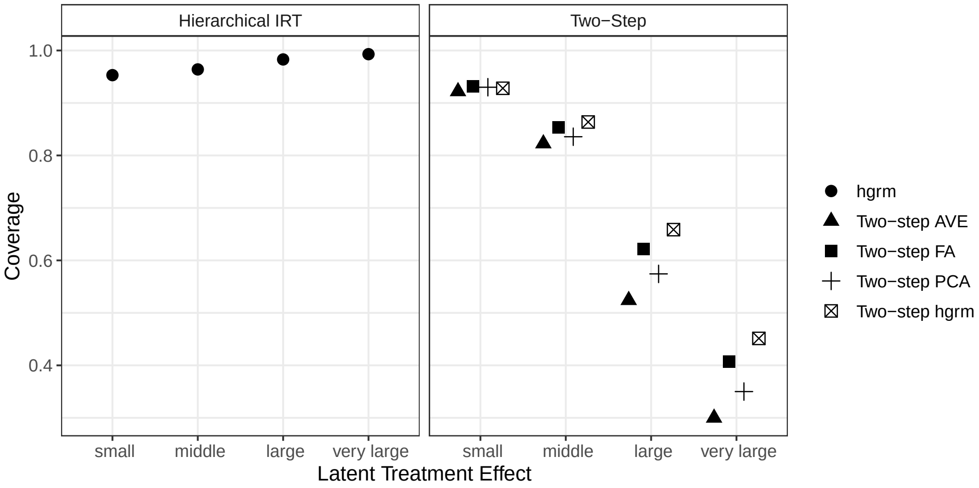


Figure SM7: Simulation results for the coverage of the confidence intervals for different estimation methods.

E.3. Consequences of Exclusion Restriction Violations

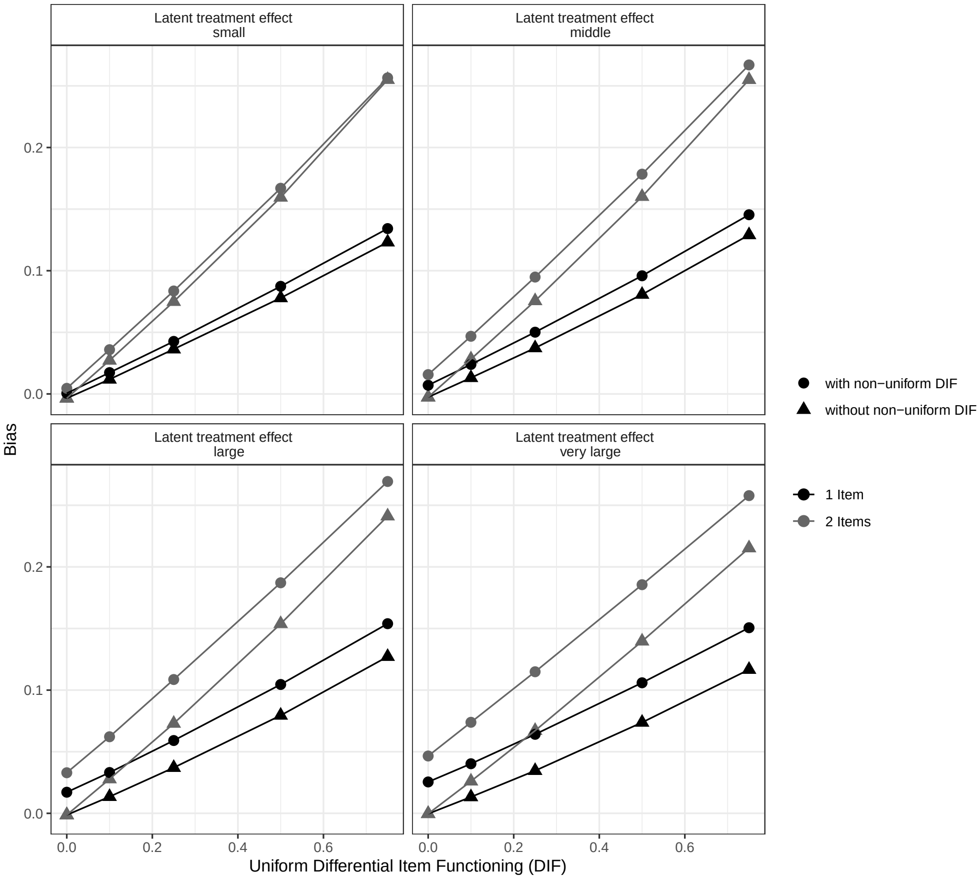


Figure SM8: Simulation results on the consequences of exclusion restriction violations.

E.4. Evaluation of DIF

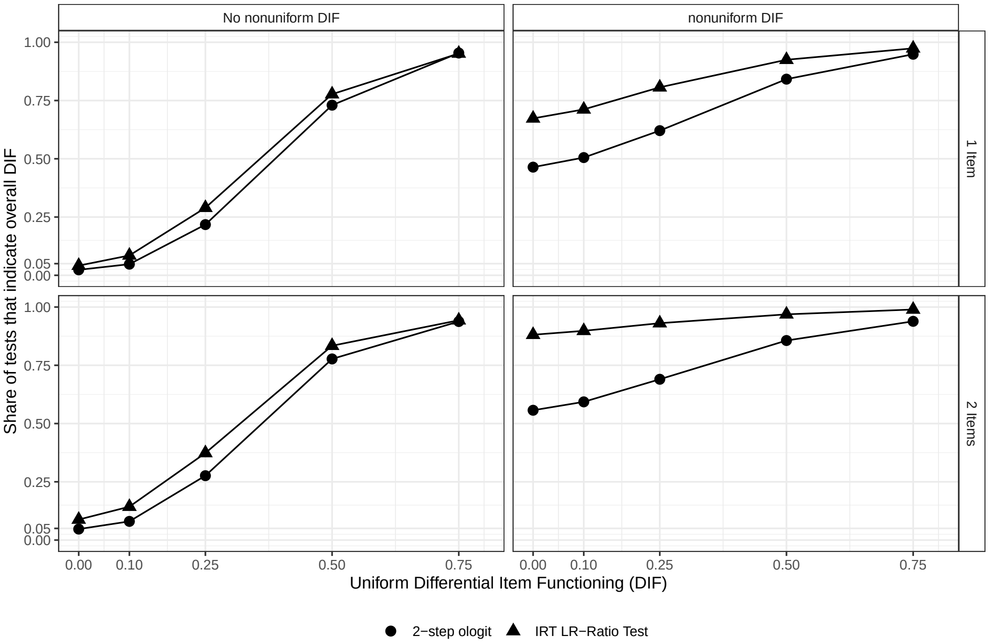


Figure SM9: Evaluation of overall differential item functioning (DIF) for simulated data. The figures show the share of overall DIF positive tests (y-axis) for varying levels of uniform DIF (x-axis), the cases of no non-uniform DIF and non-uniform DIF (columns), and the cases of one and two affected items (rows).

E.5. Bias and RMSE for different numbers of items and item categories

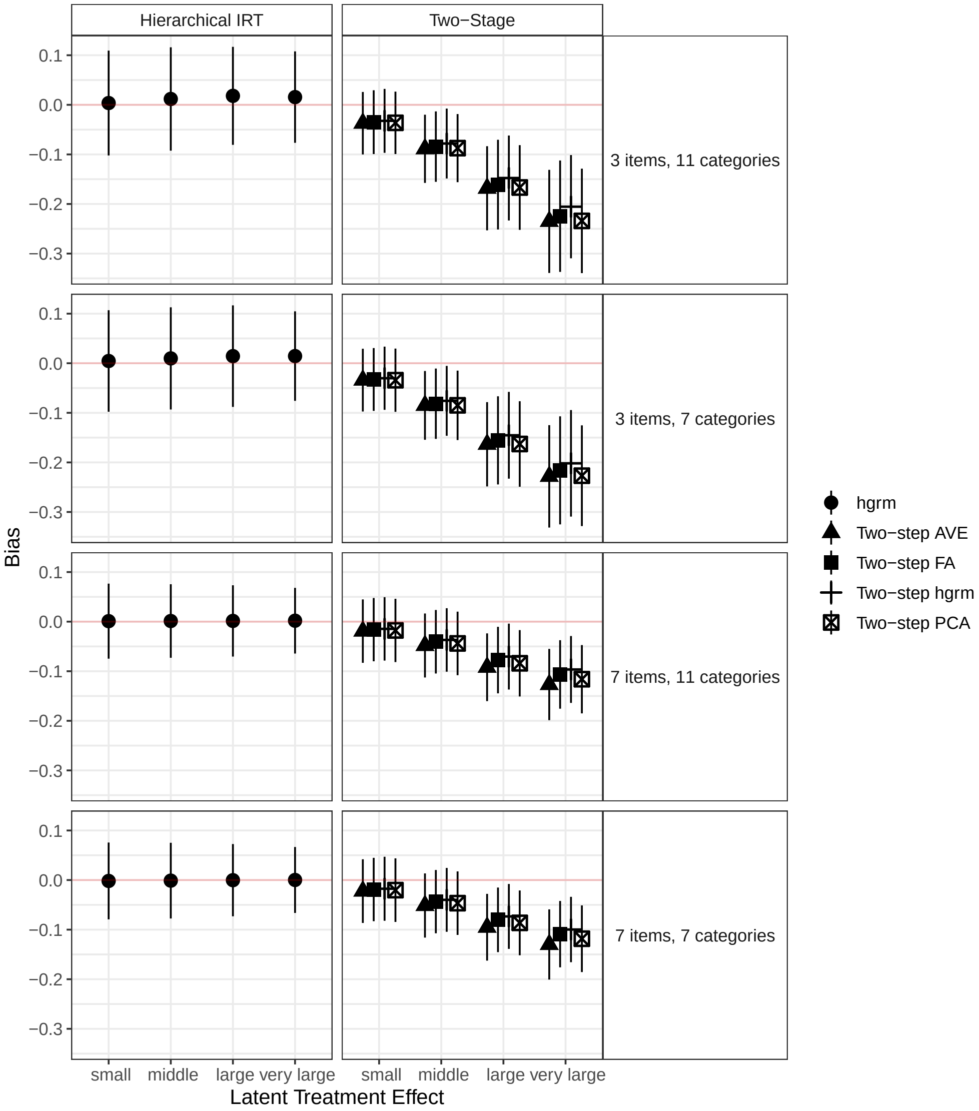


Figure SM10: Simulation results for the bias of different estimation methods for different numbers of items and item categories, without differential item functioning. The figure shows the estimated bias and variability of $\widehat{\text{LTE}} - \text{LTE}$ across simulations using one standard deviation around the mean.

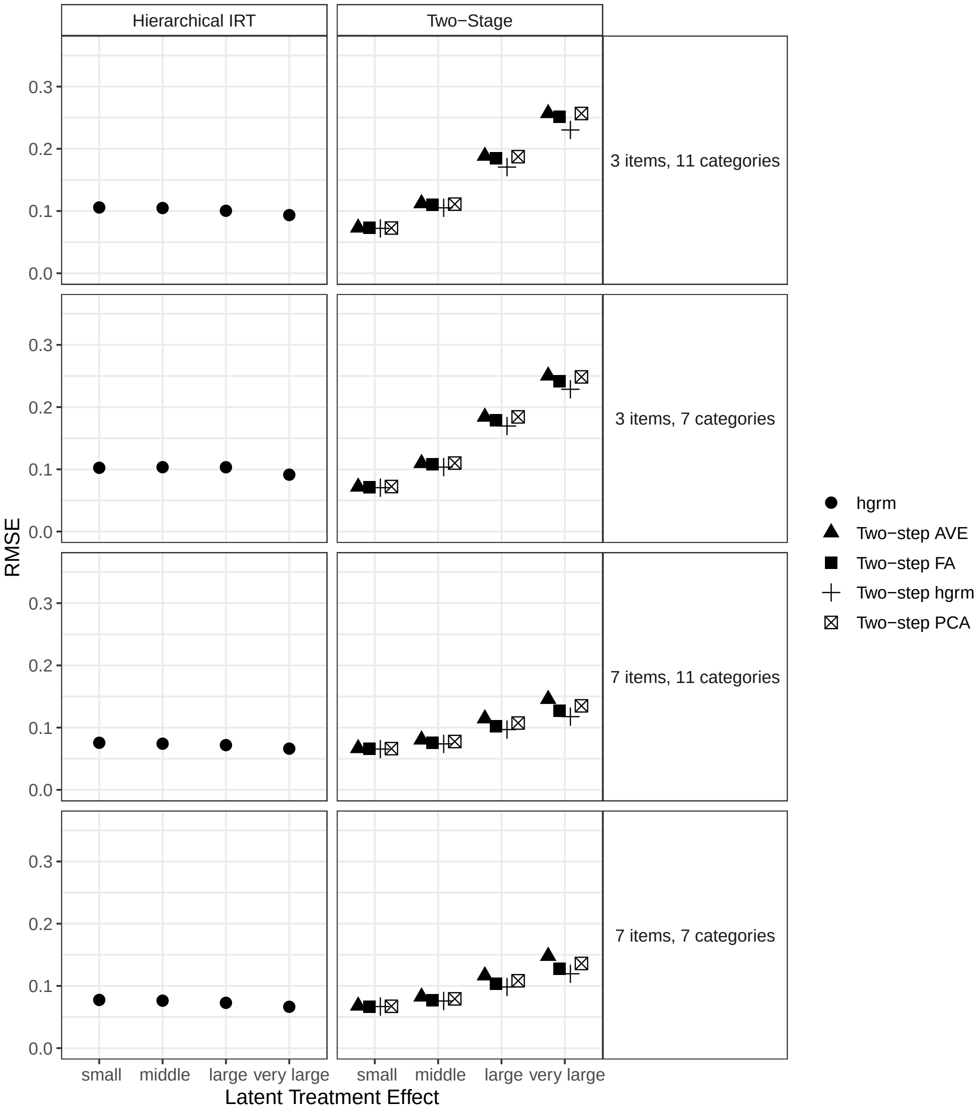


Figure SM11: Simulation results for the RMSE of different estimation methods for different numbers of items and answering categories, without differential item functioning.

E.6. Two-step approaches where the measurement model is fit using only control units

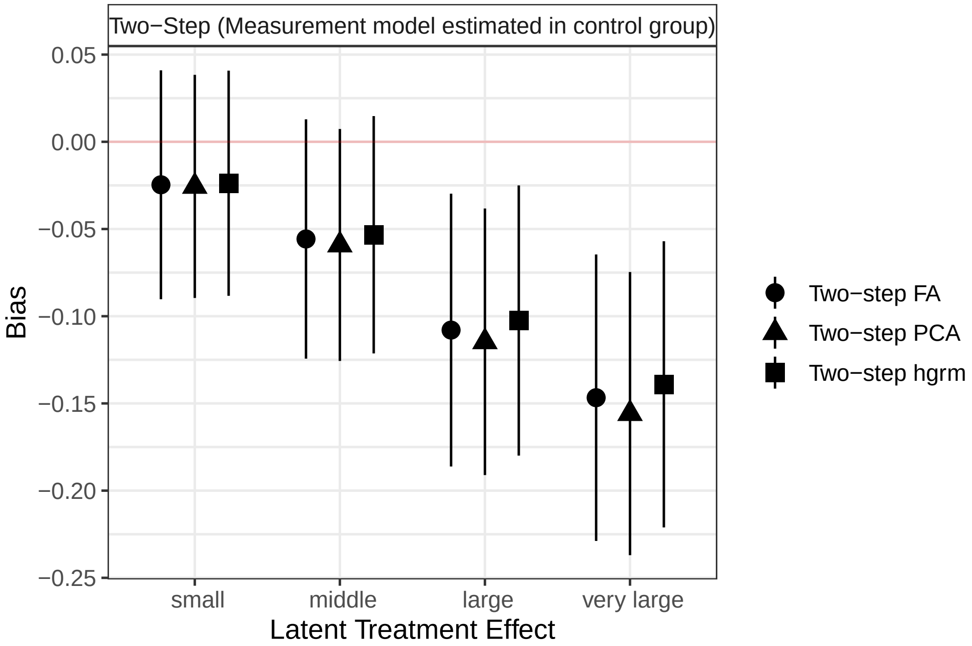


Figure SM12: Simulation results for the bias of different two-step estimation methods when the parameters of the measurement models are estimated only based on control group respondents, without differential item functioning.

F. Additional Results for Applications

F.1. Descriptive Representation and Democratic Legitimacy

F.1.1. Comparison with two-step estimates

Decision	hIRT		2-step	
Substantive Legitimacy				
Anti-Feminist	0.39	(0.07)	0.37	(0.06)
Feminist	0.1	(0.07)	0.09	(0.06)
Procedural Legitimacy				
Anti-Feminist	1.03	(0.09)	1.06	(0.07)
Feminist	0.59	(0.09)	0.51	(0.07)

Table SM2: Estimates of Latent Treatment Effects of All-Male panel versus a Gender-balanced panel for feminist and anti-feminist decisions from the hierarchical item response model (hIRT) and a two-step procedure based on factor scores (2-Step).

F.1.2. Conditional effects

We further replicate the conditional latent treatment effects for men and women on substantial legitimacy that are also reported in the main article. Clayton et al. (2019) argue that respondent gender should moderate the effect, as men depend more on cues about the substantial legitimacy of the decision. Figure SM13 shows the estimates and highlights subtle differences between results from different approaches. While the two-step approach finds a small difference in how men and women increase their perception of substantial legitimacy in the case of an anti-feminist decision (0.41 versus 0.32), the hIRT approach is more conservative about this difference. The effect estimates are closer (0.39 versus 0.36) and the uncertainty is larger. Tentative conclusions that point toward heterogeneous effects, hence, do not find support when the hIRT model is applied.

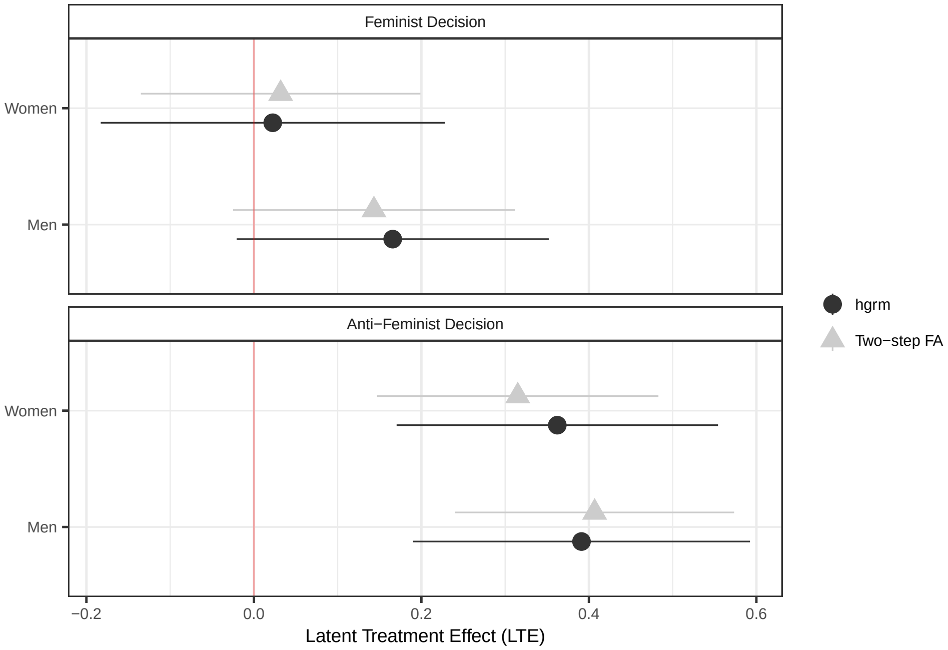


Figure SM13: Estimates of conditional latent treatment effects on substantial legitimacy of all-male versus a gender-balanced panel for men and women from the hierarchical item response model (hIRT) and a two-step procedure based on factor scores.

F.1.3. Measurement equivalence with respect to gender

In this section, we describe a test of the unconfounded measurement assumption concerning gender. As described in the main text, it could be the case the male and female respondents answer the items differently, while at the same time gender influences the latent outcomes. Our framework allows us to control for DIF concerning gender. We estimate the models again, once when controlling for DIF concerning gender and once without controlling.

Table SM3 shows that the estimates are almost identical when controlling for gender as part of the measurement. When comparing model fit using the likelihood-ratio test, we further find little support that the measurement is confounded by gender.

	Subs		Proc	
	Null	DIF Female	Null	DIF Female
(Intercept)	1.099	1.072	0.458	0.484
	(0.110)	(0.123)	(0.388)	(0.409)
AntiFem	-2.485	-2.485	-1.927	-1.925
	(0.186)	(0.218)	(0.155)	(0.155)
GBP	0.156	0.154	0.838	0.830
	(0.109)	(0.110)	(0.127)	(0.127)
GBP:AntiFem	0.434	0.437	0.622	0.624
	(0.155)	(0.156)	(0.163)	(0.163)
Female	-0.060	-0.005	-0.080	-0.129
	(0.076)	(0.405)	(0.077)	(0.103)
LR Test for overall DIF				
Chi-sq		11.042		10.704
p-val		0.09		0.1

Table SM3: Results when controlling for female as part of the measurement

F.2. Reducing exclusionary attitudes

F.2.1. Differential Item Functioning Tests for Latent Outcomes

Table SM5: Differential Item Functioning Tests for Latent Outcomes

item	Non-uniform	Overall	Uniform
Policy Index, 1 Week			
Item 1	4.69	7.74	3.05
Item 2	5.46	1.3	6.76
Item 3	0.37	1.12	0.75
Item 4	7.42	7.49	0.07
Item 5	2.53	1.02	1.51
Item 6	0.76	1.94	2.7
Policy Index, 1 Month			
Item 1	0.93	4.52	3.59
Item 2	11.35 *	15.31 *	3.96
Item 3	0.07	1.22	1.29
Item 4	11.38 *	11.48	0.1
Item 5	0.26	2.1	2.36
Item 6	16.34 **	15.66 *	0.68
Policy Index, 3-6 Month			
Item 1	0.88	0.57	1.45

Item 2	0.46	5.89	5.43
Item 3	1.29	2.55	1.26
Item 4	2.99	4.99	2
Item 5	0.26	1.89	1.63
Item 6	0.09	1.19	1.1
Prejudice Index, 1 Week			
Item 1	0.04	0.34	0.38
Item 2	10.59 *	11.3	0.71
Item 3	1.01	10.22	9.2
Item 4	3.8	6.13	2.33
Item 5	2.47	4.6	2.13
Prejudice Index, 1 Month			
Item 1	3.68	4.81	1.13
Item 2	3.38	7.44	4.06
Item 3	3.12	7.46	4.34
Item 4	7.43	9.47	2.03
Item 5	3.81	7.04	3.24
Prejudice Index, 3-6 Month			
Item 1	6.45	7.75	1.3
Item 2	5.45	5.34	0.11
Item 3	0.59	4.13	4.72
Item 4	1.92	3.88	1.96
Item 5	9.43 *	12.3	2.86
Combined Index, 1 Week			
Item 1	6.48	12.36	5.88
Item 10	7.75	8.3	0.55
Item 11	0.57	1.53	0.96
Item 2	3.06	6.91	3.85
Item 3	0.98	0.77	0.21
Item 4	6.25	6.42	0.17
Item 5	1.91	3.42	1.52
Item 6	0.76	4.33	5.09
Item 7	1.72	1.91	0.19
Item 8	4.38	4.26	0.11
Item 9	5.45	12.62	7.17
Combined Index, 1 Month			
Item 1	0.44	4.58	4.14
Item 10	1.86	3.46	1.6
Item 11	4.61	7.33	2.72
Item 2	13.14 *	15.41 *	2.27
Item 3	0.88	1.85	0.98
Item 4	4.04	4.31	0.27
Item 5	1.16	2.53	1.37

Item 6	14.9 **	14.89	0.01
Item 7	0.4	1.09	0.69
Item 8	0.45	2.46	2.01
Item 9	2.46	6.73	4.27
Combined Index, 3-6 Month			
Item 1	0.03	3.25	3.22
Item 10	0.06	0.81	0.75
Item 11	10.66	14.24	3.58
Item 2	1.76	4.83	3.07
Item 3	3.8	4.68	0.88
Item 4	3.31	4.4	1.08
Item 5	1.42	2.43	1.01
Item 6	0.32	1.28	0.96
Item 7	4.23	6.86	2.63
Item 8	2.17	3.03	0.86
Item 9	0.21	4.71	4.51

Time	Outcome	Uniform	Non-uniform	Overall
1 Week	Policy Index	None	None	None
1 Week	Prejudice Index	None	4 out of 5	None
1 Week	Combined Index	None	None	None
1 Month	Policy Index	None	3 out of 6	4 out of 6
1 Month	Prejudice Index	None	None	None
1 Month	Combined Index	None	9 out of 11	10 out of 11
3-6 Months	Policy Index	None	None	None
3-6 Months	Prejudice Index	None	4 out of 5	None
3-6 Months	Combined Index	None	None	None

Table SM4: Testing exclusion restriction using the differential item functioning test. The table reports the number of significant tests based on the χ^2-values of nested Likelihood-ratio tests.

F.2.2. Comparison with two-step estimates

Time	hIRT		2-step	
Anti-Immigrant Policy Index				
1 Week	0.09	(0.03)	0.10	(0.02)
1 Month	0.07	(0.03)	0.06	(0.03)
3-6 Months	0.07	(0.03)	0.08	(0.03)
Anti-Immigrant Prejudice Index				
1 Week	0.08	(0.04)	0.08	(0.03)
1 Month	0.07	(0.04)	0.05	(0.03)
3-6 Months	0.03	(0.04)	0.04	(0.03)
Combined Index				
1 Week	0.09	(0.03)	0.10	(0.02)
1 Month	0.07	(0.03)	0.06	(0.02)
3-6 Months	0.07	(0.03)	0.06	(0.03)

Table SM6: Estimates of latent treatment effects of the full protocol intervention for different outcome constructs from the hierarchical item response model (hIRT) and a two-step procedure based on factor scores (2-Step). Standard errors are reported in parentheses.

References

Bollen, Kenneth A. 2002. "Latent variables in psychology and the social sciences." *Annual review of psychology* 53(1): 605–634.

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap van Heerden. 2003. "The Theoretical Status of Latent Variables." *Psychological Review* 110(2): 203–219.

Campbell, David E, and Richard G Niemi. 2016. "Testing civics: State-level civic education requirements and political knowledge." *American Political Science Review* 110(3): 495–511.

Cavaille, Charlotte, and John Marshall. 2019. "Education and anti-immigration attitudes: Evidence from compulsory schooling reforms across Western Europe." *American Political Science Review* 113(1): 254–263.

Choi, Seung W, Laura E Gibbons, and Paul K Crane. 2011. "Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations." *Journal of statistical software* 39(8): 1.

Clayton, Amanda, Diana Z O'Brien, and Jennifer M Piscopo. 2019. "All male panels? Representation and democratic legitimacy." *American Journal of Political Science* 63(1): 113–129.

French, Ann W, and Timothy R Miller. 1996. "Logistic regression and its use in detecting differential item functioning in polytomous items." *Journal of Educational Measurement* 33(3): 315–332.

Kalla, Joshua L, and David E Broockman. 2020. "Reducing exclusionary attitudes through interpersonal conversation: evidence from three field experiments." *American Political Science Review* pp. 1–16.

Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Miller, Timothy R, and Judith A Spray. 1993. "Logistic discriminant function analysis for DIF identification of polytomously scored items." *Journal of Educational Measurement* 30(2): 107–122.

Pereira, Frederico Batista. 2019. "Gendered political contexts: The gender gap in political knowledge." *The Journal of Politics* 81(4): 1480–1493.

Peyton, Kyle. 2020. "Does trust in government increase support for redistribution? Evidence from randomized survey experiments." *American Political Science Review* 114(2): 596–602.

Rueda, David, and Daniel Stegmueller. 2016. "The externalities of inequality: Fear of crime and preferences for redistribution in Western Europe." *American Journal of Political Science* 60(2): 472–489.

Samejima, Fumiko. 1969. “Estimation of Latent Ability Using a Response Pattern of Graded Scores.” *Psychometrika* 34(mar): 1–97.

Swaminathan, Hariharan, and H Jane Rogers. 1990. “Detecting differential item functioning using logistic regression procedures.” *Journal of Educational measurement* 27(4): 361–370.

Van Deth, Jan W. 2016. “What is political participation?” In *Oxford research encyclopedia of politics*.

Williamson, Scott, Claire L Adida, Adeline Lo, Melina R Platas, Lauren Prather, and Seth H Werfel. 2021. “Family matters: How immigrant histories can promote inclusion.” *American Political Science Review* 115(2): 686–693.

Zhou, Xiang. 2019. “Hierarchical Item Response Models for Analyzing Public Opinion.” *Political Analysis* pp. 1–22.