UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

Formative and reflexive parameter bias under instances of continuous partial measurement noninvariance before and after item purification using the multiple indicator multiple causes model

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

Adon Rosen

Norman, Oklahoma

2022

Formative and reflexive parameter bias under instances of continuous partial measurement noninvariance before and after item purification using the multiple indicator multiple causes model

A THESIS APPROVED FOR THE

DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Hairong Song

Dr. Lauren Ethridge

Dr. David Bard

Table of Contents

List of Tables

List of Figures

Abstract

The multiple indicator multiple causes (MIMIC) model has been proposed as a powerful technique for the identification of partial measurement noninvariance (pMnI). Typically MI has been explored by comparing response patterns across groups using techniques such as the multiple group confirmatory factor analysis technique. The MIMIC model allows for the exploration of pMnI to be performed in relation to continuous covariates, however the specificity and sensitivity of the MIMIC to identify instances of continuous influenced pMnI is unexplored. This study first explores the bias that instances of continuous pMnI introduce in both formative and reflexive models when estimated within a MIMIC model framework using simulated data. Notable parameter estimation error is observed in extreme instances of both the formative and reflexive models. Next, the ability for the MIMIC model to identify and remove items which possess continuous pMnI are explored, high accuracy is obtained when instances of low and moderate MnI exist although performance degrades as the MnI increases in both magnitude and frequency. Finally, after removing items identified as MnI, parameter bias is again reevaluated in a similar framework noting reductions in parameter estimation bias in the formative model.

Keywords: Measurement invariance, Structural Equation Modeling, Multiple indicator multiple cause model

Introduction

The existence of reliable and valid data is a founding principle for any corpus of scientific knowledge. The behavioral sciences evaluate measurement invariance (MI) to ensure the validity and comparability of latent trait estimates across individuals. An example of an instance where MI is not satisfied is frequently observed in depression studies where biased response patterns exist when comparing male and female participants and their endorsement for questions such as, “I cry frequently,” (Steinberg & Thissen, 2006; Teresi et al., 2009). Within this example even with equivalent levels of depression there are frequently observed instances where females are more likely to endorse this item than their male counterparts. When such patterns arise in studies the latent trait of interest is obfuscated reducing both the validity and reliability of any statistical claims.

One of the most popular models used to assess the presence of MI is the multiple group confirmatory factor analysis model (Jöreskog, 1971). This model requires the participants to be categorized into several a priori groups (i.e. race or gender), and for the fit of theorized models to be explored across these groups. Other techniques afford the opportunity to explore instances of MI in relation to continuous variables, such as the moderated nonlinear factor analysis (Bauer, 2017), as well as the focus of this study–the multiple indicator multiple causes (MIMIC) model (Joreskog & Goldberger, 1975). While these techniques allow for the exploration of MI with continuous covariates, studies exploring the impacts of continuous covariates remain an inchoate methodological direction for psychometrics.

The multiple group confirmatory factor analysis remains one of the most popular techniques used to assess MI as this technique allows for models to be computed in parallel and utilizes easily computed statistics to assess invariance across groups. For instance the chi-square statistic, a well behaving asymptotic statistic, can be computed when comparing fit across various degrees of invariance. This is performed by comparing nested models where a model is fitted with increasingly strict parameterization and compared to a more flexible model to assess changes in model performance within and across groups. Elements of the multiple group confirmatory factor analysis have received prominent attention but several limitations have been addressed by more contemporary approaches, such as the inability to explore interactions across groups or continuous covariates when exploring MI, something both the moderated nonlinear factor analysis and the MIMIC model can incorporate into how the presence of MI is assessed.

The moderated nonlinear factor analysis technique was devised as a method to perform integrative data analysis (Bauer & Hussong, 2009). The goal of integrative data analysis is to synthesize datasets from multiple studies using unique measurement tools which measure a single theorized latent trait. This in turn leads to distinct MI issues: does the tool and the granularity of measurement bias the results; furthermore, some behavioral screeners are better positioned to identify lower or higher instances of a latent trait and combining across these tools introduces distinct methodological concerns. Sampling practices further influence the integration as studies may be focused on participants within a distinct range of ability. Motivated by all of these potential confounders, the moderated nonlinear factor analysis can incorporate studies of MI when using multiple group factors, and continuous covariates (Bauer, 2017). While this approach is appropriate when working with integrative data analysis, and has been used to explore MI in tasks similar to a multiple group factor analysis, the MIMIC model is distinguished from the moderated nonlinear factor analysis because of its ability to model instances of MI in relation to specific causal variables of interest.

The MIMIC model was originally derived to map a set of causal variables onto a set of indicator variables through one theorized latent variable (Joreskog & Goldberger, 1975). It was later utilized to explore instances of differential item functioning (DIF), a form of MI where a single item’s characteristics are explored (Muthén, 1985). The benefits of the MIMIC model for such explorations is the ability to identify instances of when and how MI may or may not be satisfied in specific items. Such explorations are in line with the “third generation” of invariance studies where the focus extends from identifying if MI exists, towards the goal of explaining why MI may or may not exist (Zumbo, 2007). The MIMIC model is uniquely positioned, when contrasted against the multiple group confirmatory factor analysis and the moderated nonlinear factor analysis techniques, given its ability to incorporate causal modeling into MI explorations.

Understanding the capabilities of these techniques to identify MI in relation to continuous covariates is important for applied researchers. This study explores the MIMIC models vulnerability to instances of measurement non invariance (MnI); followed by the models ability to identify and remove MnI items. First background on prototypical MI approaches including historical approaches is introduced, followed by an introduction to the MIMIC model and its extension to identify instances of MnI. Next, a stimulation study exploring parameter estimation error when MnI is ignored in a MIMIC model, and the MIMIC model’s ability to identify and remove MnI items is performed. This is then followed by a discussion detailing the importance of a closed system which can identify and remove instances of MnI for applied researchers.

Background on Measurement Invariance

The APA defines measurement invariance as “the situation in which a scale or construct provides the same results across several different samples or populations” (APA, 2014, p. 211). The basis of MI can be formulated in the following equation:

$$P\left(Y^{\square} \vee \eta, V\right)=P\left(Y \vee \eta\right) \text{ (1)}$$

where Y reflects a response to an item or set of items, P(Y) reflects the probability of Y occurring, η reflects an individual’s latent trait, and finally V reflects a matrix of individual traits. The above formula distinguishes that the probability of observing a response pattern is independent of an individual's traits. Studies exploring MI using a multiple group confirmatory factor analysis approach typically employ a hierarchy detailing specific levels of congruence across factor models. Studies first typically explore configural invariance, which is present when the factor structure, or the existence of the latent traits is equivalent across groups, followed by assessments of weak, strong, and strict factorial invariance (Meredith, 1993). In order to further expand on these concepts, the multiple group confirmatory factor analysis will be displayed using the following formula:

$$E\left(y_i \vee \eta_i, g\right)=v_g+\Lambda_g \eta_i \text{ (2)}$$

$$V\left(y_i \vee \eta_i, g\right)=\Sigma_g \text{ (3)}$$

$$E\left(\eta_i \vee g\right)=\alpha_g \text{(4)}$$

$$V\left(\eta_i \vee g\right)=\Psi_g \text{(5)}$$

Here $y_i$is a p × 1 vector which contains the item responses for individual i, η_i is a r × 1 vector of latent traits, and g represents a group factor indexing the group membership. The intercepts and slopes (or factor loadings) from the regression of the items on the factors within group g are contained in the p× 1 vector ν_g and p× r matrix Λ_g respectively, whereas the group-specific residual variances and covariances of the indicators are contained in the p× p matrix Σ_g. Usually, Σ_g is assumed to be diagonal, consisting only of the residual variance parameters. Finally, the r×1 vector α_g contains the group means for the factors, whereas the r×r matrix Ψ_g contains the group-specific factor variances and covariances.

Using Meredith’s hierarchy when weak invariance is present, only the factor loadings are equivalent across groups i.e. $\Lambda_g = \Lambda$. When weak factorial invariance is present mean group comparisons should be avoided, but comparisons of variance or covariance are permissible. The next form is strong invariance where both the loadings and intercepts i.e. $v_g = v$ are equivalent and permits the exploration of group mean differences. Finally, strict invariance exists when the loadings, intercepts, and residual variance is equivalent across groups i.e. $\Psi_g = \Psi_{\square}$. These explorations are typically performed by fitting a model across an entire battery, allowing for the item parameters to be explored across an entire behavioral battery; distinct from this are explorations of partial measurement invariance which explore individual item characteristics within a battery.

Partial measurement invariance (pMI) is typically explored when batteries perform well, but specific subsets of items display biased response patterns. The absence of partial MI can be formulated as:

$$P\left(y_{ij} \vee \eta , V\right) = P\left(y_{ij} \vee \eta\right) \text{ (6)}$$

In this representation y_{ij} represents an individual's probability to endorse a specific item. This formula extends the formulation of equation (1) but focuses on a specific item within a battery. Similar to Meredith's hierarchy for levels of MI, there are specific subsets of partial measurement invariance where only an item's intercept or an item's loading display biases. When items do not satisfy partial measurement invariance it suggests configural invariance may not be present. Literature suggests this phenomenon manifests when latent factors that influence response patterns are not being accounted for in the model (K. A. Bollen, 1989b). Previous research has detailed the issues ignoring pMnI can introduce in both the measurement model and downstream statistical models. When pMnI is ignored within a measurement model it increases parameter estimation error across the entire measurement model; furthermore, when pMnI is included downstream statistical tests become muddled and this is more pronounced in nonlinear tests such as a moderation test (Guenole & Brown, 2014; Hsiao & Lai, 2018; Li & Zumbo, 2009). Specific to this study, analyses will explore the impact of continuous partial Measurnment nonInvariance (cpMnI) when the group covariate variable (   ) is continuous in nature as opposed to the typically used group factor.

Historical approaches for the assessment of pMI

The presence of partial pMI is typically assessed in a post-hoc manner and has historically required group assignment. For instance, the Mantel Haenszel (MH) approach assesses pMI by exploring systematic differences of contingency tables across ranges of ability as estimated from a battery (Holland & Thayer, 1986; Mantel & Haenszel, 1959). For example, participants can be discretized into the number of items answered correctly and the group factor of interest forming a set of two-by-two contingency tables. Next, the Mantel Haenszel chi-squared statistic can be calculated by calculating biases in contingency tables across the range of correct answers. The major appeal of this approach is that it yields metrics which can be used similar to effect sizes, in the form of odds ratios, so the magnitude and significance of pMI can be obtained; furthermore, the output statistic follows a chi-squared distribution allowing for easily testing null and alternative hypotheses. One prominent limitation of the MH approach is the strict requirement of group membership, as well the ability to only explore differences in intercept (Andre A. Rupp & Bruno D. Zumbo, 2006; Li & Zumbo, 2009). An example of a method which can forgo group assignment is a logistic regression based approach (Swaminathan & Rogers, 1990). The logistic based approach requires predicting the probability of endorsement when modeled by overall test score, or an ability estimate, and including the predictors which might indicate pMI. The benefits of this approach are that interpreting the presence of pMI is equivalent to testing for moderation, and can be performed using either a t-statistic or parameter magnitude to identify levels of bias which are unacceptable. Limitations of the logistic approach include those inherent to logistic regression: for extremely difficult items, maximum likelihood has a difficult time obtaining interpretable coefficients when an item has perfect fit and when an item is rarely endorsed (Kleinbaum & Klein, n.d.).

The MIMIC model

The MIMIC model reflects a suite of tools derived from structural equation modeling (SEM). The MIMIC model extends attractive components of SEM modeling such as latent variable estimation and path analysis and incorporates these into a system of formative and reflexive models (see Figure 1A). The appeal of a formative and reflexive relationship is the ability to map a set of causal variables onto a set of indicator variables through a theorized causal system. A benefit of the systems of equations approach is the ability to incorporate error from both the formative and reflective model in order to emphasize this benefit, both the formative and reflexive models, and the system will be described.

Beginning with the reflexive model, a measurement model's distinctions will be described using language and formulation similar to that of item response theory (IRT) (Embretson & Reise, 2000):

(7)

In the above model, $p_i(\theta)$ is the probability of endorsement for an item given an individual's latent score estimate, is the item discrimination (i.e. factor loading), and b_i is the item difficulty (i.e. factor intercept). The above formula highlights how, given a set of manifest variables, IRT estimates a probability to endorse an item given an item's discrimination and difficulty estimates. Items which can better discriminate across groups at a specific trait level have higher discrimination parameters; difficulty reflects the location of the probability of endorsement being a 50% chance for a binary item. When working with binary data, the logic of IRT extends beyond the formula to read as:

$$y_i = \begin{cases} 1, & \text{if } y_i^* > \gamma_i, \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Where is a threshold parameter for $y_i^{\square}$, and assuming that:

$$y_i^{\square} = \lambda_i \eta + \epsilon_i \ (9)$$

Where $\lambda_i$is a loading parameter, and η reflects an individual's latent ability and $\epsilon_i$reflects the residual variable. This formulation highlights how, when provided with a set of binary manifest variables, a theorized normally distributed latent trait can be estimated (Joreskog & Goldberger, 1975; Muthén, 1984).

The formative model adheres to the following formulation:

$$\eta = \gamma x + \zeta \quad (10)$$

Where is a vector of the regression coefficients, x is a q x 1 vector of manifest random variables, where q is the number of observed variables and ζ reflects the residual term. While this takes the form of a causal model, previously it has been suggested that this moreso reflects a linear relationship between the independent and dependent variables (Muthén, 1985; Pearl, 2012).

The MIMIC model combines these into a system of equations resulting in the following formulation:

(11)

From this system of equations approach, error from both sets of manifest variables is incorporated, which improves the predictive and inferential nature of the model. Via the joint estimation of these two formative and reflexive models, the maximal relationship between the sets of manifest variables is obtained.

The MIMIC model for pMI assessment

The MIMIC model assesses for pMI by the inclusion of a direct path from the causal variables onto the response patterns of an individual indicator variable (see Figure 1B). By allowing for a direct path between the covariates of interest (i.e. socioeconomic status) and the response patterns (i.e., correctly endorsing an item), it allows for differences in the item's characteristics to be modeled after controlling for the latent ability estimate. Through a mediation framework, pMI is present when this direct effect is not fully mediated (Montoya & Jeon, 2020, see figure 1B). Other components of the mediation model can be used to explore for patterns of interest; for instance, a significant path from the covariates of interest on the latent variable suggests differences in the latent ability. Beyond the ability to use mediation model techniques, the MIMIC model also reduces the sample size required to perform a DIF exploration. When compared to the MH technique, where participants are binned into discrete performance bins, the MIMIC model imposes a probabilistic distribution across an entire range of the covariates distribution. This increased sensitivity has previously been a source of study (Montoya & Jeon, 2020; Woods & Grimm, 2011). As the MIMIC model is using a mediation framework to explore instances of partial MI, the causal variables can assume either a group or a continuous form, allowing for the exploration of cpMnI, distinguishing the MIMIC model from the prominent MI techniques.

In order to ensure an entire behavioral screener is MI, a purification procedure is performed. The purification procedure is an iterative and exhaustive process which requires fitting a mediated MIMIC model for every indicator variable (see figure 1B). After all mediated MIMIC models are estimated, any item where the relationship between the causal variable and the item's response pattern is not fully mediated is excluded. Following this exclusion, a new round of mediation models are trained for every item remaining in the item bank (see figure 1C). Previous research has suggested this to be best practice when implementing the MIMIC model for scale purification (Wang et al., 2009).

The MIMIC model for MI testing with continuous covariates

Studies which have examined cpMnI using the MIMIC model can be found in developmental literature (Le et al., 2019) and clinical literature (Stevens et al., 2022). Le et al. sought to validate parental hopes for their children; cpMnI was explored using parental age, child age, and socioeconomic status. Results indicated distinct differences in response patterns in relation to both age and socioeconomic status. An example from the clinical literature explored differences in alcohol use amongst an adolescent pediatric sample; item characteristics for differences in expectations in relation to alcohol use were explored across participants' age, sex, race, and socioeconomic status (Stevens et al., 2022). Results suggested differences in response patterns based on an individual's socioeconomic status and age. This brief literature review suggests several facts: first, instances of cpMnI do exist in relation to demographic variables such as socioeconomic status and second, the extant literature for instances of cpMnI is limited. One possible contributing factor for this pattern is the limitation of studies exploring impacts of cpMnI, and second, studies detailing the ability to identify and remove instances of cpMnI are limited.

Current Study

The current study seeks to fill several gaps in the extant cpMnI literature. These include exploring the impacts that the inclusion of cpMnI items introduced in formative relationships; second, how well can the MIMIC model perform item purification when working with cpMnI items; and third, how much does item purification impact parameter estimates. Furthermore, this study represents a unique direction for MI research: identifying purified itemsets within a specific causal mechanism of interest. As previously described, the typical MI study is performed in a post-hoc fashion attempting to remove bias contributed by potential nuisance variables in a study. Extending this research to identify itemsets free from bias in relation to the causal variable of interest is critical, as one potential mechanism creating item bias is an unmodeled latent trait (K. A. Bollen, 1989b). When such instances exist, the parsimony of reported findings may be impacted.

Methods

Approach overview

The goals of this simulation study were multifaceted and included: first, exploring the impact instances of cpMnI have on formative and reflexive relationships within a MIMIC model, second, exploring the capabilities of the MIMIC model to identify instances of cpMnI, and third, exploring how item purification procedures impact parameter estimates. This process required several discrete tasks. First, binary indicators and continuous causal variables were simulated under various population-wide parameters. Second, MIMIC models were trained ignoring all instances of cpMnI, and estimation bias is explored from these models in both the formative and reflexive components. Third, the identification of DIF is explored following an iterative purification procedure using the MIMIC model; the results of this process are explored using both item-wise and model-wise methods. Fourth and finally, MIMIC models are re-estimated, removing items identified as cpMnI from the third step, and parameter estimation error is again explored as in step one. The following subsections of the Method section detail each of these steps in full. All code can be found online in a GitHub repository (https://github.com/adrose/mastersThesis).

Simulation Conditions

Simulation factors were varied in 6 ways, for a total of 576 conditions. The factors included:

1. The number of examinees. This number varied the sample size of the simulated study ranging between a sample size, which meets a minimum recommended sample size standard for a structural equation model exploration (n=200) to a moderately powered exploration (n=500). The minimum recommended sample size follows recommendations from Bollen (1989) where it is recommended to have 5 observations per freely estimated parameter. The moderately powered sample size follows more contemporary recommendations for roughly 10 observations for freely estimated parameters (Christopher Westland, 2010).
2. The magnitude of the indicator variables. The magnitude of the relationship between the binary indicator variables and the theorized latent variable (i.e., reflexive model) varied between weak (Beta = .4) and strong (Beta = .8). The strength of the indicator was selected for the even and odd valued indicators separately so in total four permutations of the indicator strength were possible.
3. The number of cpMnI items. The number of items varied between 2 (10%), 4 (20%), and 6 (30%) of items which were modeled to include a direct relationship between the causal variable and the indicator. These values were taken from similar MIMIC explorations (Wang et al., 2009); however, the ceiling was lowered as this is where previous reports indicated extremely poor performance for DIF identification.

4 The magnitude of cpMnI. The magnitude of the direct effect from the causal variable to the indicator variable after controlling for the latent variable ranged between 0, .2, .4, and .6.

5 Item intercept. This condition type varied the item intercept thresholds—i.e. how high on the latent trait an examinee has to be to have a 50% probability of endorsement. Difficulties of screen items were drawn randomly from a uniform distribution ranging from [-1 to 1] or [0 to 2].

6 The magnitude of the causal relationship. The strength of the formative model included values from .2, .4, and .6.

Across all conditions the number of indicator and causal variables were held constant at 20 and 1 respectively. Across all permutations 100 datasets were simulated, in total 57600 datasets were simulated. All simulation was performed using MPlus (Linda K., Muthén & Bengt O., Muthén, 2017).

Error in parameters ignoring cpMnI

In order to explore the impact that cpMnI has on parameter estimation a single MIMIC model was estimated ignoring the potential existence of cpMnI items. This was performed by estimating a MIMIC model using all indicators (p=20) and causal (p=1) variables (see figure 1A). The model outcome was the root of the squared error, this allows for any deviations from the true parameter to be highlighted in a consistent fashion. Separate models were estimated for both the formative and the reflexive models. All results were explored using ANOVA, all factors were included and up to all four-way interactions were included as predictor variables.

Model purification accuracy

In order to identify indicator variables which exhibit cpMnI an iterative approach is applied, this approach follows previously described methodology (Wang et al., 2009) and is described briefly here. Through a mediation framework, the goal of using the MIMIC model to identify instances of cpMnI is to map response patterns onto a causal variable after controlling for a theorized latent variable. Within this framework the mediating variable is the latent trait, the independent variable is the causal variable, and the dependent variable is the response pattern of a single indicator variable (see figure 1B). The iterative nature of the item purification is twofold (see figure 1C): first, a MIMIC model is trained for every indicator variable as the independent variable, second instances of cpMnI are identified and removed from further models, step one is then repeated after removing any item which is identified as possessing cpMnI. The presence of cpMnI is identified when the relationship between the causal variable and an indicator variable is not fully mediated by the latent variable. The presence of this relationship is identified by a significant path at a predetermined α of .05. The outcomes for this exploration explored model-wise classification performance within an individual dataset across all possible iterations of the purification procedure.

Parameter recovery after purification

In order to explore parameter recovery after the purification procedure is performed MIMIC models were estimated removing any item that was identified as possessing cpMnI from the previous step. The goal here is to compare estimation error from the MIMIC models which were estimated ignoring the presence of cpMnI from those that underwent the purification procedure. These models build upon the previous set of analyses exploring estimation error in models ignoring instances of cpMnI by including an additional factor detailing if the model estimate is derived from a purified MIMIC model.

Results

Error in parameters ignoring cpMnI

Table 2 shows the results of an ANOVA relating the simulation conditions (plus all interactions) to estimation error between the true and estimated formative model parameters when ignoring impacts of cpMnI. All results are statistically significant but note that significance of effects is confounded by the number of simulations. Therefore, meaningful interpretation of the ANOVA results requires effect sizes; table 2 includes eta-squared values. Of the main effects, the largest eta squared is for the sample size (eta squared = 0.045) and the smallest was for the minimum item difficulty (eta squared = 0.003). The main effects are displayed graphically in figure 2A. The largest two-way interaction was between the magnitude of the cpMnI and the number of cpMnI items (eta squared = .013; see figure 2B); the interaction suggests that the estimation error increases faster as greater, and more instances of cpMnI are introduced. The largest three-way interaction was observed across the magnitude of the cpMnI, the magnitude of the indicator variables, and the number of cpMnI items (eta squared = 0.009). The interaction indicates the lowest error is observed in datasets with strong indicator variables where even under strong instances of cpMnI estimation error remains similar to those itemsets without cpMnI; however, when low magnitude indicator variables are present estimation error increases very rapidly when both magnitude of cpMnI increases and frequency of cpMnI items increases (see figure 2C). The four-way interaction with the largest eta squared included the variables from the three-way interaction and the magnitude of the formative relationship (eta squared < 0.001) and indicted that models with weak instances of causal relationships and large number of and strong magnitude of cpMnI show slightly reduced error but these effects are washed away as the magnitude of the indicator variable increases (see figure 2D).

The next set of analyses explored parameter estimation error from the reflexive model when ignoring cpMnI, table 3 displays the results of an ANOVA relating the simulation conditions (plus all interactions) to estimation error within these models. The largest main effect was for the strength of the indicator set (eta squared = 0.071) suggesting that as the magnitude of the indicator increases the estimation error decreases (see figure 3A). The strongest two-way interaction was observed between the magnitude of the indicators and the indicator factor (eta square = 0.030); further underscoring that the estimation error is lower in models with strong indicators (see figure 3B). The strongest three-way interaction extended the previous two-way interaction with the magnitude of the cpMnI (eta square = 0.009); suggesting that estimation error increases faster in those items which have greater instances of cpMnI (see figure 3C). The strongest four-way interaction extended the three-way interaction to include the number of items with cpMnI (eta-squared = 0.002), highlighting how estimation error increases in itemsets with larger and more frequent instances of cpMnI (see figure 3D).

Model purification accuracy

Next the model wise classification performance was explored, table 3 displays an ANOVA relating the simulation conditions (plus all interactions) to the classification performance for the model wise results. Of the main effects the largest eta squared value was observed for the magnitude of cpMnI (eta squared = 0.364), results indicate that models perform extremely well at identifying instances where cpMnI is not present but nonlinearities are observed where when increasing the magnitude of cpMnI. The greatest accuracy is seen in the 0.4 magnitude condition and lower accuracy was observed in the 0.2 and 0.6 instances (see figure 4A). The largest two-way interaction was observed between the magnitude of the cpMnI and the number of items which had cpMnI (eta squared = 0.170). The two-way interaction indicates greater accuracy when fewer and weaker instances of cpMnI exist, accuracy decreases faster when stronger instances of cpMnI are introduced (see figure 4B). The largest three-way interaction extends the previous two-way to include sample size (eta squared = 0.040). This interaction indicates that as the magnitude of the cpMnI increases the accuracy increases within the smaller sample size (see figure 4C). Finally, the largest four-way interaction extends the previous three-way to include the magnitude of the causal relationship (eta squared 0.003), this interaction indicates that greater accuracy is observed in models with stronger causal relationships (see figure 4D). In order to further elucidate the driving factor behind this decrease in accuracy true positive, true negative, false positive, and false negative rates were modeled in an additional ANOVA model. Results indicate decreminates in performance were driven primarily by increases in false positive rates (see figure 6).

Error in parameters removing cpMnI

Table 4 shows the strength of the additional purification factor (and all possible interactions) in an ANOVA modeling formative parameter estimation error. The table suggests a strong main effect of purification status (eta squared = 0.065) with the direction of the effect suggesting lower estimation error in the post-purification model (see figure 7A). There were two two-way interactions which also had large eta squared values these were: the interaction between purification status and the magnitude of the cpMnI (eta squared = 0.065; see figure 7B), and the interaction with the purification status and the number of cpMnI items (eta squared = 0.065). These two-way interactions both suggested lower error from the model which had undergone purification. This pattern of lower estimation error continues to extend through a three-way interaction which includes the purification status, the magnitude of the cpMnI, and the number of cpMnI items (eta squared = 0.026; see figure 7C). Finally the largest four-way interaction extends the three-way interaction to include the strength of the indicator variables (eta squared = 0.006) suggesting that the estimation error is the lowest in a purified item set, with a strong indicator set, and fewer instances of cpMnI items; however, when working with weak itemsets, with strong and frequent instances of cpMnI estimation error increases rapidly for models which have not undergone purification whereas in the purified model the estimation error remains lower (see figure 7D).

Table 5 shows the strongest interactions between the original simulation parameters and the additional parameter indicating if the parameter was estimated from a model which had undergone the purification procedure for the reflexive model. The eta-squared values from the ANOVA suggest small main effects for purification status (eta squared < 0.001; see figure 8A); however the direction of the effect suggests increased estimation error after the purification procedure is performed. The largest two-way interaction between the purification status and the magnitude of the cpMnI (eta squared < 0.001; see figure 8B) suggests a similar trend where the estimation error is larger in the models which have undergone the purification procedure. The largest three-way interaction extends the two-way interaction to include the individual indicator variables (eta squared = 0.001; see figure 8C) indicating that variables where cpMnI exists have considerably greater estimation error than those which were never modeled to include cpMnI. Finally, the largest four-way interaction extends the three-way interaction to include the number of cpMnI items: suggesting that the models with more frequent instances of cpMnI items have greater estimation error, although this pattern is much more exaggerated in the specific items which were model to include cpMnI (eta-squared = 0.001; see figure 8D).

Discussion

In this study the issues surrounding the inclusion of cpMnI items in a MIMIC model are highlighted. This paper highlights how estimation error increases in both formative and reflexive models as the frequency and magnitude of cpMnI items increases. The next set of analyses explored the accuracy of item purification performed in an iterative and exhaustive manner using the MIMIC model. Modelwise accuracy is inline with previous reports identifying the MIMIC model as a technique with very good purification capabilities. Following item purification estimation error was again explored by removing any item identified as cpMnI and estimating the MIMIC model in a reduced indicator variable itemset. Results from this model suggested decreased estimation error in the formative model but increased estimation error in the reflexive model. Taken together this simulation study offers an insight into potential estimation error and a technique which can alleviate the presence of poor performing items.

Estimation error in formative models

Through a causal modeling framework a formative model is used to specify the cause of an unobserved latent trait in the MIMIC model. Compared to the reflexive (measurement) model, parameter estimation error has received sparse attention. This is potentially motivated by some of the inherent limitations regarding parameter estimation and formative models such that for a formative model to be accurately identified all of the causes of the latent trait should be present in the model (Diamantopoulos, 2006; Diamantopoulos et al., 2008). In fact, it has been proposed that omission of items is similar to restricting the domain of the latent trait (MacKenzie, 2003). Further issues surrounding the implementation of formative models include the inability to include measurement error on the indicator variables within the model (Edwards & Bagozzi, 2000). The results of this simulation study offer an additional issue surrounding the formative model utilized in a MIMIC model: susceptibility to bias in the estimation of the latent trait. The results highlighted in Figure 2 indicate in the most extreme cases the root of the squared estimation to be 0.20, which underscores the ability for issues in the reflexive model to cause issues in the formative model. While isolated measurement and formative models were not explored within this study previous research exploring impacts of MnI on latent trait estimation indicate greatest bias introduced into an individual’s latent score as more frequent instances of pMnI are introduced (Andre A. Rupp & Bruno D. Zumbo, 2006). This finding offers one explanatory mechanism for the cause of the formative model estimation error, increasing error in the estimation of the latent score further increases error of models using this latent score as a dependent variable.

Another key takeaway from the ANOVA models was that the stronger the population formative model, the larger the parameter estimation error. This is in contrast with previous studies exploring Type-1 error in downstream statistical conclusions. One such example exists where data were simulated without group differences, however when latent traits were estimated in the presence of MnI *t*-tests were used to compare the simulated equivalent group and Type-1 error rate was greater than the a priori alpha (Li & Zumbo, 2009). One further aspect identified by the Zumbo & Li study is the frequency of, the direction, and the magnitude all impact downstream statistical conclusions, in this study all instances of cpMnI were aligned in both magnitude and direction, potentially inflating the parameter estimation error. Regardless, the results of this study should suggest to applied researchers the dangers that cpMnI can introduce in causal models.

Estimation error in reflexive models

Researchers employ reflexive models given their ability to obtain unbiased estimates for relations across indicator variables (K. Bollen & Pearl, 2012). Importantly, methodologists have advocated for the use of latent variable models whenever possible as they reduce the influence of measurement error (Borsboom, 2008). While these are a powerful modeling technique there are limitations when improperly specified. Previous work has detailed how parameter over and underestimation can manifest in improperly specified measurement models (Cole & Preacher, 2014; Ledgerwood & Shrout, 2011). Furthermore, in relation to instances of pMnI there exists distinct issues of parameter estimation error when such instances are ignored (Guenole & Brown, 2014). These findings underscore how when models are agnostic to MnI parameter estimation bias reaches undesirable levels as more frequent instances of pMnI are introduced. Asimilar narrative is displayed in this study. The largest three-way interaction observed when exploring parameter estimation error in the reflexive model was between the indicator, the strength of the loading, and the magnitude of cpMnI (see figure 3). This interaction displays two distinct findings: the first, a strong indicator set protects against parameter estimation error, and second in line with previous reports, more frequent and greater magnitude cpMnI increases parameter estimation error across the entire model.

Accuracy of itemset purification procedures using the MIMIC model

With the impacts of cpMnI clearly identified, the next set of analyses sought to explore the MIMIC model as a tool to identify and remove impacted items. This process used an iterative and exhaustive procedure as previously detailed (Wang et al., 2009). Model Wise accuracy metrics at minimum require a total of 20 models, in this simulation study the maximum number required to perform the purification process was 105 models. Interestingly, even with the increase in statistical comparisons the family wise error rate was similar to the prespecified alpha level. Such results are convergent with previous implementations of this purification procedure (Kim et al., 2012; Wang et al., 2009; Woods & Grimm, 2011). While instances of Type-2 error are well controlled under these practices, Type-I errors reach unacceptable rates under various conditions. In the worst case scenarios when cpMnI items were both frequent and possessed large magnitudes the MIMIC model displayed an accuracy of 60% when identifying cpMnI items. The reduction in accuracy appeared to be driven predominantly by an increase in the false positive rate (see figure5). These results were further compounded by increasing sample size offering a cautionary note regarding the null hypothesis significance testing procedure these analyses are reliant upon. Specifically the MIMIC model quickly becomes overpowered when attempting to identify significant relationships between indicator and causal variables. An adaptive significance threshold could be considered to mitigate this concern.

Comparison of unpurified with purified MIMIC models

Next, using models which had undergone the purification procedure parameter estimation error was again calculated using a purified itemset. Results suggest model purification consistently reduces estimation error in the formative model irregardless of the frequency or the magnitude of cpMnI items. In the most extreme instances where the formative models had weak indicators and large magnitude of cpMnI the greatest error was observed (rmse = 0.4) but the purification procedure reduced estimation error by more than half (rmse = .18). In fact, across the entire subset of formative models with weak indicators estimation error is reduced across all instances of cpMnI frequency. Furthermore, the estimation error is comparable when purification is performed in datasets without any cpMnI items. These results suggest item purification is a low risk preprocessing step for itemsets even with desirable characteristics.

In stark contrast with the encouraging results from the formative model, estimation error increases in the reflexive model following item purification. This effect is best highlighted in the four-way interaction in which items which were not modeled to include cpMnI also displayed elevated estimation error (see Figure 8D: items 7 & 8). One explanation for this effect is that poor performing items were more likely to remain in the model when in the worst case scenarios: weak indicator sets and large frequency of cpMnI items. Furthermore, in these undesirable conditions the removal of properly performing items becomes more frequent as the number of false positive cases increases (see figure 5). These combinations together make for instances where the estimation error in the reflexive model increases as a result of the purification procedure.

Implications for applied researchers

The predominant MI testing techniques rely upon group splits (i.e. high versus low socioeconomic status) in order to explore the quality of behavioral data. While the binarization of continuous covariates is a common practice it offers several limitations and is often advised against in the behavioral sciences (Altman & Royston, 2006). Not only would such a practice introduce researcher degrees of freedom (i.e. mean versus median split), but it is inherently less sensitive to differences which exist along a continuum: when using a median split the cost of power is equivalent to removing one third of all observations (Cohen, 1983). The methodology proposed and explored within this simulation study details a modeling technique which can accurately identify and remove poor performing items in relation to a continuous gaussian variable.

Another prominent feature of this study is the item purification procedure is performed in relation to a specific causal variable of interest. The typical MI exploration is performed in a post-hoc manner which explores response bias in potential confounding variables. While this is an important practice to reduce potential bias of demographic variables, it does not ensure the latent variable of interest is properly tuned to a causal variable. As instances of pMI are theorized to be introduced by unmodeled latent traits it is important to ensure a closely coupled relationship between the cause and indicators (Millsap, 2007). By performing item purification with the MIMIC model it allows researchers to use specific causal variables of interest to ensure tightly coupled theoretical explorations.

Limitations

This study has several prominent limitations that include: only a single causal variable was simulated, all instances of cpMnI were simulated in an identical direction and equivalent magnitude, the reliance of the null hypothesis significance test framework. The use of only a single causal variable was selected in relation to how item purification is typically performed with the MIMIC model (Montoya & Jeon, 2020) typically when the MIMIC model is used in an applied setting many causal variables are used. Previous researchers have explored how multiple instances of pMnI impact factor scores (Edwards & Bagozzi, 2000; Li & Zumbo, 2009), these studies have explored how differences in magnitude and direction of pMnI impact these scores. The results of these previous studies suggest that impacts when items are in opposite directions reduces the impact of pMnI on factor scores estimates. This study chose to only simulate identical magnitude and direction in impacted items which may have increased the effects of formative model and reflexive model parameter estimation error. Finally, and most importantly, the reliance of the null hypothesis significance testing framework certainly contributed to some issues as illustrated by the high false positive case count for models with the higher sample size. One of the biggest limitations and concerns about pMI explorations are the multitude of manners items can be identified as performing satisfactory here a strict nominal p-value was utilized when it is known that there are many alternatives albeit, no clear predominant technique (Borsboom, 2006). Other methods should be considered such as model fit statistics or nonparametric statistics.

Conclusions

This study highlighted the issues cpMnI can introduce in both formative and reflexive relationships in a MIMIC model. Importantly, formative and reflexive relationships become error prone as large and frequent instances of cpMnI are introduced but this effect is mitigated by strong indicators. Item purification represents a low cost technique researchers can apply in order to reduce parameter estimation error in formative relationships; however, caution should be applied if the goal is to identify reflexive relationships as this can potentially increase error.

References

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ : British Medical Journal*, *332*(7549), 1080.

Andre A. Rupp & Bruno D. Zumbo. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement*, *66*(1), 63–84. https://doi.org/10.1177/0013164404273942

Bauer, D. J. (2017). A More General Model for Testing Measurement Invariance and Differential Item Functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125. https://doi.org/10.1037/a0015583

Bollen, K. A. (1989a). Structural Equation Models with Observed Variables. In *Structural Equations with Latent Variables* (pp. 80–150). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118619179.ch4

Bollen, K. A. (1989b). The Consequences of Measurement Error. In *Structural Equations with Latent Variables* (pp. 151–178). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118619179.ch5

Bollen, K., & Pearl, J. (2012). *Eight Myths About Causality and Structural Models* (SSRN Scholarly Paper No. 2343821). Social Science Research Network. https://papers.ssrn.com/abstract=2343821

Borsboom, D. (2006). When Does Measurement Invariance Matter? *Medical Care*, *44*(11), S176. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and*

Perspectives, *6*(1–2), 25–53. https://doi.org/10.1080/15366360802035497

Christopher Westland, J. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, *9*(6), 476–487. https://doi.org/10.1016/j.elerap.2010.07.003

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Diamantopoulos, A. (2006). The error term in formative measurement models: Interpretation and modeling implications. *Journal of Modelling in Management*, *1*(1), 7–17. https://doi.org/10.1108/17465660610667775

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218. https://doi.org/10.1016/j.jbusres.2008.01.009

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/article/10.3389/fpsyg.2014.00980

Holland, P. W., & Thayer, D. T. (1986). *Differential Item Performance and the Mantel-Haenszel Procedure*. https://eric.ed.gov/?id=ED272577

Hsiao, Y.-Y., & Lai, M. H. C. (2018). The Impact of Partial Measurement Invariance on Testing

Moderation for Single and Multi-Level Data. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/article/10.3389/fpsyg.2018.00740

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. *Journal of the American Statistical Association*, *70*(351), 631–639. https://doi.org/10.2307/2285946

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing Measurement Invariance Using MIMIC: Likelihood Ratio Test With a Critical Value Adjustment. *Educational and Psychological Measurement*, *72*(3), 469–492. https://doi.org/10.1177/0013164411427395

Kleinbaum, D. G., & Klein, M. (n.d.). *Logistic Regression*. https://doi.org/10.1007/978-1-4419-1742-3

Le, B. M., Sakaluk, J. K., Day, L. C., & Impett, E. A. (2019). How gender, age, and socioeconomic status predict parenting goal pursuit. *Journal of Social and Personal Relationships*, *36*(10), 3313–3338. https://doi.org/10.1177/0265407518818375

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, *101*(6), 1174–1188. https://doi.org/10.1037/a0024776

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica*, *30*(2), 343–370.

Linda K., Muthén & Bengt O., Muthén. (2017). *Mplus User's Guide*. https://www.statmodel.com/html_ug.shtml

MacKenzie, S. B. (2003). *The dangers of poor construct conceptualization*. https://doi.org/10.1177/0092070303031003011

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2007). Invariance in Measurement and Prediction Revisited. *Psychometrika*, *72*(4), 461–473. https://doi.org/10.1007/s11336-007-9039-7

Montoya, A. K., & Jeon, M. (2020). MIMIC Models for Uniform and Nonuniform DIF as Moderated Mediation Models. *Applied Psychological Measurement*, *44*(2), 118–136. https://doi.org/10.1177/0146621619835496

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B. (1985). A Method for Studying the Homogeneity of Test Items with Respect to Other Relevant Variables. *Journal of Educational Statistics*, *10*(2), 121–132. https://doi.org/10.3102/10769986010002121

Pearl, J. (2012). The causal foundations of structural equation modeling. In *Handbook of structural equation modeling* (pp. 68–91). The Guilford Press.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402–415. https://doi.org/10.1037/1082-989X.11.4.402

Stevens, A. K., Janssen, T., Belzak, W. C. M., Treloar Padovano, H., & Jackson, K. M. (2022). Comprehensive measurement invariance of alcohol outcome expectancies among adolescents using regularized moderated nonlinear factor analysis. *Addictive Behaviors*, *124*, 107088. https://doi.org/10.1016/j.addbeh.2021.107088

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., Lai, J.-S., Choi, S. W., Hays, R. D., Reeve, B. B., Reise, S. P., Pilkonis, P. A., & Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science*, *51*(2), 148–180.

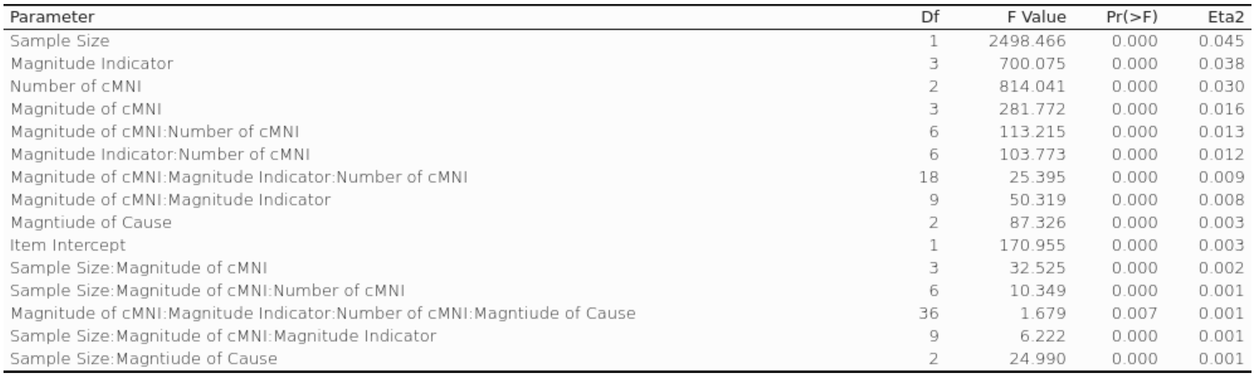
Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC Method With Scale Purification for Detecting Differential Item Functioning. *Educational and Psychological Measurement*, *69*(5), 713–731. https://doi.org/10.1177/0013164409332228

Woods, C. M., & Grimm, K. J. (2011). Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models. *Applied Psychological Measurement*, *35*(5), 339–361. https://doi.org/10.1177/0146621611405984

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

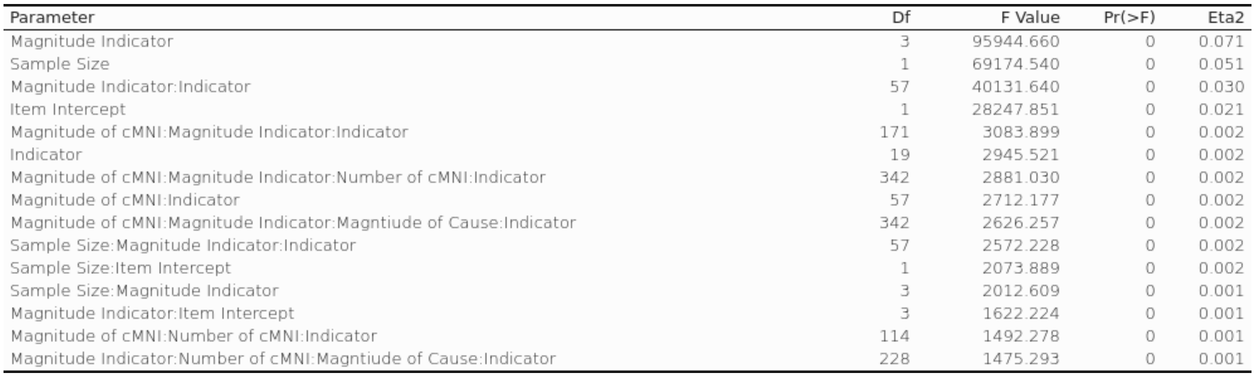
Tables

Table 1: Predictors with the largest eta-squared from an ANOVA modeling estimation error from formative models ignoring cpMnI.



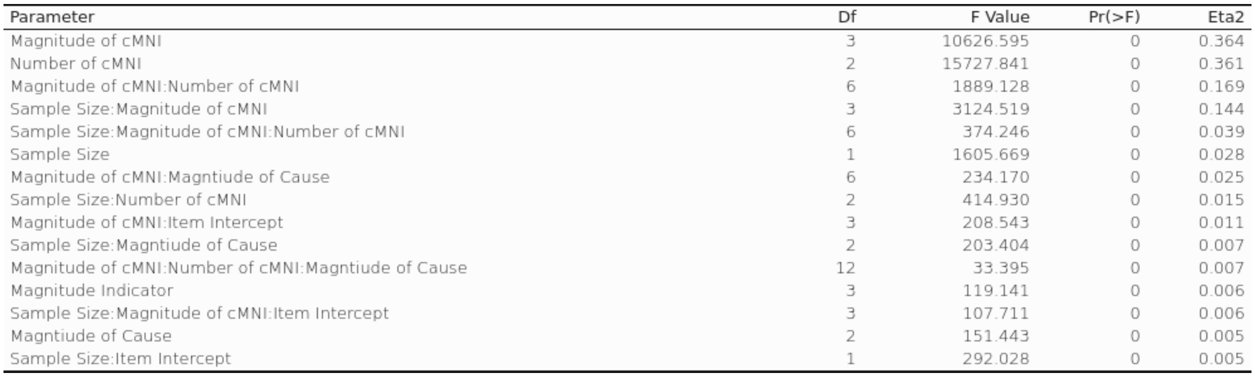
Parameter	Df	F Value	Pr(>F)	Eta2
Sample Size	1	2498.466	0.000	0.045
Magnitude Indicator	3	700.075	0.000	0.038
Number of cMNI	2	814.041	0.000	0.030
Magnitude of cMNI	3	281.772	0.000	0.016
Magnitude of cMNI:Number of cMNI	6	113.215	0.000	0.013
Magnitude Indicator:Number of cMNI	6	103.773	0.000	0.012
Magnitude of cMNI:Magnitude Indicator:Number of cMNI	18	25.395	0.000	0.009
Magnitude of cMNI:Magnitude Indicator	9	50.319	0.000	0.008
Magntiude of Cause	2	87.326	0.000	0.003
Item Intercept	1	170.955	0.000	0.003
Sample Size:Magnitude of cMNI	3	32.525	0.000	0.002
Sample Size:Magnitude of cMNI:Number of cMNI	6	10.349	0.000	0.001
Magnitude of cMNI:Magnitude Indicator:Number of cMNI:Magntiude of Cause	36	1.679	0.007	0.001
Sample Size:Magnitude of cMNI:Magnitude Indicator	9	6.222	0.000	0.001
Sample Size:Magntiude of Cause	2	24.990	0.000	0.001

Table 2: Predictors with the largest eta-squared from an ANOVA modeling estimation error from reflexive models ignoring cpMnI.



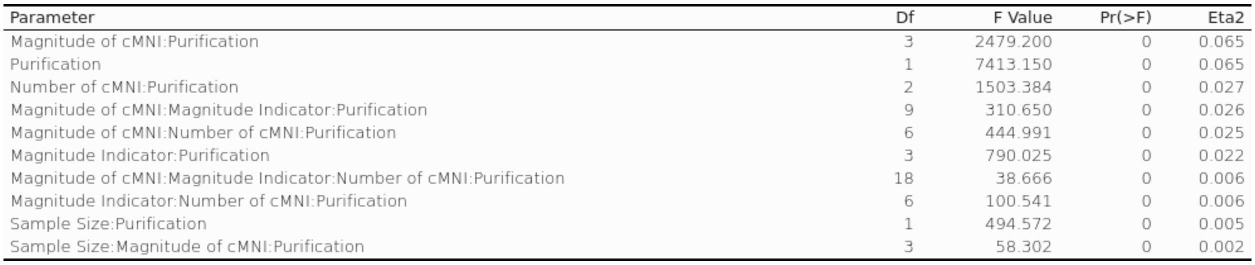
Parameter	Df	F Value	Pr(>F)	Eta2
Magnitude Indicator	3	95944.660	0	0.071
Sample Size	1	69174.540	0	0.051
Magnitude Indicator:Indicator	57	40131.640	0	0.030
Item Intercept	1	28247.851	0	0.021
Magnitude of cMNI:Magnitude Indicator:Indicator	171	3083.899	0	0.002
Indicator	19	2945.521	0	0.002
Magnitude of cMNI:Magnitude Indicator:Number of cMNI:Indicator	342	2881.030	0	0.002
Magnitude of cMNI:Indicator	57	2712.177	0	0.002
Magnitude of cMNI:Magnitude Indicator:Magntiude of Cause:Indicator	342	2626.257	0	0.002
Sample Size:Magnitude Indicator:Indicator	57	2572.228	0	0.002
Sample Size:Item Intercept	1	2073.889	0	0.002
Sample Size:Magnitude Indicator	3	2012.609	0	0.001
Magnitude Indicator:Item Intercept	3	1622.224	0	0.001
Magnitude of cMNI:Number of cMNI:Indicator	114	1492.278	0	0.001
Magnitude Indicator:Number of cMNI:Magntiude of Cause:Indicator	228	1475.293	0	0.001

Table 3: Predictors with the largest eta-squared from an ANOVA modeling accuracy of model wise identification of cpMnI.



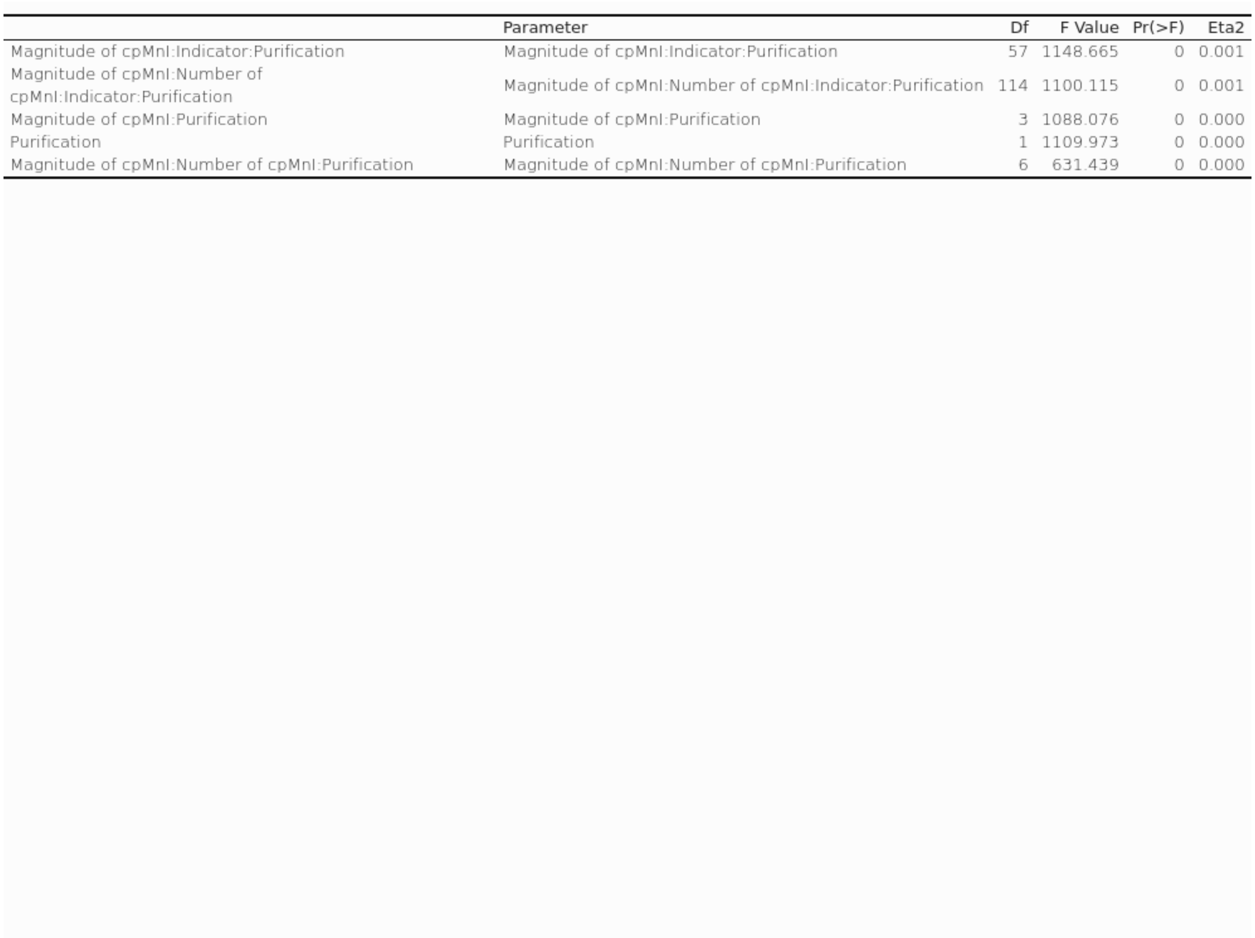
Parameter	Df	F Value	Pr(>F)	Eta2
Magnitude of cMNI	3	10626.595	0	0.364
Number of cMNI	2	15727.841	0	0.361
Magnitude of cMNI:Number of cMNI	6	1889.128	0	0.169
Sample Size:Magnitude of cMNI	3	3124.519	0	0.144
Sample Size:Magnitude of cMNI:Number of cMNI	6	374.246	0	0.039
Sample Size	1	1605.669	0	0.028
Magnitude of cMNI:Magntiude of Cause	6	234.170	0	0.025
Sample Size:Number of cMNI	2	414.930	0	0.015
Magnitude of cMNI:Item Intercept	3	208.543	0	0.011
Sample Size:Magntiude of Cause	2	203.404	0	0.007
Magnitude of cMNI:Number of cMNI:Magntiude of Cause	12	33.395	0	0.007
Magnitude Indicator	3	119.141	0	0.006
Sample Size:Magnitude of cMNI:Item Intercept	3	107.711	0	0.006
Magntiude of Cause	2	151.443	0	0.005
Sample Size:Item Intercept	1	292.028	0	0.005

Table 4: Predictors with the largest eta-squared from an ANOVA comparing pre- and post-purification reflexive model parameter estimation error.



Parameter	Df	F Value	Pr(>F)	Eta2
Magnitude of cMNI:Purification	3	2479.200	0	0.065
Purification	1	7413.150	0	0.065
Number of cMNI:Purification	2	1503.384	0	0.027
Magnitude of cMNI:Magnitude Indicator:Purification	9	310.650	0	0.026
Magnitude of cMNI:Number of cMNI:Purification	6	444.991	0	0.025
Magnitude Indicator:Purification	3	790.025	0	0.022
Magnitude of cMNI:Magnitude Indicator:Number of cMNI:Purification	18	38.666	0	0.006
Magnitude Indicator:Number of cMNI:Purification	6	100.541	0	0.006
Sample Size:Purification	1	494.572	0	0.005
Sample Size:Magnitude of cMNI:Purification	3	58.302	0	0.002

Table 5: Predictors with the largest eta-squared from an ANOVA comparing pre- and post-purification reflexive model parameter estimation error.



	Parameter	Df	F Value	Pr(>F)	Eta2
Magnitude of cpMnI:Indicator:Purification	Magnitude of cpMnI:Indicator:Purification	57	1148.665	0	0.001
Magnitude of cpMnI:Number of cpMnI:Indicator:Purification	Magnitude of cpMnI:Number of cpMnI:Indicator:Purification	114	1100.115	0	0.001
Magnitude of cpMnI:Purification	Magnitude of cpMnI:Purification	3	1088.076	0	0.000
Purification	Purification	1	1109.973	0	0.000
Magnitude of cpMnI:Number of cpMnI:Purification	Magnitude of cpMnI:Number of cpMnI:Purification	6	631.439	0	0.000

Figures:

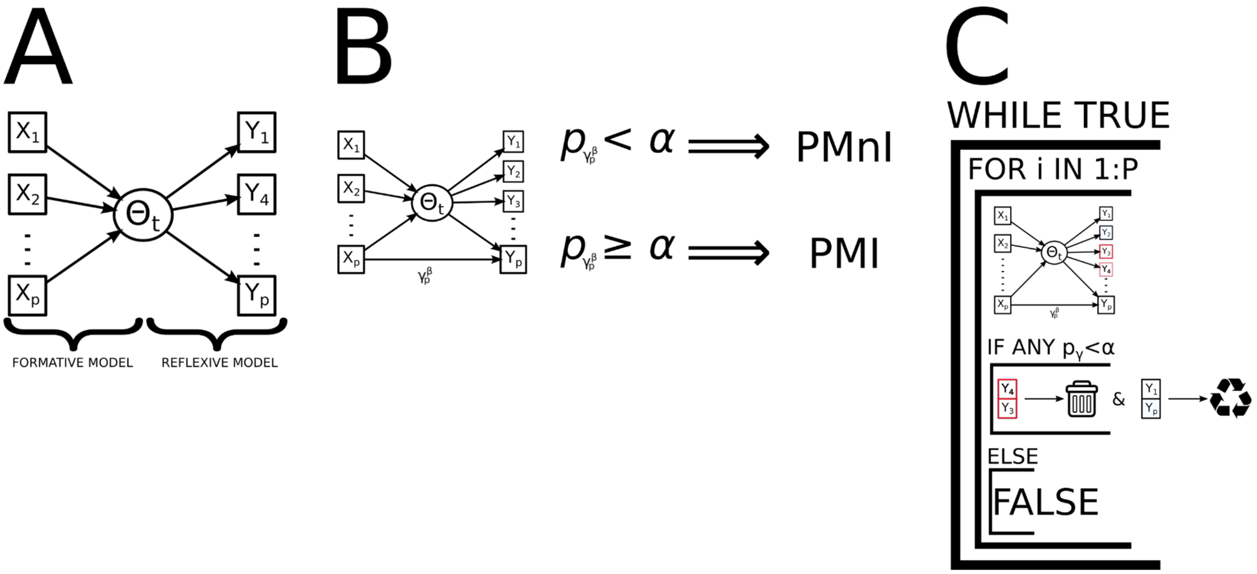


Figure 1: An overview of the MIMIC model, and how it is applied to study cpMnI

A: The composition of the formative and reflexive model is displayed, importantly the indicator variables are connected through a single latent variable. ***B*****:**The formulation of the MIMIC model to explore instances of MnI is highlighted, when the path from the indicator variable X_P to Y_P is significant then an item is noninvariant. ***C*****:** The iterative purification process is displayed, at the highest level the iterative procedure stops when no more noninvariant items are detected, nonivaraint items are detected within the *for* loop, where every item is modeled with a direct path from a causal variable. After all models are trained, items which are identified as noninvaraint are removed, and the process is repeated until no items are flagged as noninvaraint.

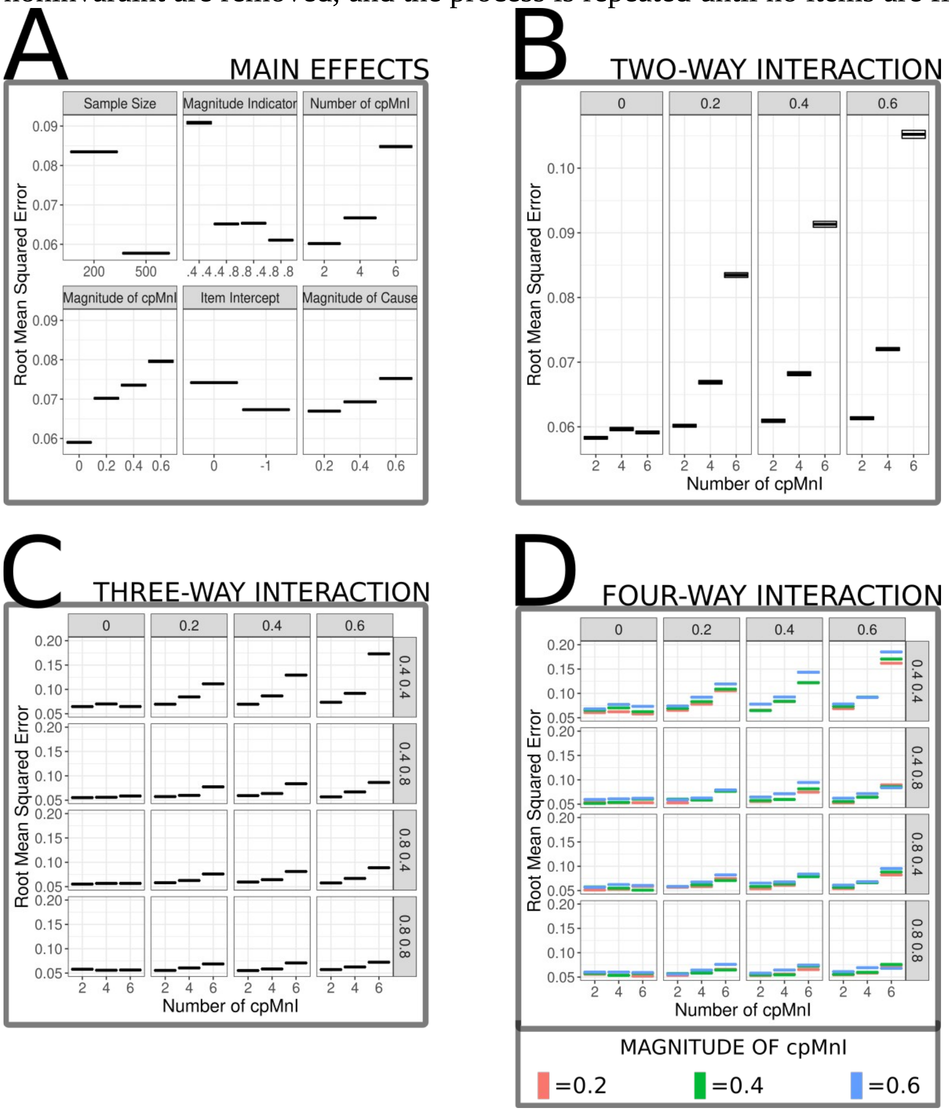


Figure 2: Estimation error from formative model ignoring instance of cpMnI

A: Mean estimation error values (+/- S.E.M.) for all main effects modeled in the ANOVA. ***B*:**Mean estimation error values (+/- S.E.M.) from the largest two-way interaction in the model which included the number of cpMnI items (x-axis) and the magnitude of cpMnI (facets). ***C*:** Mean estimation error values (+/- S.E.M.) from the largest three-way interaction in the model which included the number of cpMnI items (x-axis) and the magnitude of cpMnI (horizontal facets) and also the strength of the indicator variables (vertical facets). ***D*:** Mean estimation error values (+/- S.E.M.) from the largest four-way interaction in the model which included the number of cpMnI extends the three-way interaction to include the magnitude of the causal relationship described by the colors.

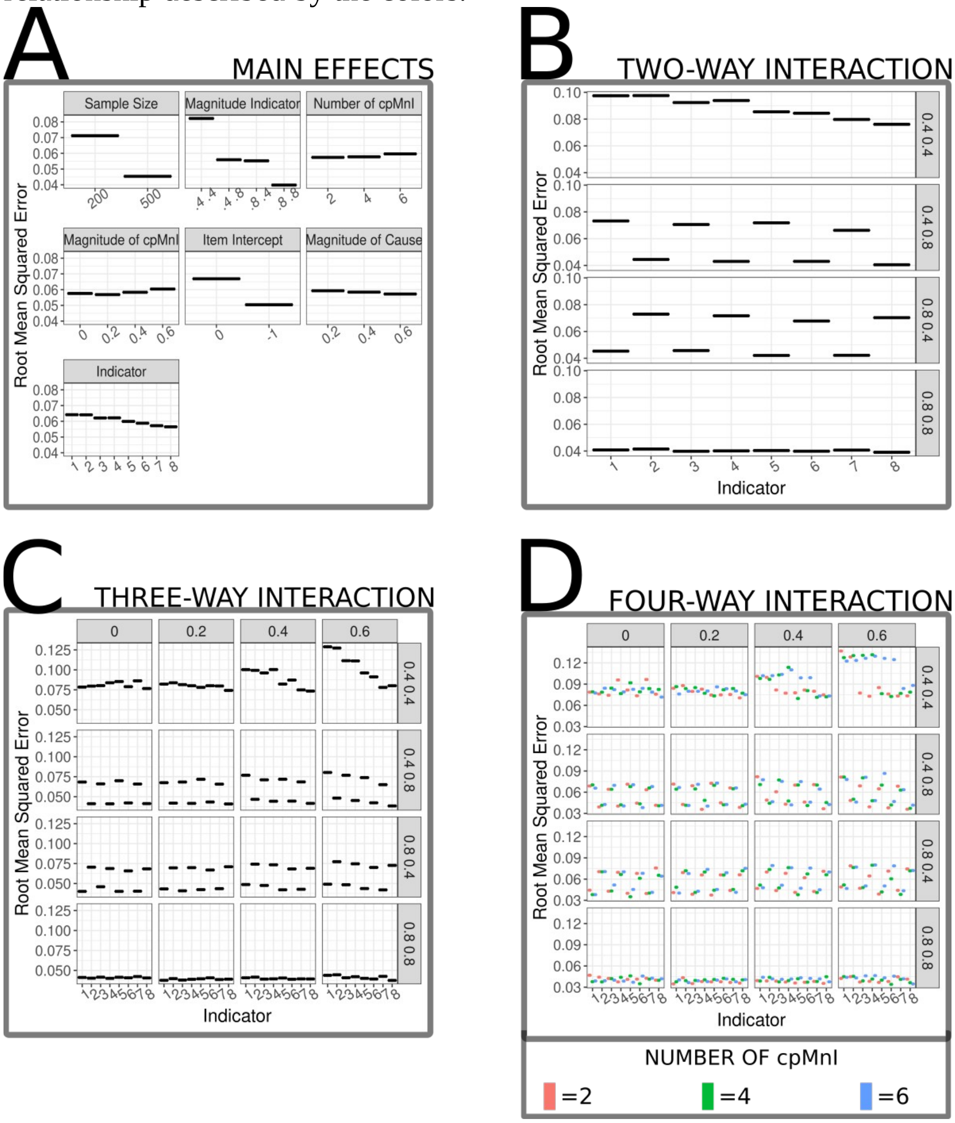


Figure 3: Estimation error from reflexive model ignoring instance of cpMnI

A: Mean estimation error values (+/- S.E.M.) for all main effects modeled in the ANOVA. ***B*****:**Mean estimation error values (+/- S.E.M.) from the largest two-way interaction in the model which included the indicator variable (x-axis) and the strength of the indicator variable (facets). ***C*****:** Mean estimation error values (+/- S.E.M.) from the largest three-way interaction in the model which included the indicator variable (x-axis) strength of the indicator variable (horizontal facets) and also the magnitude of cpMnI (vertical facets). ***D*****:** Mean estimation error values (+/- S.E.M.) from the largest four-way interaction in the model which extends the three-way interaction to include the number of cpMnI items described by the colors.

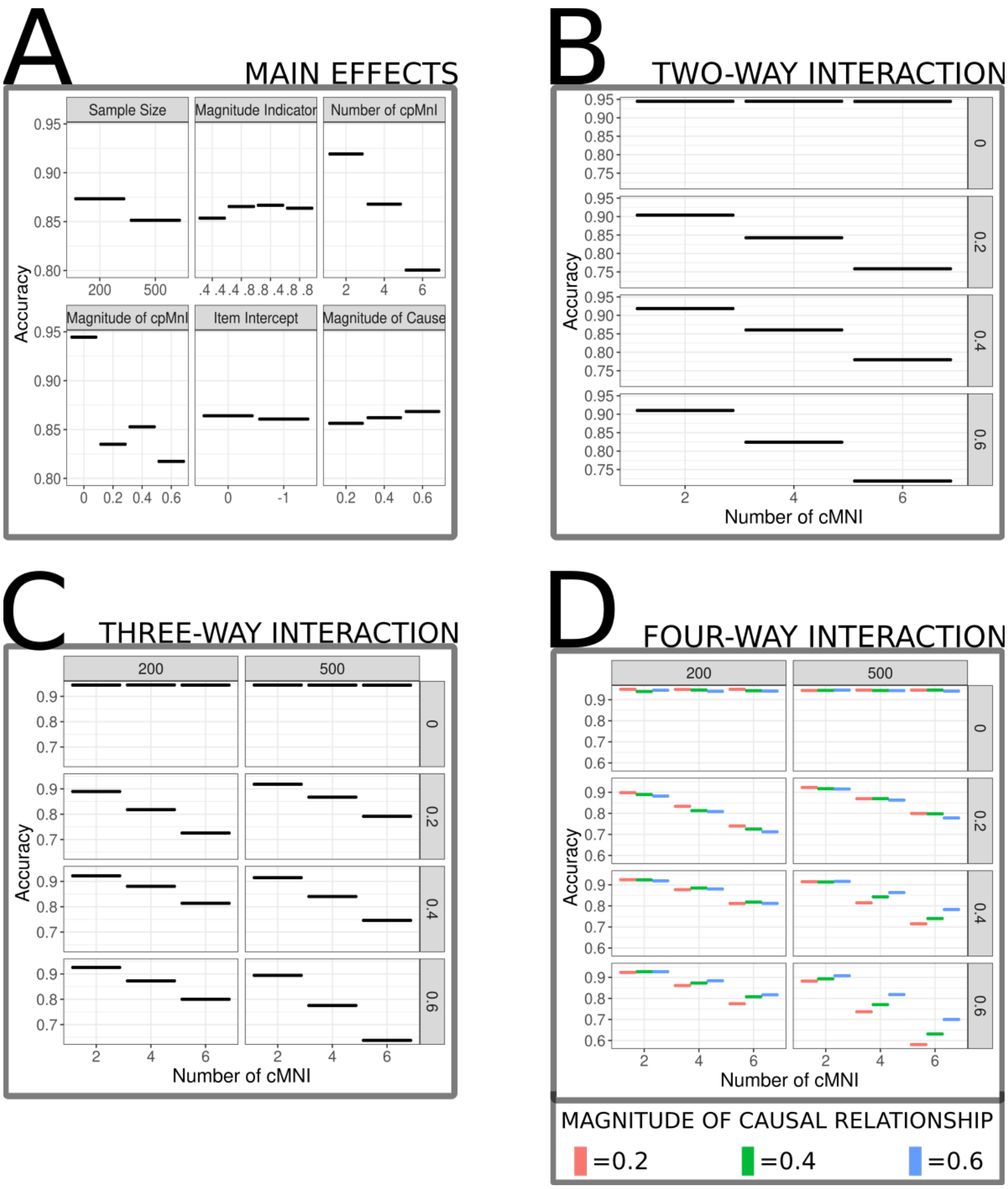


Figure 4: Model Wise accuracy for cpMnI Identification

A: Mean accuracy (+/- S.E.M.) for all main effects modeled in the ANOVA. ***B*:**Mean accuracy (+/- S.E.M.) from the largest two-way interaction in the model which included the number of cpMnI items (x-axis) and the magnitude of cpMnI (facets). ***C*:** Mean accuracy (+/- S.E.M.) from the largest three-way interaction in the model which included the number of cpMnI items (x-axis), sample size (horizontal facets), and the magnitude of cpMnI (vertical facets). ***D*:** Mean accuracy (+/- S.E.M.) from the largest four-way interaction in the model which extends the three-way interaction to include the magnitude of the causal relationship described by the colors

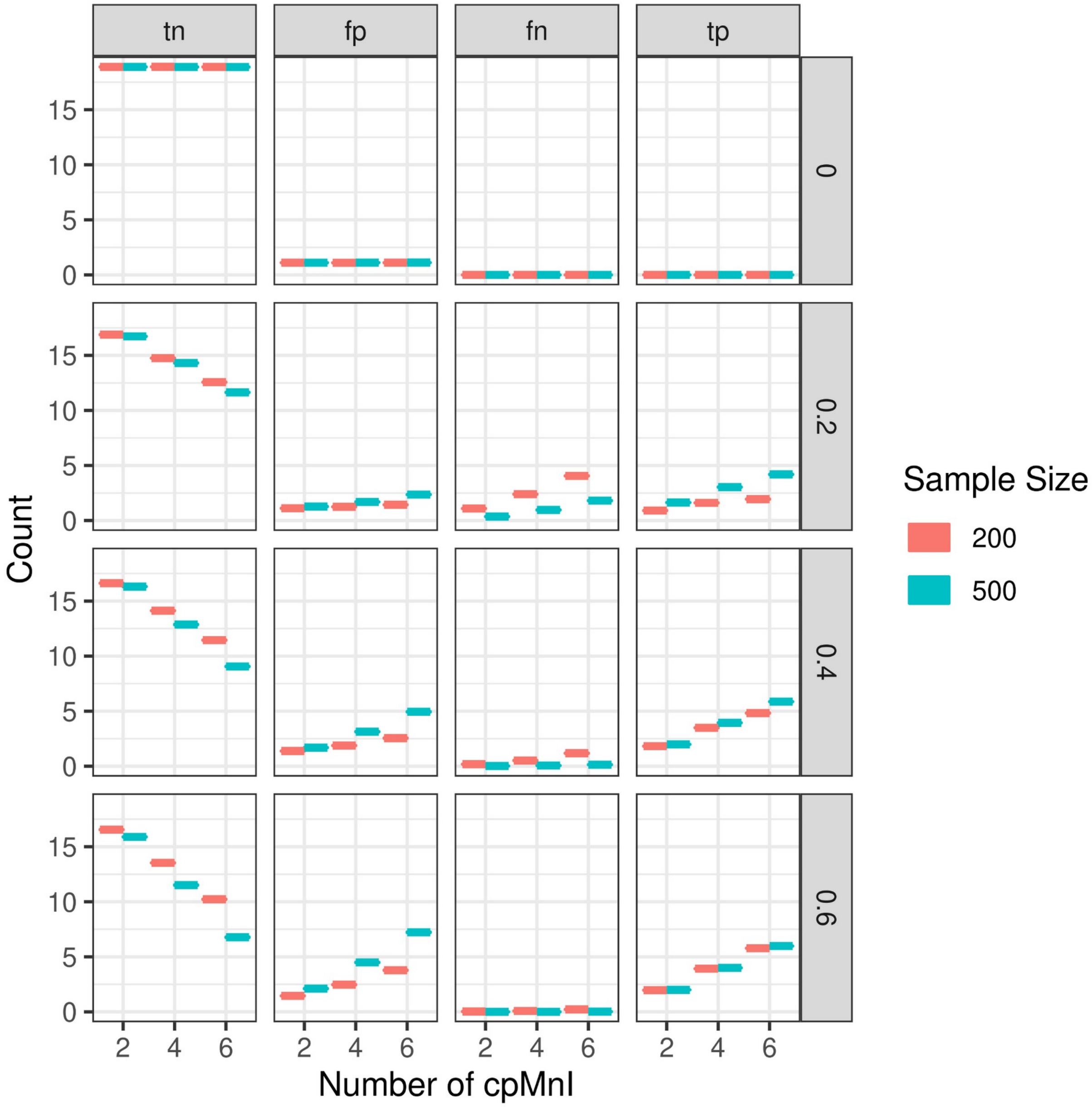


Figure 5: Model Wise cell counts for cpMnI Identification

Displayed are the true negative (tn), false positive (fp), false negative (fn), and the true positive (tp) average counts faceted by the magnitude of cpMnI. The x-axis represents the number of items which were modeled to include cpMnI, colors illustrate the sample size of the simulated datasets.

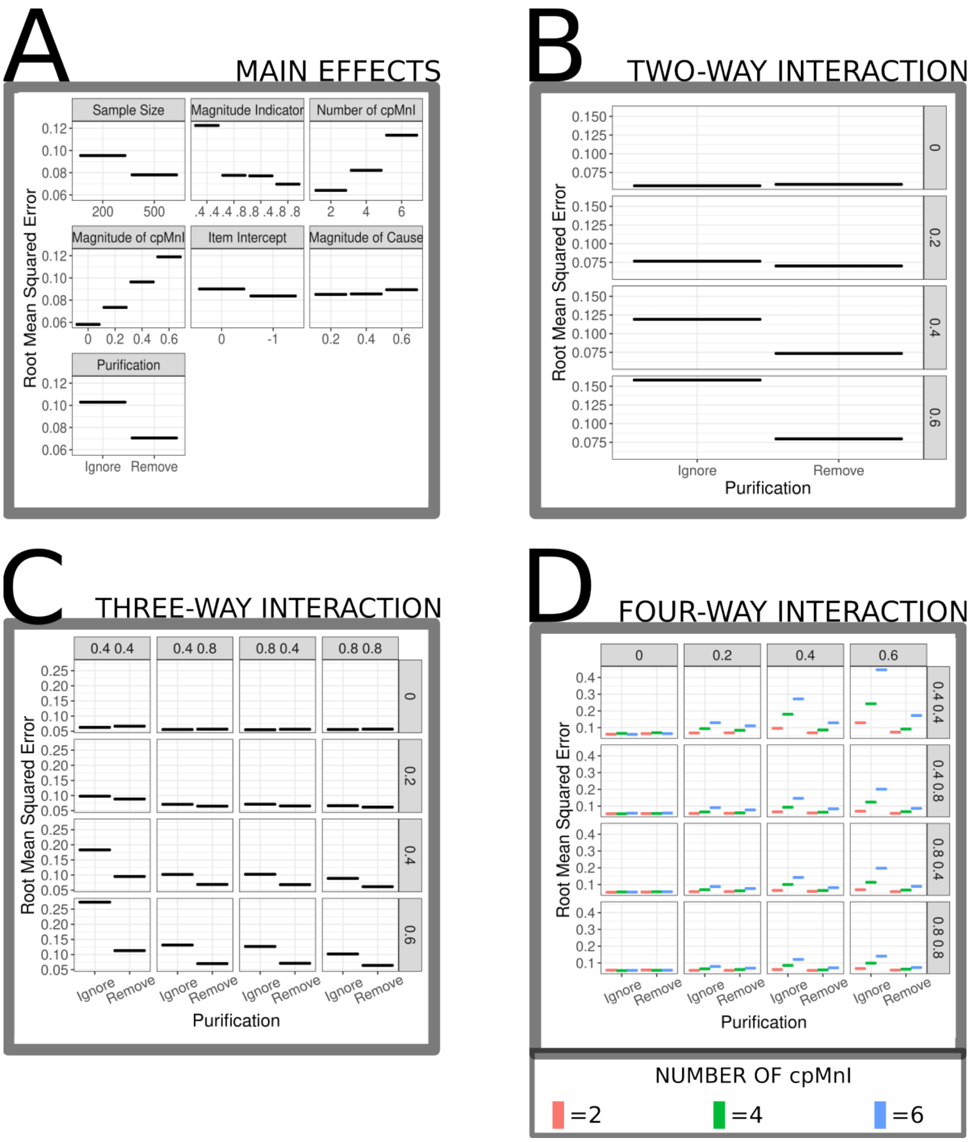


Figure 7: Comparison of estimation error from formative model between models ignoring and following the purification procedures

A: Mean estimation error values (+/- S.E.M.) for all main effects modeled in the ANOVA. ***B*****:**Mean estimation error values (+/- S.E.M.) from the largest two-way interaction in the model which included the purification status (x-axis) and the strength of the causal relationship (facets). ***C*****:** Mean estimation error values (+/- S.E.M.) from the largest three-way interaction in the model which included the purification status (x-axis) strength of the causal relationship (horizontal facets) and also the strength of the indicator variables (vertical facets). ***D*****:** Mean estimation error values (+/- S.E.M.) from the largest four-way interaction in the model which extends the three-way interaction to include the number of cpMnI items described by the colors.

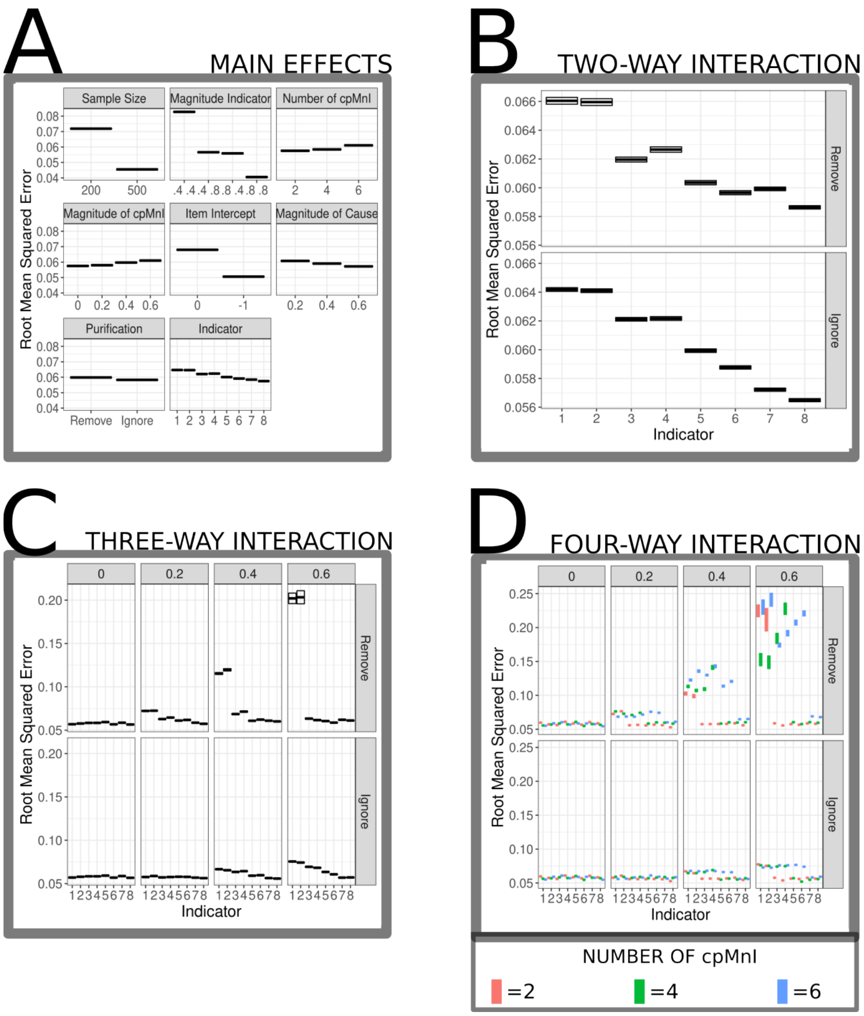


Figure 8: Comparison of estimation error from reflexive model between models ignoring and following the purification procedures

A: Mean estimation error values (+/- S.E.M.) for all main effects modeled in the ANOVA. ***B*****:**Mean estimation error values (+/- S.E.M.) from the largest two-way interaction in the model which included the indicator (x-axis) and the purification status (facets). ***C*****:** Mean estimation error values (+/- S.E.M.) from the largest three-way interaction in the model which included the indicator (x-axis) magnitude of the cpMnI (horizontal facets) and also the purification status (vertical facets). ***D*****:** Mean estimation error values (+/- S.E.M.) from the largest four-way interaction in the model which extends the three-way interaction to include the number of cpMnI items described by the colors.