AN IRT MODEL-BASED RELIABLE CHANGE INDEX WITH EMPIRICAL PRIORS: AN EXTENSION USING A MULTIPLE GROUP APPROACH WITH FINITE SAMPLE SIZES

SARAH CAMPBELL

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

October 2025

Abstract

The reliable change index (RCI; Jacobson & Truax, 1991) is a popular tool for assessing whether individuals have changed between treatments. Recently, an Item Response Theory (IRT)-based RCI that incorporates group mean information through the use of expected a posteriori (EAP) estimation has been adopted, showing promising results (Chalmers & Campbell, 2025). This paper extends the previous RCI_{IRT} work by (1) using finite sample sizes for model calibration and parameter estimation and (2) adopting a multiple-group (MG) approach to modelling sample data. Results showed that even with slight methodological changes, the results are similar to the previous studies, in that incorporating empirical priors improves rates of detecting individual change when true change is present. Larger calibration sample size has an impact on model parameter recovery, but not person parameter recovery. Finally, results favour the use of the MG approach with EAP group-informed priors when underlying group heterogeneity is expected.

Table of Contents

List of Tables

List of Figures

Introduction

The question of whether clinical patients have improved, declined, or stayed the same is addressed by tools that measure individual symptom change. Precise change measurement is a crucial yet often overlooked aspect of clinical and mental health research. Over- or under-estimation of the magnitude of change can lead to incorrect inferences, confusion around patient improvements, delayed effective treatment, and wasted clinic resources. Thus, accurate tools that guide clinicians are essential. Current measurement tools, namely, the reliable change index (RCI) by Jacobson and Truax (1991) for assessing individual change, can be inaccurate for some clinical uses. Specifically, the classical test theory (CTT) assumptions that underlie the RCI may oversimplify individual variability in that measurement error is assumed to be constant for all individuals and that their associated group membership (e.g. control vs. treatment groups, or pre-test vs. post-test groups) is often not factored into measurement, leading to less accurate conclusions.

Recent integration of the item response theory (IRT) approach with the RCI has shown promising results, overcoming the limitation of CTT's oversimplified individual variability assumption (Jabrayilov et al., 2016). As well, an IRT-based RCI that includes group-level mean priors has been shown to further reduce bias with increased precision for detecting individual change (Chalmers & Campbell, 2025). However, previous studies used the true population model when assessing individual change, which is unreflective of real-world contexts where finite sample sizes are used to calibrate the model. Additionally, the single-group estimation approach used in previous studies ignores potential differences in the clinical sample data, where heterogeneity is often expected.

In this paper, I build on the previous IRT-based RCI studies by both replicating and extending earlier work through the introduction of two methodological refinements: (1) calibrating the models with finite sample sizes instead of using the true population models for a more realistic approach, and (2) using a multiple-group (MG) approach to allow different group characteristics to be jointly estimated from sample data (e.g., as in randomized control trials). Doing so makes the simulation design more aligned with practical applications, where sample sizes are often limited and distinct differences exist in groups.

Specifically, three methods of RCI were compared using Monte Carlo simulations: the CTT-based RCI, the IRT-based RCI using weighted maximum likelihood (WML) estimation (Jabrayilov et al., 2016), and the IRT-based RCI with expected a posteriori (EAP) estimation that includes varying group mean priors (Chalmers & Campbell, 2025).

This paper is organized as follows. The introduction covers the assessment of individual change in the context of both CTT and IRT, with the focus on two IRT scoring methods: WML and EAP. Then, the current study's modified methodological approaches and the research questions are outlined. Later, the simulation study method and variables chosen are discussed, after which the results are summarized. Finally, a discussion around the results drawn from this study and future directions are presented.

Theoretical Approaches: RCI in CTT

Individual change measurements inform clinicians of whether their patient is responding to treatment, help monitor how their symptoms have changed, and aid decisions about the future course of treatment (Boswell et al., 2015). Among the approaches to assess individual change, the reliable change index (RCI; Jacobson & Truax, 1991) is the most widely used. The RCI evaluates the statistical reliability of an individual's change scores from a two-time-point, pre- and post-treatment framework, using CTT assumptions; henceforth, Jacobson and Truax's (1991) RCI will be referred to as RCI_{CTT}.

RCI is calculated by the individual's post-pre score difference divided by the pooled measurement standard error, where the threshold of significant change is often evaluated from the z-score and resulting p-value. The concept of the RCI is not new, as the formula is largely derived from the long-established CTT methodology that most researchers are familiar with (Jabrayilov et al., 2016; Maassen, 2000; McAleavey, 2024). Due to its conceptual familiarity, execution simplicity, and reasonable effectiveness for detecting individual change with insufficient data (McAleavey, 2024), RCI_{CTT} has dominated the field of assessing reliable change for the past three decades. Jacobson and Truax (1991) has over 13 thousand citations and RCI_{CTT} is a commonly required statistic by many journal editors for published clinical studies (Bauer et al., 2004).

However, several limitations come with CTT's simplified assumptions rooted in RCI_{CTT}. The next section explains RCI_{CTT}'s computation and the limitations that lead to the potential preference for IRT model-based approaches.

CTT Method for Computing RCI

RCI_{CTT} consists of two components: the post-pre score difference as the numerator and the standard error of measurement (*SEM*) as the denominator. Under the assumption that the distribution for pre-test and post-test scores are both normal, RCI_{CTT} follows a standard normal distribution, $RCI_{CTT} \sim N(0, 1)$, and can be used to test the null hypothesis that no "true" change occurred between the two measurement time points. RCI_{CTT} is calculated as follows:

$$RCI_{CTT} = \frac{X_{post} - X_{pre}}{\sqrt{SEM_{X_{post}}^2 + SEM_{X_{pre}}^2}} \quad (1)$$

$$RCI_{CTT} = \frac{X_{post} - X_{pre}}{\sqrt{2}SEM_{X_{pre}}} \quad (2)$$

where X_{post} is the individual's post-test sum score; X_{pre} is the individual's pre-test sum score; and *SEM* is the standard error of measurement of either the pre- or post-test score. Strictly speaking, both the pre- and posttest measures have their own unique *SEM*s. However, in CTT, equal measurement precision is often assumed across time, resulting in the *SEM*s being the same for both the pre- and post-test (Jabrayilov et al., 2016). If the *SEM* pre- and post-scores are treated as equal, the denominator can be computed based on the pooled pre- and post-*SEM* (i.e. $SEM^2_{pooled} = SEM^2_{X_{post}} + SEM^2_{X_{pre}}$). However, SEM_{pre} is commonly used instead of computing the pooled *SEM*, resulting in Equation 2. But if the CTT assumptions are relaxed and measurement precision is not assumed to be equal at pre- and post-test, Equation 1 may still be adopted.

The *SEM* is computed based on the pre- (or post-) test standard deviation and the reliability estimate, $SEM_{X_{pre}} = SD_{X_{pre}}\sqrt{1 - r_{xx}}$. The reliability estimate (r_{xx}) can be that of test-retest reliability (as suggested by Jacobson & Truax, 1991), coefficient alpha (used by Jabrayilov et al., 2016), McDonald's (1999) omega (used by Lee et al., 2022), or other suitable estimates. The computed RCI results in a standardized score, which can be used to obtain a *p*-value given the normal sampling distribution assumption under the null hypothesis of no change. For example, using a two-tailed test with an alpha rejection level of .05, $\left|RCI_{CTT}\right| \geq 1.96$ indicates a statistically significant change.

RCI-CTT Limitations

Since RCI_{CTT} is defined in CTT, CTT's assumptions may restrain RCI's accuracy in assessing change, presenting challenges when applied in clinical practices. Among the challenges of RCI_{CTT}, the constant standard error assumption is considered to be the primary limitation among many authors (e.g., Hays et al., 2021; Jabrayilov et al., 2016; Stratford et al., 1996). Assuming constant standard errors implies that the magnitude of variation is the same across the pre- and post-test scores and individuals. This assumption is unrealistic in practice as measurement precision is not equal across the full test range. Specifically, those who score at the extreme ends of the test have less room for variation due to floor and ceiling effects, hence skewing the sampling distribution of the measurement error.

Jabrayilov et al. (2016) demonstrated why this issue can be problematic with simulation data, showing that using constant *SEM* for all individuals across time points results in underestimating the errors at the tail-ends of the distribution (i.e. leading to conservative Type I error detection rates; rejecting less than the nominal alpha level) and overestimating the *SEM*s distributed in the middle (i.e. leading to liberal Type I error detection rates; rejecting more than the nominal alpha level). The constant *SEM* assumption is a simplified way of computing reliable change at the expense of accuracy.

Some of the other shortcomings are the population-dependent nature of CTT-based statistics (e.g., *SD* and reliability estimates depend on the sampled score distributions), which hinders the generalizability of the inferences made across one population or timepoint to the next, the violation of normal sample

assumptions when clinical data often have skewed distributions, and the limitation of two timepoints. Though some of these challenges can theoretically be overcome, for example, multiple timepoints are possible in the context of CTT (just not in Jacobson & Truax's RCI; see McAleavey's 2022 reliable trend index), they are often unnecessarily complicated and require further assumptions; hence, model-based approaches are often preferred instead as these offer more controlled ways of capturing individual and group-level changes.

Theoretical Approaches: RCI in IRT

With the above and other limitations, the CTT-based RCI by Jacobson and Truax (1991) may only be useful in a few situations (McAleavey, 2024). Thus, numerous critics have proposed alternative versions, hoping to improve the original index. Alternative versions of the RCI include those based in CTT, such as adjusting for the presence of practice effects in fields assessing cognitive change (Chelune et al., 1993) and accounting for regression to the mean effects (Hageman & Arrindell, 1999), as well as modern model-based approaches, like those based on IRT models. The IRT-based RCI was first proposed by Jabrayilov and colleagues (2016) to address the assumption of constant measurement error across individuals by allowing the measurement precision to vary by person, providing a more precise estimate of individual change.

IRT offers a robust alternative for modelling individual change. Although CTT is an easily conceptualized model, considering little information is needed, greater precision is required to capture the complexity of psychological symptoms. IRT addresses many of CTT's limitations, such as the assumption of constant measurement error and timepoint restraints, by modelling latent traits at the item level. This section covers the basics of IRT, its advantages over CTT, the computation of RCI in the context of IRT, and estimator considerations.

IRT Explained

Compared to CTT, which focuses on the total observed score, IRT uses the item as the unit of measurement. This approach allows one to tease out the item and individual characteristics that may

otherwise be concealed by CTT total scores. Specifically, IRT models the probability of an individual with a given level of latent trait endorsing an item response.

IRT is based on the assumed existence of latent traits (denoted as θ), which are variables or constructs the test aims to measure that cannot be observed but are instead inferred based on observed data. The latent trait is assumed to be continuous and often follows the assumption of monotonicity, meaning the probability of endorsing an item only increases as the latent trait increases, or vice versa (i.e., the probability of endorsement only decreases as the trait increases).

IRT consists of a series of models of varying complexity based on the inclusion of item difficulty, item discrimination, and other parameters. The model choice depends on the information one wishes to obtain, the interpretation possibilities, and the response data type (e.g. dichotomous or Likert scale data). In this study, the graded response model (GRM; Samejima, 1969) was adopted. The GRM is well-suited for Likert scale type data (i.e., ordered polytomous responses), which aligns with the understanding that psychological constructs often exist on a continuum instead of in binary categories (Preston & Reise, 2014).

IRT Strengths within RCI

IRT presents multiple attractive resolutions compared to CTT. Since each item in IRT has its own error term (Lord, 1980), the approach allows for measurement accuracy to vary across different trait levels (Magno, 2009), avoiding the naive constant error assumption and providing a more precise understanding of how well items perform across the latent trait continuum. The innate flexibility of IRT models allows multidimensional frameworks (Wang & Weiss, 2018) and the inclusion of multiple timepoints instead of just two (Morgan-Lopez, 2022). IRT can also distribute accurate weights for each item instead of the unweighted assumption from RCI_{CTT} (Morgan-Lopez, 2022). Another advantage of the IRT models is that, as long as the measures themselves are structurally equivalent, the pre- and post-tests do not have to be the same measure. This flexibility allows clinicians to draw items from the same scale to produce two tests that have different yet comparable items, potentially avoiding the influence of

practice effects. With these and other advantages, IRT has been slowly incorporated into clinical research (Thomas, 2011).

IRT Method for Computing RCI

The IRT-based RCI is a slight modification of RCI_{CTT} to accommodate unique measurement errors. As well, instead of comparing individual sum scores, IRT allows for the use of the individual's estimated latent trait, theta (denoted as $\hat{\theta}$). The IRT-based RCI formula is:

$$RCI_{IRT} = \frac{\hat{\theta}_{post} - \hat{\theta}_{pre}}{\sqrt{SE^2_{\hat{\theta}_{post}} + SE^2_{\hat{\theta}_{pre}}}} \quad (3)$$

where $\hat{\theta}$ is the individual's estimated latent trait and $SE_{\hat{\theta}}$ is the estimated standard error of $\hat{\theta}$. Since IRT trait scores are latent, one must first estimate the $\hat{\theta}$ from observed data. $\hat{\theta}$ can be estimated in several ways; popular methods include the maximum-likelihood (ML; Lord & Novick, 1968), weighted maximum-likelihood estimation (WML; Warm, 1989), and expected and maximum a posteriori estimates (EAP and MAP; Bock & Aitkin, 1981; Bock & Mislevy, 1982). The estimates of $SE_{\hat{\theta}}$ terms depend on the method used to estimate the latent traits. For instance, Jabrayilov et al. (2016) used WML to estimate the $\hat{\theta}$ and the $SE_{\hat{\theta}}$ were obtained via the Fisher information function, while Hays et al. (2021) used EAP to estimate the $\hat{\theta}$ and the $SE_{\hat{\theta}}$ were obtained via the expected a posteriori standard deviations. Nevertheless, the major difference between the RCI_{CTT} and RCI_{IRT} equations is the constant *SEM* in RCI_{CTT} and the *SE*s that vary based on the item parameters in RCI_{IRT}.

IRT Estimator Considerations

In IRT, the estimation method for $\hat{\theta}$ is a major consideration. In historical software, ML was typically the default estimation criterion: it finds the value of $\hat{\theta}$ that maximizes the probability of

observing a person's specific responses (i.e. maximizing the likelihood function, $L(u|\,\theta)$) given the response data and item parameters.

One concern with ML is that estimates at the extreme ends (i.e., people who score the lowest or highest on a scale) cannot converge, providing limited information about individuals' estimates around the tails (de Ayala, 2022). Therefore, estimators like WML or EAP are often preferred, as they are designed to handle extreme scores via weighting or incorporating prior distributions. WML is an adaptation of ML developed to reduce estimation bias (Warm, 1989) that provides finite estimates at the extreme ends. WML reduces bias by incorporating a weighted term in the estimation of $\hat{\theta}$, maximizing the weighted likelihood function, $L(u \mid \theta)\, w(\theta)$. The weighted term is defined as $w(\theta) = \sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher information function. Adding the weighted term shrinks the estimates toward where the test information is the highest (which is often the mean but not always; for example, if items are extremely difficult, the test may be more informative at higher θ values and more weight will be ascribed to those θ values).

Alternatively, EAP (Bock & Mislevy, 1982) is a Bayesian estimation method which combines the likelihood of a person's pattern of item responses with a prior specification of the population distribution, which is often the standard normal distribution, $\theta \sim N(0, 1)$. The estimates are systematically regressed towards the prior population mean, which can be useful when one has knowledge of the population's behaviour. EAP is computed as follows. Let $\hat{\theta}_{EAP}$ be the expected value of the posterior distribution, $L(u|\,\theta)$ be the likelihood of observing the set of item responses u given parameter θ, and $g(\theta)$ be the prior normal distribution, $\theta \sim N(\mu, \sigma^2)$:

$$\hat{\theta}_{EAP} = \frac{\int_{-\infty}^{\infty} \theta L(u|\theta) g(\theta) d\theta}{\int_{-\infty}^{\infty} L(u|\theta) g(\theta) d\theta} \tag{4}$$

The performance of these different estimators of $\hat{\theta}$ were not formally compared in the context of RCI_{IRT} until Jabrayilov (2016). Jabrayilov (2016) compared ML, WML, and EAP (here, the EAP prior

distribution was a standard normal). Jabrayilov concluded that when dealing with longer test lengths, EAP is preferred since its estimation is more efficient; however, ML or WML is preferred for shorter test lengths with low discrimination. Although the recommendations for estimators were inconclusive in general, Jabrayilov used WML as their primary IRT estimator for subsequent chapters. Jabrayilov et al. (2016) omitted the use of the EAP estimator and focused solely on the comparison between CTT and IRT with the WML estimator. Jabrayilov et al.'s (2016) results showed that CTT performed better with shorter test lengths of five items, while IRT should be reserved for longer test lengths of 20 items; however, the difference is minimal.

Theoretical Approaches: RCI in IRT with Group-informed Priors

The non-empirical, informative prior, $\theta \sim N(0, 1)$ used in Jabrayilov's (2016) paper, may not allow full utilization of EAP estimators. Chalmers and Campbell (2025) proposed using group-informed empirical EAP prior instead of the informative standard normal prior to better reflect the assumptions behind changes after treatment. Specifically, they allowed the mean of the prior distribution to vary, reflecting different magnitudes of mean differences between pre- and post-treatment. Their method showed better results than WML, especially when mean differences are implied in the context of clinical change. Their arguments for incorporating group mean differences as a prior are elaborated in this section.

EAP Estimation with Group-informed Priors

It should be highlighted that RCI is intended to assess individual change and not the average changes of groups. Therefore, ideally, the computation should have individuals' pre-post difference scores evaluated against their own error estimates. However, it turns out that how individual changes are evaluated is directly informed by the group-level population parameters, such as reliability and *SEM* (Chalmers & Campbell, 2025; Wang et al., 2020), which are subsequently used to estimate an individual's RCI. When only one group-level statistic is obtained from a combination of, say, "healthy", "mildly dysfunctional", and "highly dysfunctional" groups, it would result in biased estimates for individuals in the subgroups. Therefore, it is important to estimate group-level parameters separately, especially in

clinical settings where group-level changes are often assumed (e.g., the treatment group is expected to, on average, decrease in symptom severity) (Chalmers & Campbell, 2025).

Leveraging this idea, Chalmers and Campbell (2025) argued that Bayesian empirical priors, implemented with the EAP estimation method, can be used to inform these group-level changes. Since EAP naturally incorporates additional prior information, such as past experiences or assumptions (de Ayala, 2022), group-level changes, such as the mean difference between the control and treatment group or between the pre- and post-test group, can be directly incorporated. Therefore, rather than using a non-informative prior, such as the one used in the WML method with Jeffreys prior, group-mean differences are used as empirically informative priors that allow the natural assumption of post-test group means to differ in the process of obtaining $\hat{\theta}$. Specifically, the modified $RCI_{IRT-EAP}$ includes a standard normal prior $g(\theta_{pre}) \sim N(0, 1)$ on components with $\hat{\theta}_{pre}$, whereas the components with $\hat{\theta}_{post}$ included priors of $N(\mu, \sigma^2)$, where μ equals the empirical group mean differences between pre- and post-treatment groups (Chalmers & Campbell, 2025). To evaluate whether EAP with group-informed priors is superior to non-informative priors (i.e., WML), Chalmers and Campbell (2025) first consider a scenario based on true population parameters, where the group mean differences are derived from known population values. In other words, the group mean differences were not empirically estimated but artificially imposed. However, in applied settings, these group-mean difference priors are unknown and must be empirically estimated from a calibration sample.

Chalmers and Campbell (2025) showed that EAP with group-informed priors show better parameter recovery than WML, especially when the prior magnitudes are similar to how much an individual changed. Moreover, when individual change is present, even when the prior magnitudes are different from the individual changes, including non-zero priors resulted in better power. The authors concluded that EAP with group-informed priors should be used when one expects group-level differences, while WML still produces good results when no group change is present.

Limitations of the Previous Methodological Approach and Current Study

While both IRT-based RCI (i.e., WML from Jabrayilov et al., 2016 and EAP with group-informed priors from Chalmers & Campbell, 2025) showed promising results estimating individual change, there presents two notable limitations that restrict their applicability to real-world clinical contexts. These limitations relate to (1) the assumption that the model was based on true population parameters, and (2) the use of a single-group estimation approach when group differences are likely to exist.

Unrealistic Model Based on True Population Parameters

In Jabrayilov et al.'s (2016) simulation study, the true population item and person parameters were used to build the model. Chalmers and Campbell (2025) did the same to maintain methodology consistency and to ensure estimates closely reflect true parameters for the initial investigation, which is useful for understanding theoretical model behaviour. This approach implicitly assumes that the calibration process is based on an infinitely large sample size, which is unrealistic in applied contexts. In practice, one cannot estimate individual change scores with the true population model. Rather, models are calibrated with finite sample sizes, where sampling variability introduces parameter estimation bias. This fact is particularly problematic in clinical applications, where small sample sizes are common.

Simplified Assumptions Through Single-Group Estimation

In Chalmers and Campbell's (2025) simulation study, a single-group estimation approach was used to estimate the person population parameters. Doing so assumes that the sample data arise from one homogeneous population distribution, ignoring the potential differences between groups within the population. However, in clinical settings, group differences are expected. For example, in randomized control trials (RCTs), the control and the treatment groups are expected to have different improvement rates, and in clinic databases, patients of varying degrees of symptom traits have different progress trajectories. In these cases, each group has their own population parameters. Therefore, when a single-group approach is applied to model data that have group differences, individual θ estimates can be biased because the individual estimates are not estimated based on their own group's prior distribution, but instead based on a pooled prior distribution.

Current Study

This current study aims to address the above two methodological limitations by presenting a more realistic methodological approach that reflects applied clinical situations. Specifically, (1) model calibration was done on finite sample sizes instead of using true population parameters to investigate the sample size requirement for accurate parameter estimation, and (2) a MG approach was used instead of the single-group model to ensure that the underlying group differences in the sample data were correctly reflected in individual estimates.

First, this study used finite sample sizes to calibrate the IRT model. Rather than using true population parameters as priors to estimate individual change scores, the model is first calibrated on finite sample sizes, after which individual θs are estimated from these estimated model parameters for participants who were not involved in the model calibration. Since sample size affects the model parameter's estimation accuracy, the calibration sample size requirement for accurate parameter recovery should be considered. Hence, this study explored the influence of calibration sample size on obtaining accurate RCI estimates, particularly focused on smaller sample sizes, as these may better reflect the realities of clinical data availability. As well, due to IRT's large sample size requirement, how well the models fit under smaller sample sizes and whether CTT would outperform IRT in these situations can also lead to practical recommendations.

Second, this study adopts a MG approach instead of a single-group approach, in which separate population means and standard deviations are estimated for each group. Specifically, two groups (i.e., the control and treatment groups) with distinct population prior distributions, both assumed normal, were simultaneously estimated using the MG approach. Modelling each individual's θ estimates with their associated group's population distribution not only yields more accurate estimates, but also allows for directly interpretable model comparison on the same scale by placing both groups on a shared metric (Kolen & Brennan, 2014). Moreover, aside from the advantage of not needing to link separate models on the same metric post hoc for comparison, MG approaches are also useful in clinical datasets, where

comparisons might be made between groups with different sample sizes. Since the MG approach estimates all parameters jointly, the estimates are more stable than if the groups were estimated separately.

The above two methodological refinements make the simulation design more relevant to applied settings, where sample sizes are often limited and underlying group differences are expected.

Multiple Group Approach in Estimating Person-Parameters in GRM

Before moving forward, the MG approach is explained in the context of GRM, using Marginal Maximum Likelihood Estimation (MMLE) with the EM algorithm. The GRM was used to both calibrate the model and simulate the item responses. In IRT, the specified model tells the estimation algorithm the mathematical relationship between the individual latent trait and the item parameters that determine the probability of a specific response. For the GRM, this mathematical relationship that defines the response probability is approached with a series of dichotomous models. Specifically, the probability that a response y to item j in category k or higher is:

$$P_j^*(y \geq k|\theta) = \frac{1}{1+exp[-\alpha_j(\theta - b_{jk})]} \tag{5}$$

where $P_j^*(y \geq k|\theta)$ is the probability of choosing a response in category k or higher, θ is the latent trait, α is the discrimination parameter for item j, b is the threshold parameter, which represents the trait level needed for a 50% chance of responding to category k or higher for item j. There are $k - 1$ threshold parameters per item (e.g., in the case of five ordered categories, four threshold parameters will be obtained). This logistic function is similar to that of a two-parameter logistic model (2PL) when the subscript threshold category k is dropped. To obtain the probability of endorsing a specific category, the difference between the probability of endorsing category k or higher and the probability of endorsing higher than category k is (de Ayala, 2022; Thissen et al., 2001):

$$P_j(y = k|\theta) = P_j^*(y = k|\theta) - P_j^*(y = k + 1|\theta) \tag{6}$$

Extending the single-group GRM above to a multiple-group approach allows each group to be separately calibrated. That is, instead of fitting separate GRM models for each group and then linking the groups on the same scale post hoc to compare the models, for direct model comparison, the MG approach simultaneously calibrates each group by adding a subscript g to Equation 5, which allows for direct comparison across models:

$$P^{*}_{jg}(y = k|\theta_g) = \frac{1}{1+exp[-\alpha_{jg}(\theta_g - b_{jkg})]} \tag{7}$$

In this paper, we held item parameters (discrimination and threshold parameters) equal across groups and time points, but allowed the estimated latent trait distributions, $g(\theta_g)$, to vary. That is, each group g has its own population distribution:

$$\theta_g \sim N(\mu_g, \sigma^2_g) \tag{8}$$

The item parameters and the group-level hyperparameters (μ_g, σ^2_g) are estimated with MMLE using the EM algorithm. The marginal maximum likelihood function for group g person i's response vector is:

$$l(u_i|\Psi, \Phi_g) = \int P(u_i|\theta, \Psi) g(\theta|\Phi_g) d\theta \tag{9}$$

where u_i is the response pattern for person i in group g, Ψ is the set of item parameters and $\Phi_g = (\mu_g, \sigma^2_g)$ contains the hyperparameters of group g, which are the parameters of each group's latent population distribution $g(\theta|\mu_g, \sigma^2_g)$. After the group-level hyperparameters are estimated, the resulting estimated latent trait distributions $g(\theta|\hat{\mu}_g, \hat{\sigma}^2_g)$ are then used as empirical priors to estimate the individual latent trait scores using EAP. In this study, the two groups (i.e., control and treatment group), have distinct prior distributions. Specifically, the control group's informative priors follow a standard normal $N(0, 1)$ in the pre-test and post-test, while the treatment group takes on a normal distribution with varying means

of either $N(0, 1)$, $N(-\ 0.2, 1)$, $N(-\ 0.5, 1)$, or $N(-\ 0.8, 1)$ in the post-test while assuming that $N(0, 1)$ holds in the pre-test.

The EAP method is used to estimate the individual latent trait scores. Since EAP is a Bayesian estimation method, it allows the estimated hyperparameters to be included as a prior to calculate the posterior estimates of the individual latent traits. Let $\hat{\theta}_{EAP}$ be the expected value of the posterior distribution, $L(u_i|\ \theta)$ be the likelihood of observing the sample response u_i given parameter θ, and $g(\theta)$ be the prior distribution, $\theta \sim N(\mu_g, \sigma_g^2)$, where the values depend on the group association:

$$\hat{\theta}_{EAP} = \frac{\int_{-\infty}^{\infty} \theta L\left(u_i|\ \theta\right) g(\theta_g)\, d\theta}{\int_{-\infty}^{\infty} L\left(u_i|\ \theta\right) g(\theta_g)\, d\theta} \tag{10}$$

where $L(u_i|\theta)$ describes the likelihood of the individual's response pattern at each level of θ, and $g(\theta_g)$ is the prior distribution based on the hyperparameters in the MMLE stage. EAP estimation allows the estimated hyperparameters to be included as a prior to update the individual $\hat{\theta}$ estimates, whereas other estimation methods, like ML or WML, do not use priors.

Research Questions

As mentioned, this study differs from the previous studies in that sample data is first generated to calibrate the model, and then the empirically calibrated model is used to compare across RCI methods, instead of relying on the true population model for calibration. This approach enables the evaluation of how well each calibrated model performs across different estimation methods and sample sizes. Therefore, this study investigated the following research questions:

1. Which RCI method — the traditional RCI_{CTT} (Jacobson & Truax, 1991), Jabrayilov et al.'s (2016) RCI_{IRT} with WML, or Chalmers and Campbell's (2025) RCI_{IRT} with EAP informed priors — performs better in detecting individual change, as measured by Type I error rates and power rates,

under different conditions (sample size, test length, item type, prior strength and individual true change), given their empirically calibrated models?

2. How does sample size influence the MG model parameter recovery in terms of item and group mean latent trait distribution recovery?
3. How well can the MG approach discern the heterogeneity in the sample, evaluated by the recovery of the post-test group mean's latent trait distribution?
4. Which IRT-based estimation method (WML or EAP with varying priors) performs better in terms of person parameter recovery, conditioned by sample size, test length, and individual's true latent trait change?

I hypothesize that results would closely align with previous literature (Chalmers & Campbell, 2025; Jabrayilov et al., 2016), even with the methodological and factor levels changes. I expect that IRT would outperform CTT across all conditions, except when CTT's inflated Type I error rates in the middle of the theta range influence power detection (Chalmers & Campbell, 2025). For IRT-EAP models, I expect higher power than IRT-WML, better theta parameter recovery, especially when the individual's true change rates are aligned with the prior strength, and Type I error rates being more stable and slightly below the nominal alpha level across the full range of theta. Lastly, I expect calibration sample size to positively influence model and person parameter recovery, though I do expect that the smaller sample sizes would not result in optimal recoveries.

Methods

A Monte Carlo simulation study was conducted to compare the three variants of the RCI: the CTT version by Jacobson and Traux (1991), the IRT version with WML by Jabrayilov and colleagues (2016), and the IRT version with EAP by Chalmers and Campbell (2025) that was modeled with the MG approach, to determine which method performs better under different conditions (i.e., calibration sample size, test length, item type, person latent trait change, and group mean difference prior). As well, a particular focus is on the calibration sample size's role in parameter recovery and the MG approach's ability to detect and correctly classify heterogeneity in the sample.

The simulation was conducted in R (R Core Team, 2024). The SimDesign (Chalmers & Adkins, 2020) and mirt (Chalmers, 2012) packages were used to simulate the data and estimate the single and multiple-group IRT models for the respective RCI methods. See Appendix A for the complete simulation code with explanations and Appendix B for the post-analysis code with explanations: https://github.com/sarahgcampbell/MA-Thesis-Appendix.

Data Generation

The calibration sample's item response data was generated such that both the control and treatment groups had their own set of pre-test and post-test data with different population latent trait means. Specifically, both groups' population latent traits are normally distributed, with the control group's pre- and post-test group latent means set as $N(0, 1)$, and the treatment group's pre-test group latent means set as $N(0, 1)$. The treatment group's post-test group latent means varied depending on the condition's effect size magnitude $N(\mu_{T,\,post}, 1)$, where T implies the treatment group. The treatment group's post-test means are independent variables and are elaborated in the following section.

These group latent means and standard deviations (μ_g, σ_g^2), are estimated within the MMLE step in Equation 9, to produce the estimated latent trait distributions $g(\theta|\hat{\mu}_g, \hat{\sigma}_g^2)$ that were subsequently used as empirical priors in Equation 10 to estimate the individual latent trait scores. In the data generation step, the true group means were set to the prespecified values for each condition, so that during estimation, the MG model can freely estimate the group's latent trait means μ_g from the sample data, without treating the latent trait means as fixed.

In data generation, the correlation between the pre- and post-test latent trait distributions is set as 0.9 so that within an individual, their pre- and post-test latent trait scores are highly related, but still allow some individual change. Note that the empirical estimate of the correlation parameter was not included in the EAP computations (i.e., a joint or correlated EAP prior was not used in the calculation of the posterior distribution), as doing so would cause a strong mean regression effect and therefore would not allow

sufficient flexibility in estimating pre- and post-test differences. Instead, scoring was done separately for each time point. Measurement invariance was also assumed between pre- and post-test. In other words, the discrimination and difficulty parameters were equal for both groups and across pre-post time points.

Independent Variables

Five design factors were manipulated across all three methods: treatment group's post-test latent mean ($\mu_{T,\,post} = 0, -0.2, -0.5, -0.8$), model calibration sample size ($N = 100, 300, 1000$), test length (10, 20 and 50 items), item type (homogeneous and heterogeneous items), and person latent trait (θ) change between pre- and post-test ($\Delta\theta = 0, -1/3, -2/3, -1$).

The treatment group's post-test group latent means varied across four effect-size conditions: no change, small, medium, and large changes; $N(0, 1)$, $N(-\ 0.2, 1)$, $N(-\ 0.5, 1)$, $N(-\ 0.8, 1)$. Post-test group latent means were negative, indicating a decrease in symptom severity. Note that the treatment group's post-test latent means can also be conceptualized as the group latent mean differences between the pre- and post-test groups. The no-change group is included for comparison to Jabrayilov's (2016) and Hays et al.'s (2021) IRT-based RCI with EAP estimates using a prior of $N(0, 1)$.

Three levels of sample size per group were used to calibrate the IRT models: small ($N = 100$), medium ($N = 300$), and large ($N = 1000$). Sample sizes of 100 are not ideal for an IRT model's estimation precision; however, in clinical trial settings, sample sizes under 100 are quite common (Califf et al., 2012; Schuster et al., 2021). Reise and Yu (1990) suggest a sample size of at least 500 for adequate estimation precision for GRM; however, 500 is rather large for most applied settings. Kieftenbeld and Natesan (2012) also recommended 500, but a minimum of 300, as parameter recovery seems to have less of an increase between the two sample sizes; thus, 300 was decided as a medium. Sample sizes of 1000 are more appropriate for estimation precision, though more unrealistic in applied settings.

Three conditions of test length were used to reflect clinical Likert-type scale measures: 10, 20 and 50 items. The test lengths were modified slightly from the previous studies (Chalmers & Campbell, 2025; Jabrayilov et al., 2016), where 5-, 10-, and 20-item tests were used. Although shorter tests are frequently

employed to minimize test burden in clinical settings (Kruyen et al., 2013), I argue that change scores from a 5-item test are more likely to be judged subjectively by clinicians. As a result, the application of any form of RCI would typically be reserved for longer clinical test lengths. Moreover, shorter tests show large measurement error, even with high-quality items (Emons et al., 2007), putting the quality of change detection into question. In fact, since the power rates (i.e., the correct classification of true change) for 5-item tests were less than 50%, 5-item tests were not recommended for detecting change in the context of RCI (Jabrayilov et al., 2016). Some test examples that reflect the test length choice in this study are the Patient Health Questionnaire-9 with 9 items (Kroenke et al., 1999), the Beck Depression Inventory with 21 items (Beck et al., 1961), and the Outcome Questionnaire-45 with 45 items (Lambert et al., 1996).

Two sets of item parameters were used for the GRMs: homogeneous and heterogeneous item discrimination and threshold parameters. The homogeneous item set reflects tests that measure a narrower range of traits with higher discrimination, such as depression or alcohol problems (Reise & Waller, 2009). The items' discrimination parameter (*a*) was sampled from a uniform distribution between [1.5, 2.5] and the four threshold parameters (*b*) were defined as$[b_1,\ b_2,\ b_3,\ b_4] = \bar{b} + [-1, -0.5, 0.5, 1]$, where $\bar{b}$ was sampled from a uniform distribution between [0, 1.25]. The heterogeneous item sets reflect tests that measure a wider range of traits with lower discrimination, such as quality of life (Reise & Waller, 2009). The items' discrimination parameter (*a*) was sampled from a uniform distribution between [1, 2.5] and the four threshold parameters (*b*) were defined as$[b_1,\ b_2,\ b_3,\ b_4] = \bar{b} + [-0.5, -0.2, 0.2, 0.5]$, where $\bar{b}$ was sampled from a uniform distribution between [-1.5, 2.5]. The two item type parameters are the same as the ones in Jabrayilov and colleagues (2016).

Four magnitudes of person latent trait change between pre- and post-treatment were used: $\Delta\theta = 0$, -1/3, -2/3, -1, where $\Delta\theta = 0$ indicates that the person's symptom severity did not change between treatments, $\Delta\theta = -1/3$ means the person's symptoms decreased by 1/3 of the population standard deviation (*SD*), $\Delta\theta = -2/3$ means the symptoms decreased by 2/3 *SD*, and $\Delta\theta = -1$ means the symptoms decreased by a full population *SD* unit. These values were adjusted from the previous studies' levels of 0,

-0.5, -1, and -1.5 (Chalmers & Campbell, 2025; Jabrayilov et al., 2016), where a 1.5 *SD* decrease in individual symptom improvement after treatment is quite large and uncommon. For example, a review of meta-analyses of different psychiatric treatments' effect sizes reported an average medium effect size of around 0.50 (Huhn et al., 2014), and another review found a small effect size of 0.35 averaged across disorders, where medium to large effect sizes were only present for a few disorders (Leichsenring et al., 2022). Therefore, I limited the plausible range of individual change to 1 *SD*.

The design was fully crossed with three methods: CTT, IRT-WML with a single group approach, and IRT-EAP with a MG approach. In the latter, two groups were modelled: the control group with a prior of $N(0, 1)$ and the treatment group with varying group latent mean post-test priors $N(\mu_{T,post},\ 1)$. Four group latent mean priors were included for the treatment group. Additionally, three sample sizes, three test lengths, two item types, and four person latent trait changes were included. The design resulted in a total of 1152 cells.

The theta pre-test values were between -3 and +3, with 0.5 increments, totaling 13 theta levels. Each theta level had 500 item sets, resulting in 6500 item sets to be analyzed with each RCI method. True post-test theta was calculated by $\theta_{post} = \theta_{pre} +$ the person change condition. Each study cell was replicated 1000 times.

Dependent Variables

The primary dependent variable was the binary classification of statistically significant individual change (i.e., changed or not changed), where a significant change detection of a two-tailed test was set to $\alpha = .10$, which matches a one-tailed $\alpha = .05$ cutoff for improvement while still allowing the detection of deterioration in practice. This two-tailed design was adopted as opposed to the one-tail design for consistency with how the procedure is typically applied in practice, even though the simulation design assumed only improvement from individuals. In other words, an RCI score that is more extreme than $\pm\ 1.645$ is classified as changed.

The "changed" and "non-changed" samples were then aggregated across replications to calculate the detection rates (i.e., Type I error and power rates) for each condition. Type I error rates, the rate of falsely classifying the person as changed when they did not, were calculated for samples that had no true person change ($\Delta\theta = 0$). Power rates, the rate of correctly identifying a true change when one is present, were calculated for samples that had a true person change ($\Delta\theta < 0$), across three conditions ($\Delta\theta$ = -1/3, -2/3, -1).

Model parameter recovery (i.e., the item parameter recovery and the group mean recovery) for the MG approach was also assessed with root mean square deviation (RMSD) and bias. Bias was calculated as the average of the difference between the estimated parameter value and the true parameter value, whereas RMSD was calculated as the square root of the average of the squared differences between the estimated and true parameters. Note that the item discrimination and difficulty parameters were held equal across both control and treatment groups and across pre-post time points, so the observed individual change would not be influenced by changing item parameters. As well, the post-test mean for the control group was assumed to be close to $\mu_{C,post} = 0$ (C implies control group) and the post-test mean for the treatment group was assumed to be close to $\mu_{T,post}$, where $\mu_{T,post}$ depended on the condition. These constraints were placed to ensure that the evaluation of the models solely reflects their ability to recover true changes, rather than confounding effects from the item parameters.

Lastly, individual latent trait recovery was calculated for both the WML and EAP estimation methods. The IRT-WML model represents the scenario where individual changes were estimated based on a single model calibration, regardless of group association, whereas the IRT-EAP model adopted the MG approach. Therefore, two sets of item parameters were calibrated: the control and treatment groups. The control group model assumed that individuals were estimated under the assumption that the group's mean did not change across time points. The treatment group model assumed that individuals were estimated under the assumption that the group's mean improved across time points.

Results

MG Model Recovery

The MG approach's model recovery was obtained for the post-test group means and the item parameters. This first section focuses on the recovery of the true post-test group mean in terms of bias and RMSD. True post-test group means were specified for model calibration, where group mean $\mu_{T,\,post}$ is either 0, -0.2, -0.5, or -0.8.

In general, bias was small to negligible across test length and sample sizes, with the only negative bias occurring in the 50-item conditions and increasing slightly as priors increased (see Table 1). As for RMSD, the larger the sample size, the smaller the RMSD (see Table 2). However, the larger the prior, the slightly bigger the RMSD, though the difference is negligible (e.g., for a sample size of 100, when $\mu_{T,\,post} = -0.8$, RMSD = 0.08, but when $\mu_{T,\,post} = 0$, RMSD = 0.06). Test length also did not seem to affect the RMSD.

Regarding the recovery of the item parameters, bias remains small to negligible across all conditions of test length, sample size, and item type, except for the 50-item homogenous condition, where negative bias is present for the discriminant parameter ($a = -0.13$) and positive bias is present for the threshold parameters ($b_1 = 0,\ b_2 = 0.04,\ b_3 = 0.12, b_1 = 0.16$) (see Table 3). On the other hand, as sample size increases, RMSD decreases across all item parameters, with RMSD below 0.1 when sample size is 1000 and RMSD between 0.8 and 0.13 when sample size is 300 (see Table 4).

Person Latent Trait Recovery for WML and EAP

For person latent trait score recovery for WML and EAP estimators, sample size does not seem to affect parameter recovery in terms of bias (see Table 5). But as prior strength increases, bias also decreases. An informative EAP prior of $N(0, 1)$ has the worst recovery of bias around 0.13. A prior of $N(-0.2, 1)$ and $N(-0.5, 1)$ has similar recovery to that of WML.

Note that as the control group model's estimates were the same as the treatment group model's $N(0, 1)$ condition — as expected, as both adopted the same informative prior $N(0, 1)$ — only the treatment group condition's estimates were included.

As test length increases, bias also decreases. WML has lower bias than that of EAP with priors of $N(0, 1)$ and $N(-\ 0.2, 1)$, while EAP has lower bias than WML when the prior strengths are medium or large (see Table 6). In terms of person true latent trait change (i.e., $\Delta\theta$) between post- and pre-test, as $\Delta\theta$ increases in magnitude, bias also increases (see Table 7). However, when $\Delta\theta \neq 0$, including a strong prior generally decreases bias, especially when person latent trait change was medium to large (see Table 7). In general, the bias of parameter recovery for both WML and EAP methods was positive, except when $\Delta\theta = 0$ under the EAP condition and $\Delta\theta = -\ 1/3$ for EAP with a large prior, there was a slight negative bias (see Table 7).

Regarding the results for RMSD, as sample size increases, RMSD also decreases for both IRT methods, though only marginally (see Table 8). EAP has an overall lower RMSD than WML across all sample sizes and prior-strength conditions (see Table 8). As test length increases, RMSD also decreases for both IRT methods (see Table 9). Furthermore, EAP's RMSD values for 10-item tests were similar to that of WML's 20-item test, again demonstrating EAP's strength (see Table 9). Lastly, with respect to person change, as $\Delta\theta$ increases, RMSD also increases for both IRT methods; RMSD is lower for EAP compared to WML when $\Delta\theta$ is medium to large; EAP's RMSD is similar to WML when $\Delta\theta$ is small to none; and EAP's RMSD is the lowest when $\Delta\theta$ is similar to the prior strength (see Table 10).

Type I Error Rates

This section focuses on the false positive detection rates in the $\Delta\theta = 0$ condition. That is, how often each method empirically obtained statistically significant (using $\alpha = .10$) person latent trait change when no true change occurred.

Figure 1 shows the comparison between the three estimation methods: CTT, IRT-WML, and IRT-EAP with the $N(0, 1)$ informative prior. As expected, CTT has high Type I error rates in the middle

of the theta range, while WML outperforms $EAP_{N(0,1)}$ for the heterogeneous sets and the middle of the theta range for the homogeneous sets. Type I error rates for both IRT methods approach the nominal alpha level as test length increases. $EAP_{N(0,1)}$ remain slightly below the nominal alpha level across all conditions, with the homogeneous group and the longer test length groups having the best precision.

The larger the sample size, the closer the detection rates are to the nominal alpha level for the IRT-based RCIs (see Figures 2 - 4). The improvements in detection rates were most notable for WML, while EAP's detection rates remain fairly stable across all sample sizes.

For WML, a sample size of 100 seems to be too small for a stable detection rate, although its performance is still better than CTT (see Figure 2). As well, WML has detection rates slightly above the nominal alpha level in the middle of the theta range for the small sample size condition (see Figure 2). A sample size of 300, however, produces similar Type I error rates to the sample size of 1000. As well, the slightly higher detection rates seen in the middle range of the $N = 100$ condition do not occur in the $N = 300$ and $N = 1000$ conditions, where the detection behaviour was either at the nominal alpha level or just slightly inflated (see Figures 3 and 4).

Looking more closely at EAP with varying priors (see Figure 5), all values of $\mu_{T,post}$ priors have slightly lower detection rates than the nominal alpha level. For both item sets, as the test length increases, the detection rate stabilizes around the alpha level. Larger values of $\mu_{T,post}$ have better detection rates. This finding is especially apparent in shorter test length conditions, where the prior strength benefited detection. Therefore, having larger priors positively biases the detection rates.

Power Rates

This last section focuses on the power rates in the $\Delta\theta \neq 0$ condition. That is, how often each method empirically detected a significant latent trait change when true change is nonzero.

In general, power rates drastically increase with the increase in test length and person latent trait change in magnitude (see Figures 6, 7 and 8). EAP consistently outperforms WML across all conditions for larger priors (i.e., $N(-0.5, 1)$ and $N(-0.8, 1)$), with $N(-0.8, 1)$ having the most power. WML

shows similar power to the $N(-\ 0.2, 1)$ condition in EAP, and generally outperforms EAP with a standard normal prior $N(0, 1)$ for the heterogeneous item set and in the middle of the homogenous item set. Power rates were mostly less than 50% when person change was small, except for the homogeneous item sets with 50 items. Power rates are lower for heterogeneous groups compared with homogenous groups. As well, power rates only marginally improve with sample size (see Figures 9 and 10 for a comparison between sample sizes of 100 and 1000).

Discussion

This paper extended the previous two studies (Chalmers & Campbell, 2025; Jabrayilov et al., 2016), comparing RCI methods: CTT, IRT-WML, and IRT-EAP with group-informed priors, by refining the methodological approach. Specifically, finite sample sizes were used for model calibration to reflect a more realistic approach in applied settings. As well, an MG approach was used to jointly calibrate items for the control and treatment groups on a common scale, leading to improvements in item parameter precision and enabling direct cross-group comparisons.

Results showed that EAP estimation with varying informed priors still showed superiority over the WML estimation method and CTT, especially when true individual change is present. The ambiguous conclusion of whether WML or EAP was a better estimator choice discussed by Jabrayilov (2016) may be due to their use of $N(0, 1)$ informative priors. Chalmers and Campbell (2025), and again in this paper, showed that $N(0, 1)$ informative priors do not always have the best detection of significant individual change, but including medium to strong priors in the presence of true change shows the superiority of the EAP estimation method. In fact, WML has similar power to that of EAP with a group mean prior of $N(-\ 0.2, 1)$, which was also shown in Chalmers and Campbell (2025). Therefore, if medium to large effects are detectable within the sample data, then including them as empirical priors is recommended.

The Effects of Calibration Sample Size

Larger calibration sample sizes positively influenced the population mean recovery and item parameter recovery, but not the person latent trait score recovery (see Table 5 and Table 8 for person trait score recovery in terms of bias and RMSD, respectively). It is promising to know that even with smaller

sample sizes, the estimates of the individual's trait scores are closely aligned with larger sample size estimates. This result reflects Reise and Yu's (1990) study on GRM parameter recovery, where they saw that the RMSD and the correlation between true and estimated parameters had little impact when sample sizes were varied (their sample sizes were 250, 500, 1000 and 2000). They further suggested that this finding may be due to how the MMLE algorithm recovers well with small sample sizes, in that person parameters are estimated independently after item parameter estimates have been fixed through calibration (Kieftenbeld & Natesan, 2012). However, this finding was, nevertheless, a surprise. Sample size influences item parameter recovery, and subsequently has an impact on person parameter recovery as the estimated item parameters are used to estimate the person parameters. So it is odd that sample size has little impact on person parameter recovery. The influence of sample size on person parameter recovery was illustrated in several studies showing how calibration sample size reduces item and person parameter precision (i.e., larger RMSD values) and impacts overall model fit (e.g., Hulin et al., 1982; Pekmezci & Avsar, 2021). Moreover, sample size determination for IRT models is still an area of study (e.g., Schroeders & Gnambs, 2024). Therefore, the divide in findings indicates finer examinations.

Larger sample sizes for model calibration are always recommended. Though the results here showed that for Type I error rates, significance tests of the EAP-based RCI remain stable and close to the nominal alpha level across sample sizes. Impressively, because EAP model estimates were from a MG group approach, the sample size used for the model estimates was split between two groups, unlike the sample size in the WML single group approach. That is, the control group calibration had $N/2$ and the treatment group calibration also had $N/2$, while WML used the full sample size to calibrate its model. Bayesian approaches are more suitable for smaller samples due to the use of priors. Even with a smaller sample size of 300, WML can produce relatively good Type I error rates, which are similar to those of a sample size of 1000 (see Figures 2 - 4). This result is surprising because IRT is considered a large-scale modelling approach, and modest to low precision was expected. Previous studies on IRT parameter estimates suggest that there may be a point of diminishing returns as sample size increases (de Ayala, 2022; Wasserman & Braken, 2003). Although it is tempting to suggest that a sample size between 300 and

1000 would contain the point of diminishing returns for WML parameter recovery, these findings should not be considered as a generalized recommendation. It is known that sample size estimates for IRT models depend heavily on the context, such as item response type, model choice, estimation method, and research discipline (Schroeders & Gnambs, 2025); therefore, sample size requirements should be taken on a case-by-case basis.

MG Approach's Effectiveness in Estimating True Group Mean Parameters

The MG approach should only be considered useful if it is able to accurately discern the heterogeneity in the sample. To examine the accuracy and applicability of the MG approach, this study embedded four levels of group mean differences (i.e., 0, -0.2, -0.5, -0.8) within the treatment group's sample item response data, and allowed the model to estimate these group-specific hyperparameters, $g(\theta) \sim (\widehat{\mu_g}, \widehat{\sigma_g^2})$. These estimated hyperparameters are then included as priors within the estimation of post-test individual latent trait scores through EAP. The group-mean parameter recovery showed promising results. Bias was small to negligible. As the prior increased for the 50-item condition, the negative bias also slightly increased, though the largest bias was only -0.03 for a mean prior of -0.8 (see Table 1). RMSD was also small and decreased with increasing test length and sample size, with the largest RMSD being 0.09 for a mean prior of -0.8 and the smallest RMSD being 0.02 (see Table 2). The group mean's parameter recoveries were generally stable across varying magnitudes of prior strengths, for instance, with the 100 sample size and 10-item test length condition, RMSD were between 0.07 and 0.09. Overall, the results indicate that the MG approach can recover the true group mean parameters well.

The Effects of Test Length on Model Parameter Recovery

Test length has some influence over EAP's model parameter recovery. As test length increases, bias increases for the homogenous condition. This tendency is clear in Table 1, where group means recovery was only negatively biased in the 50-item condition. Similarly, for item parameter recovery in the 50-item homogeneous conditions, negative bias appeared for the discrimination parameter and positive bias for the threshold parameter (refer to Table 3). It has been suggested that when the number of

quadrature points for MMLE is too low (which results when there are many items and thus many item parameters), discrimination parameter estimates tend to be negatively biased (Johnson, 2007), which was seen in Table 3. Increasing the number of quadrature points instead of fixing the quadrature points across conditions may help decrease item parameter bias.

The Comparison Between WML and EAP Method

The MG approach with EAP estimator outperforms WML estimator in cases where medium to large individual change is present, most notably in the power detection rate results (see Figures 9 and 10). RMSD for person parameter recovery is also consistently smaller for EAP than WML, even with informed priors with the mean centred at 0 (Tables 8-10). Moreover, when the prior strength is similar to the true person ability change ($\Delta\theta$), bias and RMSD also tend to be the lowest for EAP. Therefore, I recommend using an MG approach with group-informed priors when group difference is suspected in the sample data to accurately model the underlying characteristics of the groups, so that individuals' reliable change may be correctly estimated.

Limitations and Future Studies

One area of future study is to increase the amount of sample size variations. The sample sizes in this study were deliberately chosen in hopes of better reflecting the applied settings where access to individuals' pre- and post-test data may be limited. For example, a review on clinical trials done across the mental health and medicine domains found that almost all trials had fewer than 1000 participants, and over 60% had fewer than 100 participants (Califf et al., 2012). Within the treatment of depression for randomized controlled trials, the median sample size was 106 (Schuster et al., 2021). This study only considered three sample size variations, but it could benefit from a wider selection of sample sizes. Future work could constrain the other conditions (such as item type and test length) and focus on a wider range of sample sizes, or adopt a Monte Carlo simulation-based sample size planning for tailored conditions (Chalmers, 2025; Schroeders & Gnambs, 2025).

A potential point of inquiry is why the study varied only the means in the EAP group-level prior distributions, while holding the standard deviation (*SD*) constant at 1. The *SDs* were not varied in this

study because the fully crossed cells were already large, and any *SD* values could be linearly transformed into equivalent mean values with the assumed normal distribution. For example, with a normal distribution of $\theta \sim N(\mu,\ \sigma)$, the σ could be rescaled to $SD = 1$ by dividing all values by σ, resulting in an equivalent mean of μ/σ. Therefore, even with a fixed *SD*, results still generalize to different normal distribution combinations.

IRT is, nevertheless, a modelling approach developed for large-scale educational settings, so application to the clinical context may present some unique challenges. Reise and Rodriguez (2016) outlined a few considerations, and here I highlight one that could be of interest to future studies, which is the concern of skewed scale scores and latent trait distributions for clinical constructs (Reise & Rodriguez, 2016). Psychological constructs, such as pain or depression, are arguably non-normally distributed in the population (Preston & Reise, 2014). When estimating the item parameters, normality is often assumed for the latent variable prior $g(\theta)$. Skewed priors only lead to high estimation errors for the latent trait scores (Sass et al., 2008). With skewed latent trait distributions, one may question the model's robustness, as in whether the change estimates are still reliable when latent trait distributions are skewed but modelled as if they are normal. Several IRT models have been explored to address the issue of skewed theta distributions (e.g., RC-IRT that corrects for nonnormality; Woods, 2006; or adopting a log-logistic instead of a logistic model; Lucke, 2015), though the application in RCI has yet to be explored. This issue highlights the need to thoughtfully evaluate how latent traits are conceptualized before applying IRT models to psychological data. Doing so ensures that the model assumptions align with the psychological construct.

Lastly, another area to explore is the multidimensional IRT extension of the RCI methods, as many psychological assessment inventories are believed to be inherently multidimensional (Brouwer et al., 2013; Reise, 2012; Reise et al., 2013). When unidimensional models are fitted to multidimensional data, the assumption of local independence is violated, as items are no longer only correlated through the one latent trait, but correlated between bundles of items. However, multidimensional methods have not

yet been fully explored in RCI (though see Wang & Weiss, 2018 for a recent application in multidimensional contexts).

Conclusion

This paper presented a replication and extension of previous studies that compared the three RCI methods: CTT, single group IRT-WML, and MG IRT-EAP with varying informative priors, to examine the calibration sample size for accurate parameter recovery and the utility of the MG approach with EAP estimations. Model parameter recovery is most affected by the sample size and test length, though person parameter recovery was generally unaffected by sample size. The literature on person parameter recovery's relationship with sample size is mixed; therefore, it warrants further studies that are context dependent. Furthermore, results showed that even a sample size of 300 can produce similar Type I error rates as those from a sample of 1000, although larger sample sizes are always recommended. Type I error rates were mostly affected by sample size and test length, whereas power depended mainly on test length and person change. The MG model showed great parameter recovery results and is recommended for clinical applications where underlying group heterogeneity is often embedded in the sample. Finally, EAP is preferable when medium to large individual changes are expected, whereas WML may be more suitable when no change is assumed. However, in clinical contexts, change is typically anticipated.

References

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, *82*(1), 60–70. https://doi.org/10.1207/s15327752jpa8201_11

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561-571.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/bf02293801

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444. https://doi.org/10.1177/014662168200600405

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research*, *25*(1), 6-19. https://doi.org/10.1080/10503307.2013.817696

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory–II: G is the key. *Psychological Assessment*, *25*(1), 136–145. https://doi.org/10.1037/a0029228

Califf, R. M., Zarin, D. A., Kramer, J. M., Sherman, R. E., Aberle, L. H., Tasneem, A. (2012). Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *JAMA*, *307*(17):1838–1847. doi:10.1001/jama.2012.3424

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2025). Spower: A General-Purpose Monte Carlo Simulation Power Analysis Program. [Manuscript submitted for publication].

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chalmers, R. P., & Campbell, S. (2025). *Including Empirical Prior Information in the Reliable Change Index*. *Applied Psychological Measurement*, *0*(0), 1-17. https://doi.org/10.1177/01466216251358492

Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*(1), 41–52. https://doi.org/10.1037/0894-4105.7.1.41

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Second edition. Routledge. https://doi.org/10.4324/9780203771587

de Ayala, R. J. (2022). *The theory and practice of item response theory*. Second edition. New York, NY: Guilford Publications.

Emons, W. H., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological methods*, *12*(1), 105-120. https://doi.org/10.1037/1082-989X.12.1.105

Hageman, W. L., & Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level analysis. *Behaviour Research and Therapy*, *37*(12), 1169-1193. https://doi.org/10.1016/s0005-7967(99)00032-7

Hays, R. D., Spritzer, K. L., & Reise, S. P. (2021). Using item response theory to identify responders to treatment: Examples with the patient-reported outcomes measurement information system (PROMIS.) physical function scale and emotional distress composite. *Psychometrika*, *86*(3), 781–792. doi: 10.1007/s11336-021-09774-1

Huhn, M., Tardy, M., Spineli, L. M., Kissling, W., Förstl, H., Pitschel-Walz, G., Leucht, C., Samara, M., Dold, M., Davis, J. M., & Leucht, S. (2014). Efficacy of pharmacotherapy and psychotherapy for adult psychiatric disorders: a systematic overview of meta-analyses. *JAMA psychiatry*, *71*(6), 706–715. https://doi.org/10.1001/jamapsychiatry.2014.112

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study. *Applied Psychological Measurement*, *6*(3), 249-260. https://doi.org/10.1177/014662168200600301

Jabrayilov, R. (2016). *Improving individual change assessment in clinical, medical, health psychology* [Doctoral dissertation]. Tilburg University.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559-572. https://doi.org/10.1177/0146621616664046

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037//0022-006x.59.1.12

Johnson, M. S. (2007). Marginal Maximum Likelihood Estimation of Item Response Models in R. *Journal of Statistical Software*, *20*(10), 1–24. https://doi.org/10.18637/jss.v020.i10

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*(5), 399–419. https://doi.org/10.1177/0146621612446170

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices (3rd ed.). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-0317-7

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (1999). Patient Health Questionnaire-9 (PHQ-9) [Database record]. *APA PsycTests*. https://doi.org/10.1037/t06165-000

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13*(3), 223-248. https://doi.org/10.1080/15305058.2012.703734

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy: An International Journal of Theory and Practice*, *3*(4), 249-258.

Lee, M. K., Peipert, J. D., Cella, D., Yost, K. J., Eton, D. T., Novotny, P. J., Sloan, J. A., & Dueck, A. C. (2022). Identifying meaningful change on PROMIS short forms in cancer patients: A comparison of item response theory and classic test theory frameworks. *Quality of Life Research*, *32*(5), 1355–1367. https://doi.org/10.1007/s11136-022-03255-3

Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. P. A. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: an umbrella review and meta-analytic evaluation of recent meta-analyses. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, *21*(1), 133–145. https://doi.org/10.1002/wps.20941

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge. https://doi.org/10.4324/9780203056615

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading.

Lucke, J. F. (2015). Unipolar item response models. In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (ed. S. P. Reise and D. Revicki), pp. 272–284. Routledge: New York.

Maassen, G. H. (2000). Principles of defining reliable change indices. *Journal of Clinical and Experimental Neuropsychology*, *22*(5), 622-632. https://doi.org/10.1076/1380-3395(200010)22:5;1-9;FT622

Maydeu-Olivares, A. (2021). Assessing the accuracy of errors of measurement. Implications for assessing reliable change in clinical settings. *Psychometrika*, *86*(3), 793-799. doi: 10.1007/s11336-021-09806-w

McAleavey, A. A. (2022). R*eliableTrendIndex: Utilities for (more reliable) reliable change analysis*. In https://github.com/andrewmcaleavey/ReliableTrendIndex

McAleavey, A. A. (2024). When (not) to rely on the reliable change index: A critical appraisal and alternatives to consider in clinical psychology. *Clinical Psychology: Science and Practice*. https://doi.org/10.1037/cps0000203

McDonald, R.P. (1999). *Test Theory: A Unified Treatment* (1st ed.). In *Psychology Press eBooks*. https://doi.org/10.4324/9781410601087

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *Social Science Research Network*. *1*(1), 1-11. http://files.eric.ed.gov/fulltext/ED506058.pdf

Morgan-Lopez, A. A., Saavedra, L. M., Ramirez, D. D., Smith, L. M., & Yaros, A. C. (2022). Adapting the multilevel model for estimation of the reliable change index (RCI) with multiple timepoints and multiple sources of error. *International journal of methods in psychiatric research*, *31*(2), e1906. https://doi.org/10.1002/mpr.1906

Pekmezci, F. B., & Avşar, A. Ş. (2021). A guide for more accurate and precise estimations in simulative unidimensional IRT models. *International Journal of Assessment Tools in Education*, *8*(2), 423-453. https://doi.org/10.21449/ijate.790289

Preston, K. S. J., & Reise, S. P. (2014). Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement*, *74*(3), 377-399. https://doi.org/10.1177/0013164413507063

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges. *Psychological medicine*, *46*(10), 2025–2039. https://doi.org/10.1017/S0033291716000520

Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative approaches to addressing non-normal distributions in the application of IRT models to personality measures. *Journal of Personality Assessment*, *100*(4), 363–374. https://doi.org/10.1080/00223891.2017.1381969

Reise, S. P., Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Reise, S. P., Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*. *27*(2), 133-144. https://doi.org/10.1111/j.1745-3984.1990.tb00738.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, *34*(17), 1–97. https://doi.org/10.1007/bf03372160

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, *21*(1), 65–88. https://doi.org/10.1080/08957340701796415

Schroeders, U., Gnambs, T. (2025). Sample-size planning in item-response theory: A tutorial. *Advances in Methods and Practices in Psychological Science*, *8*(1), https://doi.org/10.1177/25152459251314798

Schuster, R., Kaiser, T., Terhorst, Y., Messner, E. M., Strohmeier, L. M., & Laireiter, A. R. (2021). Sample size, sample size planning, and the impact of study context: systematic review and recommendations by the example of psychological depression treatment. *Psychological medicine*, *51*(6), 902–908. https://doi.org/10.1017/S003329172100129X

Stratford, P. W., Binkley, J., Solomon, P., Finch, E., Gill, C., & Moreland, J. (1996). Defining the minimum level of detectable change for the Roland-Morris questionnaire. *Physical Therapy*, *76*(4), 359–365. https://doi.org/10.1093/ptj/76.4.359

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item Response Theory for Item Scored in More Than Two Categories. In Thissen, D., & Wainer, H. (Eds.), *Test scoring*. (pp.141-184). Lawrence Erlbaum Associates Publishers.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A Review. *Assessment*, *18*(3), 291-307. https://doi.org/10.1177/1073191110374797

Wang, C., & Weiss, D. J. (2018). Multivariate hypothesis testing methods for evaluating significant individual change. *Applied Psychological Measurement*, *42*(3), 221–239. https://doi.org/10.1177/0146621617726787

Wang, C., Weiss, D. J., & Suen, K. Y. (2020). Hypothesis testing methods for multivariate multi-occasion intra-individual change. *Multivariate Behavioral Research*, *56*(3), 459–475. https://doi.org/10.1080/00273171.2020.1730739

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/bf02294627

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology. Assessment psychology* (2nd ed., Vol. 10, pp. 50-81). Hoboken, NJ: John Wiley & Sons.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, *11*(3), 253–270. https://doi.org/10.1037/1082-989X.11.3.253

Tables and Figures

Table 1

Bias for the Post-test Group Mean Between the True Parameters and the Calibrated Model's Estimates Grouped by Test Length and Sample Size

Conditions		EAP Estimators				
Sample Size	Test Length	Control $N(0, 1)$	Treatment $N(0, 1)$	Treatment $N(-\ 0.2, 1)$	Treatment $N(-\ 0.5, 1)$	Treatment $N(-\ 0.8, 1)$
100	10	0.00	0.00	0.00	0.00	-0.01
100	20	0.00	0.00	0.00	0.00	0.00
100	50	0.00	0.00	-0.01	-0.02	-0.02
300	10	0.00	0.00	0.00	0.00	0.00
300	20	0.00	0.00	0.00	0.00	0.00
300	50	0.00	0.00	-0.01	-0.02	-0.03
1000	10	0.00	0.00	0.00	0.00	0.00
1000	20	0.00	0.00	0.00	0.00	0.00
1000	50	0.00	0.00	-0.01	-0.02	-0.03

Note. The control group model was estimated with an informative prior $N(0, 1)$, which is the same as the first treatment group model $N(0, 1)$. The other three treatment group models represent informative priors with increasing levels of post-pre mean difference magnitude.

Table 2

RMSD for the Post-test Group Mean Between the True Parameters and the Calibrated Model's Estimates Grouped by Test Length and Sample Size

Conditions		EAP Estimators				
Sample Size	Test Length	Control $N(0,1)$	Treatment $N(0,1)$	Treatment $N(-0.2,1)$	Treatment $N(-0.5,1)$	Treatment $N(-0.8,1)$
100	10	0.07	0.07	0.07	0.08	0.09
100	20	0.06	0.06	0.06	0.07	0.08
100	50	0.05	0.05	0.06	0.07	0.08
300	10	0.04	0.04	0.04	0.05	0.05
300	20	0.03	0.03	0.04	0.04	0.05
300	50	0.03	0.03	0.03	0.04	0.06
1000	10	0.02	0.02	0.02	0.03	0.03
1000	20	0.02	0.02	0.02	0.02	0.03
1000	50	0.02	0.02	0.02	0.03	0.04

Note. The control group model was estimated with an informative prior $N(0,1)$, which is the same as the first treatment group model $N(0,1)$. The other three treatment group models represent informative priors with increased levels of post-pre mean difference magnitude.

Table 3

Bias for the EAP Item Parameters Recovery

Condition		Item Parameters				
Item Type	Test Length	a	b_1	b_2	b_3	b_4
hetero	10	0.01	0.00	0.00	0.00	0.01
hetero	20	0.01	0.00	0.00	0.00	0.00
hetero	50	0.00	0.00	0.00	0.00	0.01
homo	10	0.01	0.00	0.00	0.00	0.01
homo	20	0.00	0.00	0.00	0.01	0.01
homo	50	-0.13	0.00	0.04	0.12	0.16

Note. 'hetero' represents the heterogeneous item type with a wider range of traits with lower discrimination. 'homo' represents the homogenous item type with a narrower range of traits with higher discrimination. The a parameter is the item discrimination parameter and b is the item threshold parameter.

Table 4

RMSD for the EAP Item Parameters Recovery

Condition	Item Parameters				
Sample Size	a	b_1	b_2	b_3	b_4
100	0.21	0.13	0.14	0.17	0.20
300	0.13	0.08	0.08	0.10	0.12
1000	0.08	0.05	0.05	0.06	0.07

Note. The a parameter is the item discrimination parameter, that is, how well the items can differentiate between similar individuals with similar latent trait levels. The b parameters is the item threshold parameter for the GRM with 5 categories.

Table 5

Person Parameter Recovery for WML and EAP in terms of Bias with respect to Sample Size

Condition	Person Parameters				
Sample Size	WML	$EAP_{N(0,1)}$	$EAP_{N(-0.2,1)}$	$EAP_{N(-0.5,1)}$	$EAP_{N(-0.8,1)}$
100	0.07	0.13	0.1	0.06	0.02
300	0.08	0.13	0.1	0.06	0.02
1000	0.08	0.13	0.1	0.06	0.02

Note. The EAP estimates were all from the treatment group model. The control group model estimates were not included as the values were exactly the same as the informative condition, $N(0, 1)$.

Table 6

Person Parameter Recovery for WML and EAP in terms of Bias with respect to Test Length

Condition	Person Parameters				
Test Length	WML	$EAP_{N(0,1)}$	$EAP_{N(-0.2,1)}$	$EAP_{N(-0.5,1)}$	$EAP_{N(-0.8,1)}$
10	0.11	0.18	0.14	0.08	0.02
20	0.08	0.13	0.11	0.07	0.03
50	0.04	0.07	0.06	0.04	0.01

Table 7

Person Parameter Recovery for WML and EAP in terms of Bias with respect to Person True Change

Condition	Person Parameters				
Person Change $\Delta\theta$	WML	$EAP_{N(0,1)}$	$EAP_{N(-0.2,1)}$	$EAP_{N(-0.5,1)}$	$EAP_{N(-0.8,1)}$
-1	0.16	0.27	0.24	0.20	0.15
-2/3	0.10	0.17	0.14	0.10	0.06
-1/3	0.04	0.08	0.05	0.01	-0.03
0	0.00	0.00	-0.02	-0.06	-0.10

Table 8

Person Parameter Recovery for WML and EAP in terms of RMSD with respect to Person Sample Size

Condition	Person Parameters				
Sample Size	WML	$EAP_{N(0,1)}$	$EAP_{N(-0.2,1)}$	$EAP_{N(-0.5,1)}$	$EAP_{N(-0.8,1)}$
100	0.50	0.42	0.41	0.40	0.40
300	0.48	0.42	0.41	0.40	0.40
1000	0.48	0.41	0.41	0.40	0.40

Table 9

Person Parameter Recovery for WML and EAP in terms of RMSD with respect to Person Test Length

Condition	Person Parameters				
Test Length	WML	$EAP_{N(0,1)}$	$EAP_{N(-0.2,1)}$	$EAP_{N(-0.5,1)}$	$EAP_{N(-0.8,1)}$
10	0.61	0.51	0.50	0.49	0.48
20	0.49	0.42	0.41	0.41	0.40
50	0.36	0.32	0.32	0.31	0.31

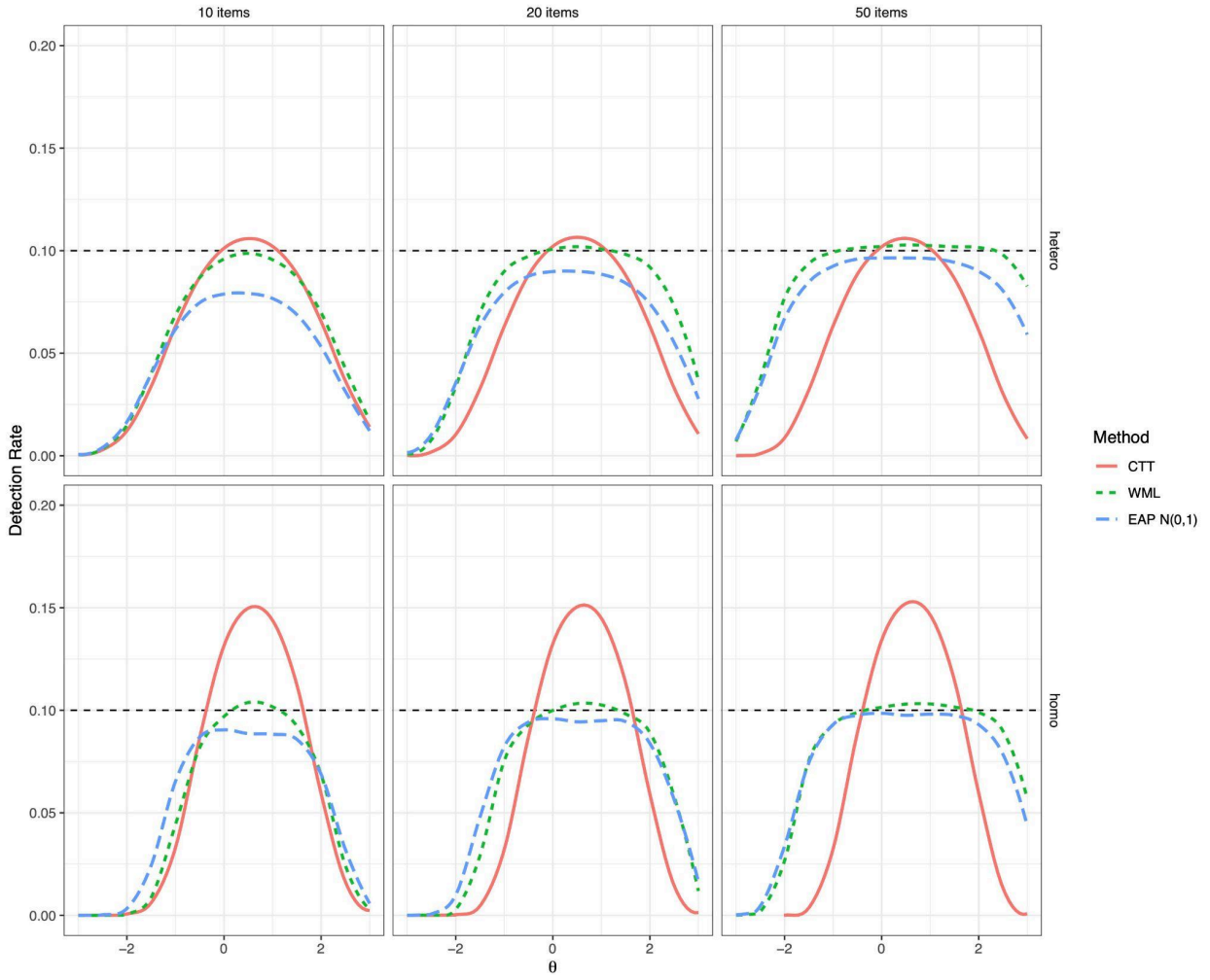
Table 10

Person Parameter Recovery for WML and EAP in terms of RMSD with respect to Person True Change

Condition	Person Parameters				
Person Change $\Delta\theta$	WML	$EAP_{N(0,1)}$	$EAP_{N(-0.2,1)}$	$EAP_{N(-0.5,1)}$	$EAP_{N(-0.8,1)}$
-1	0.54	0.50	0.49	0.47	0.45
-2/3	0.43	0.43	0.42	0.40	0.39
-1/3	0.38	0.38	0.37	0.37	0.37
0	0.36	0.36	0.36	0.37	0.39

Figure 1

Type I Error Rates for CTT, IRT with WML Estimator, and IRT with EAP $N(0, 1)$ *Informative Estimator Compared Across Item Type (Homogenous or Heterogenous) and Test Length*



Note. The dashed horizontal black line is the nominal $\alpha = .10$ rate.

Figure 2

Type I Error Rates for CTT and IRT (WML and EAP) for a Sample Size of 100

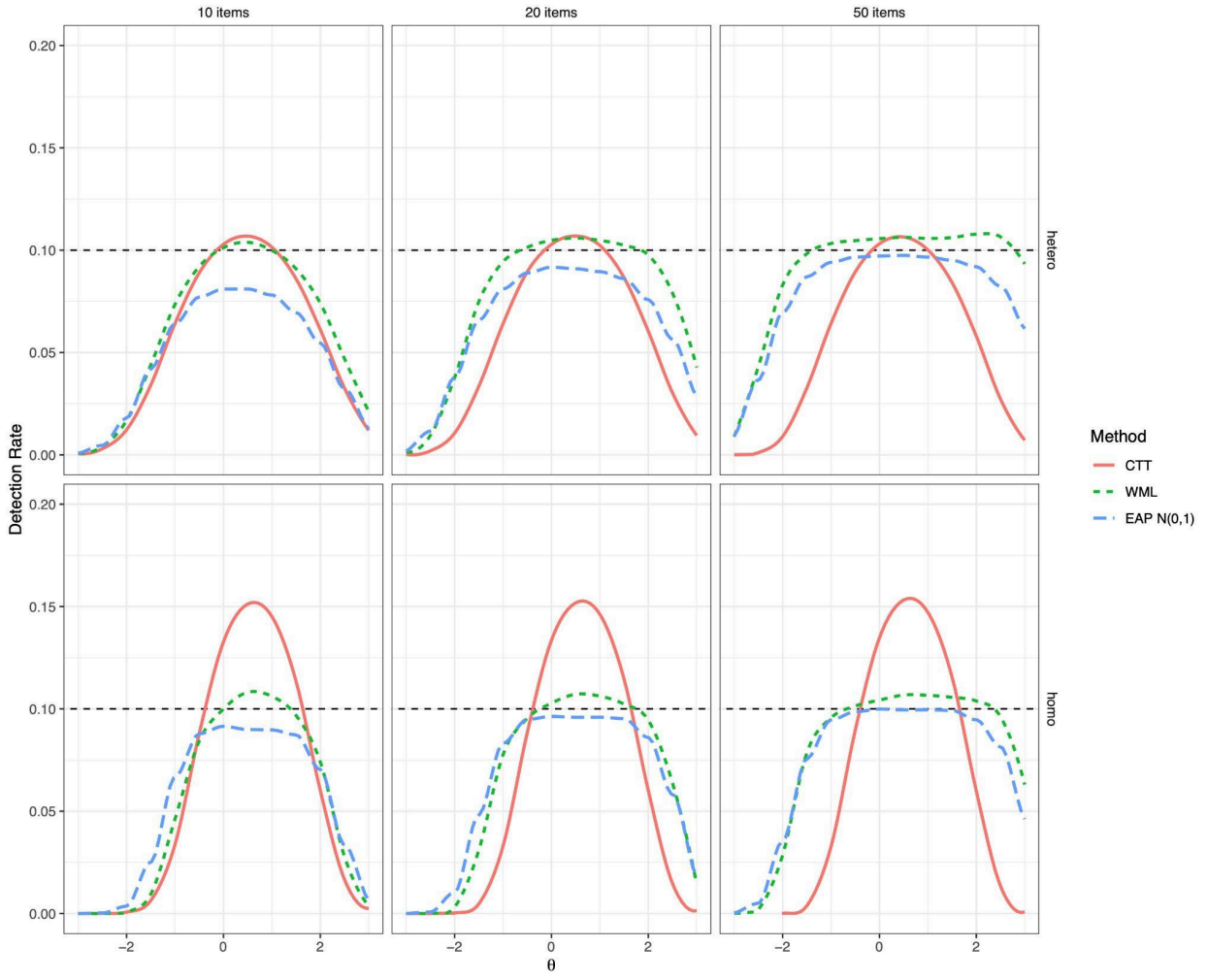


Figure 3

Type I Error Rates for CTT and IRT (WML and EAP) for a Sample Size of 300

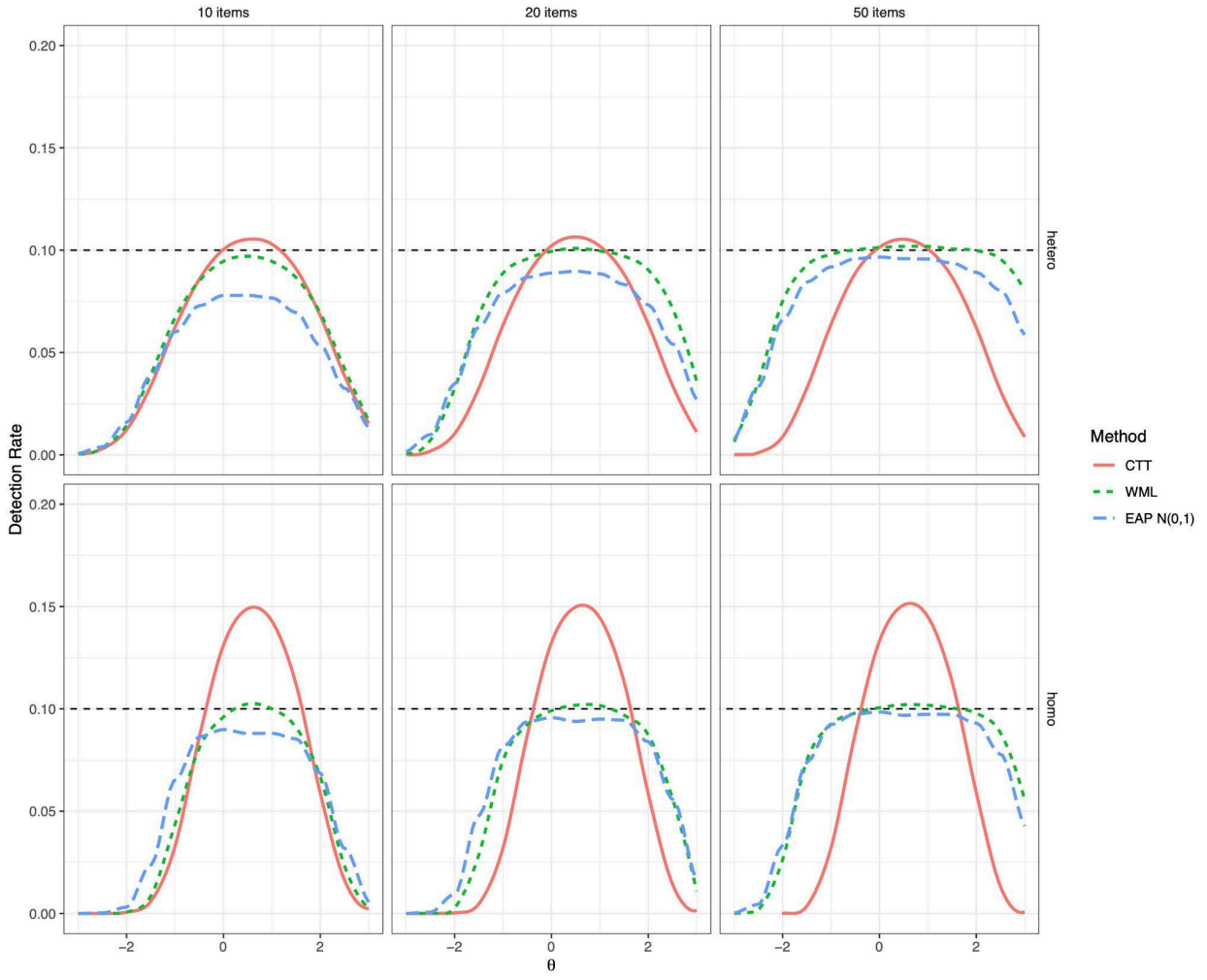


Figure 4

Type I Error Rates for CTT and IRT (WML and EAP) for a Sample Size of 1000

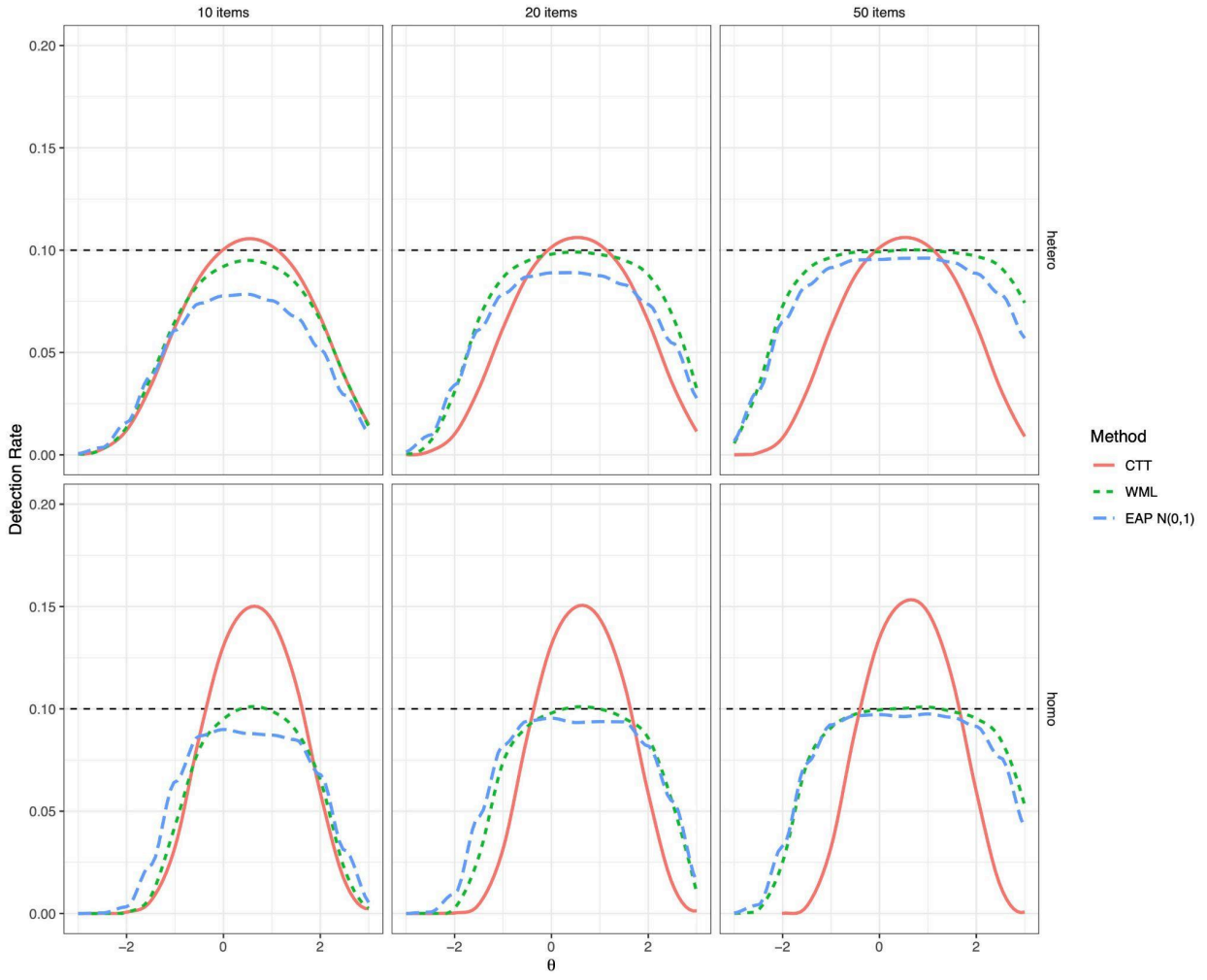


Figure 5

Type I Error Rates for EAP with Varying Magnitudes of Priors

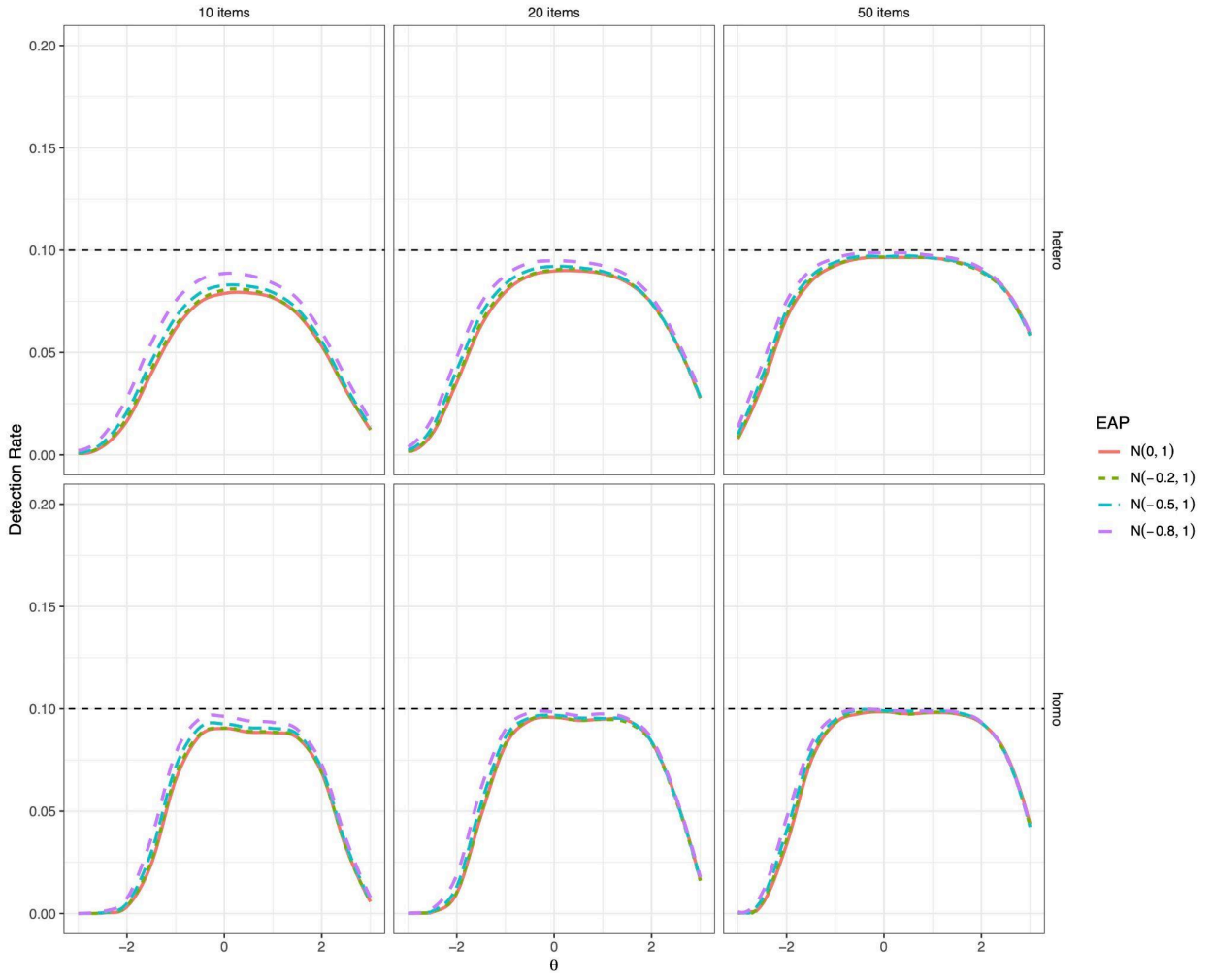


Figure 6

Power Rates for the 10-item Test

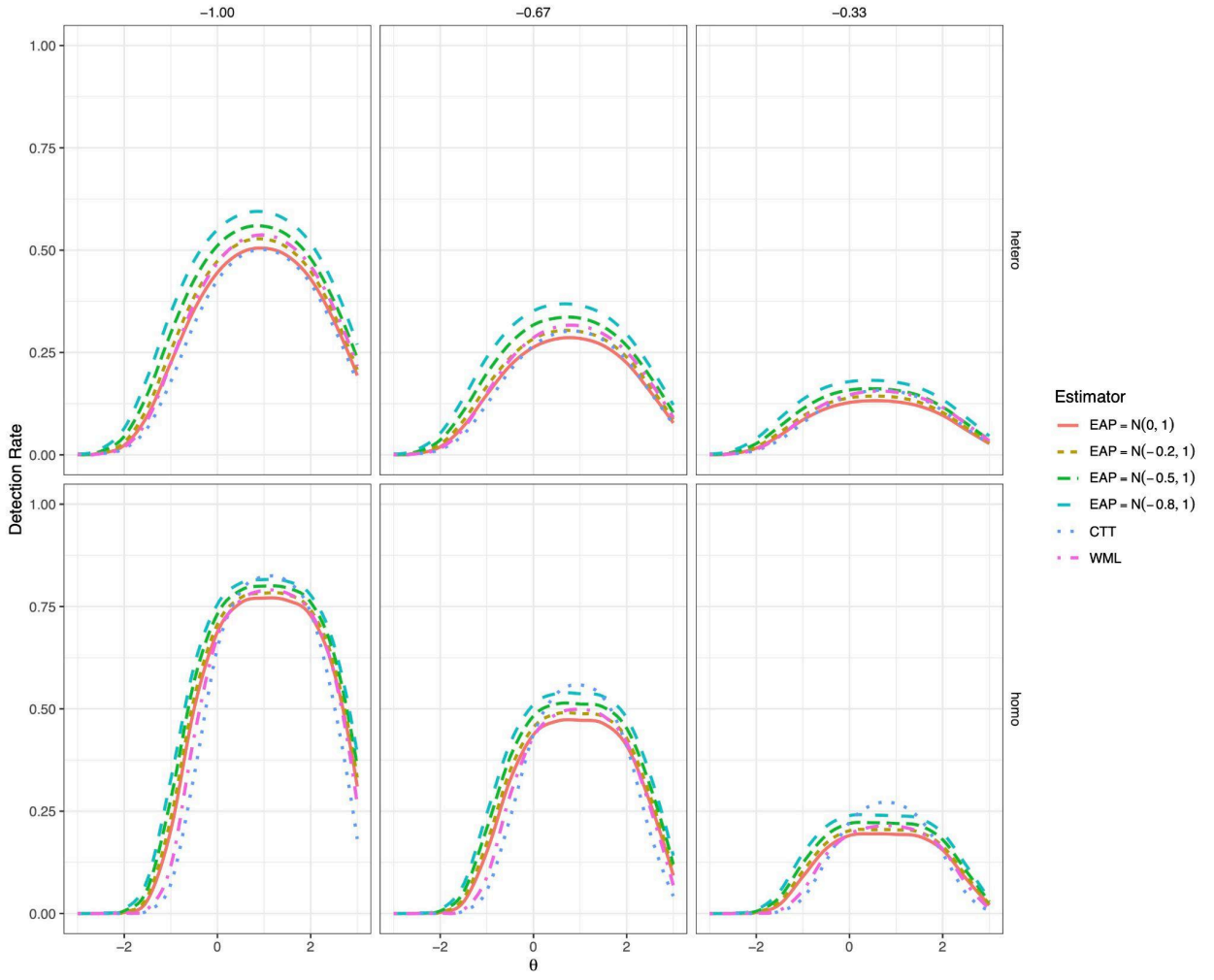


Figure 7

Power Rates for the 20-item Test

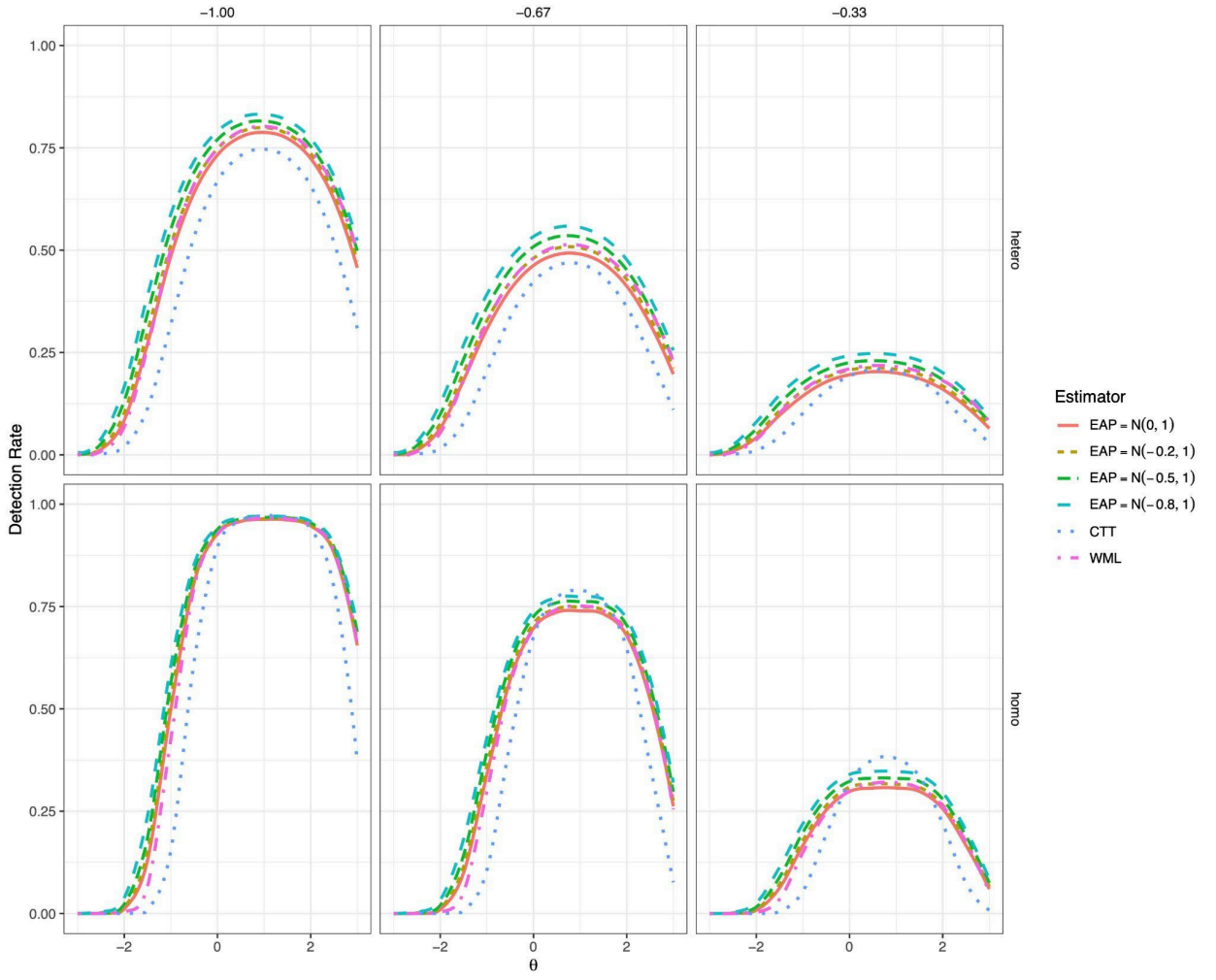


Figure 8

Power Rates for the 50-item Test

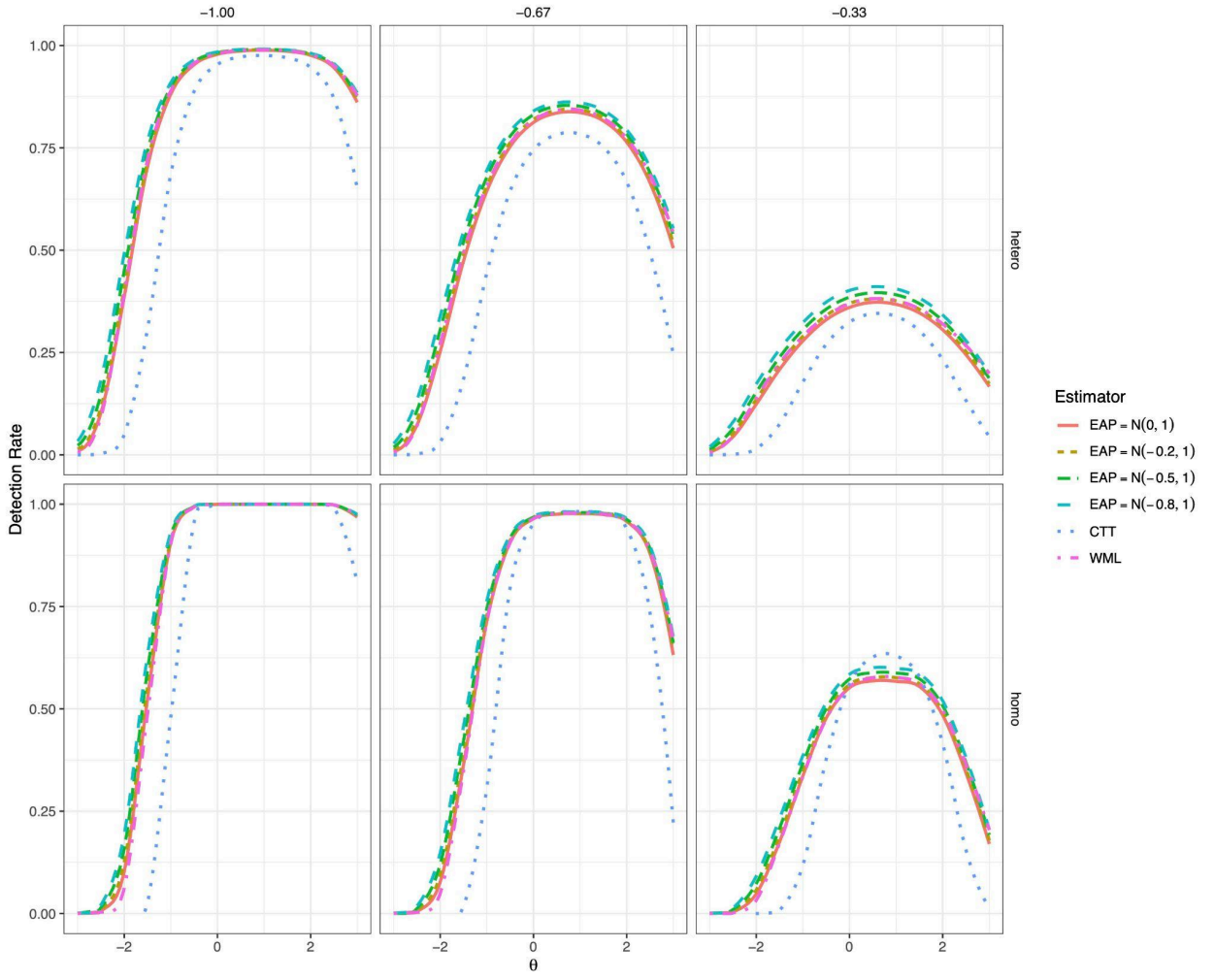


Figure 9

Power Rates for Sample Size of 100 Conditions Marginalized over Test Length

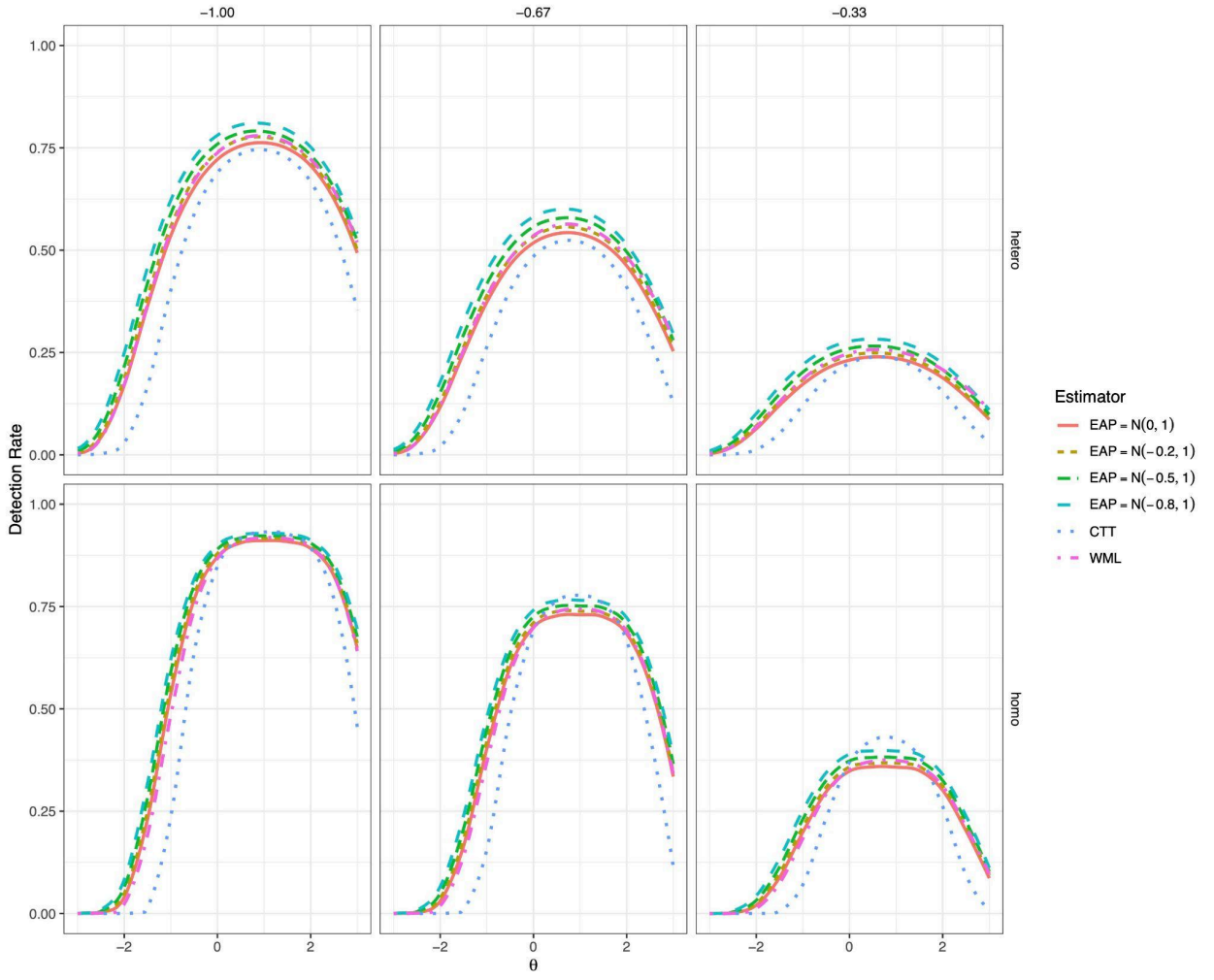


Figure 10

Power Rates for Sample Size of 1000 Conditions Marginalized over Test Length