Psychometrika (2026), **0**, 1–25
doi:10.1017/psy.2026.10146

PSYCHOMETRIC SOCIETY

APPLICATION AND CASE STUDIES - ORIGINAL

Investigating Within-Person Variation in Response Processes for Likert-Scale Data with a Response-Level Mixture IRTree Model

Jinwen Luo[1], Sijia Huang[2] and Minjeong Jeon[3]

[1]South China Normal University, China; [2]School of Education, Indiana University Bloomington, USA; [3]Education, University of California, Los Angeles, USA

Corresponding author: Sijia Huang; Email: sijhuang@iu.edu

(Received 4 November 2025 UTC; revised 18 August 2026 UTC; accepted 25 August 2026 UTC)

Abstract

Self-report scales are central to psychological and behavioral science, yet most analyses assume a single response process. In practice, respondents may follow cognitively distinct processes to answer the same item, and a respondent may switch processes across items. Ignoring between-person and/or within-person heterogeneity in item response process may bias item parameter estimates and person scores. In this study, we introduce a response-level mixture model, namely, MIX-R. The proposed model represents two or more distinct response processes using item response trees and specifies the probability that a specific process is used with a prevalence model, which includes additive person and item effects and optional covariates. We fit MIX-R in a fully Bayesian framework. Applied to three-category items from the verbal aggression questionnaire, MIX-R reveals distinct process-specific item functioning so that expected scores of the same item vary across response processes. As shown in our simulation studies, MIX-R exhibits better model fit when it is consistent with the data-generating process and does not overfit. In addition, it accurately recovers person-level and item-level process prevalence and item parameters. We conclude this study by discussing its limitations and outlining directions for future research.

Keywords: item response tree; mixture models; response processes; within-person heterogeneity

1. Introduction

Self-report scales using Likert-type items are common in the social and behavioral sciences. Examples range from the three-category items in the verbal aggression questionnaire (0 = *No*, 1 = *Perhaps*, and 2 = *Yes*; De Boeck & Wilson, 2004) to four-category items in the student survey of the Programme for International Student Assessment (PISA; 0 = *Strongly disagree*, 1 = *Disagree*, 2 = *Agree*, and 3 = *Strongly agree*). Respondents' cognitive process of arriving at a particular response category (i.e., the response process) can be complex, often involving multiple stages rather than a single, immediate reaction (Tourangeau et al., 2000). For instance, when responding to a three-category item in the verbal aggression questionnaire, respondents may first decide whether they agree with the statement (i.e., if select *No*); if so, they then proceed to determine the extent to which they agree and choose between *Perhaps* and *Yes* (Jeon & De Boeck, 2016).

When responding to a Likert-type item, respondents may use different response processes. This is similar to using different problem-solving strategies to answer the same question. For a three-category

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

verbal aggression item, in addition to the process described above, respondents may view *Perhaps* as a qualitatively distinct option and select it when they do not want to give a definitive response. If they are willing to reveal their attitude, they then choose between *No* and *Yes*.

In the current edition of *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), response processes are identified as one of the five sources of validity evidence and have thus received increasing attention from the research community. Examining response processes can also inform refinements to the content and design of test items (e.g., Embretson, 2023; Lenzner et al., 2023). For instance, items may be revised through rewording, or new items may be added to better capture the intended constructs (Bao et al., 2024).

It is important to note that, although related, response processes are conceptually distinct from response styles. A response process refers to the set of cognitive steps through which a respondent arrives at an answer. The empirical uses of these processes depend jointly on respondents and items. Response processes can be effectively represented using IRTree models, which accommodate multiple cognitive stages and latent traits (e.g., Böckenholt, 2012; De Boeck & Partchev, 2012; Jeon & De Boeck, 2016). Response styles, by contrast, are person-level tendencies to use certain response options across items, such as extreme or middle categories. They are usually modeled as latent traits separate from the target trait and not tied to a specific item structure (e.g., Huang et al., 2024; Kim & Bolt, 2021). Thus, respondents with similar response style tendencies may still follow different response processes when answering the same item.

1.1. IRTree for response process

IRTree models represent a family of item response theory (IRT) models with sequential decision tree structures. The tree structures describe the cognitive decisions respondents make when answering Likert-type items (e.g., Böckenholt, 2012; De Boeck & Partchev, 2012; Jeon & De Boeck, 2016). Specifically, an assumed response process is broken down into stages, with each stage represented by an internal decision node. These internal nodes may rely on different respondent traits such as the target latent trait, the item aims to measure and/or latent traits for response styles. The original polytomous item is therefore represented by several binary pseudo-items, and the probability of observing a specific response to the original item becomes the product of conditional probabilities of internal decisions that lead to the observed response. IRTree models have been used to study responses styles (e.g., Plieninger & Meiser, 2014) and omitted item responses (e.g., Debeer et al., 2017).

To illustrate the IRTree model, consider a three-category verbal aggression item, *A bus fails to stop for me. I would want to curse.* The aforementioned two-stage response process can be described by the tree structure shown in Figure 1(a). Let θ denote respondents' level of verbal aggression, the target trait measured by the questionnaire. In Stage 1, respondents decide if they agree with the statement, depending on their θ level. This stage includes one internal node, denoted by $Y_1^*(\theta)$, which can be viewed as a binary pseudo-item. If $Y_1^*(\theta) = 0$, the respondents reach a terminal option 0 = *No*, and if $Y_1^*(\theta) = 1$, the respondents move to Stage 2. In Stage 2, they further decide to what extent they agree with the statement. This stage involves another internal node, denoted by $Y_2^*(\theta)$, that also relies on θ. When $Y_2^*(\theta) = 0$, the observed response is 1 = *Perhaps*, and when $Y_2^*(\theta) = 1$, the observed response is 2 = *Yes*. The one-to-one correspondence between the three response options and responses to the two binary pseudo-items are summarized in the first two columns of Table 1. In this linear response process (LPT), the three response options show only quantitative differences. In other words, the middle category lies between *No* and *Yes* on the attitude continuum.

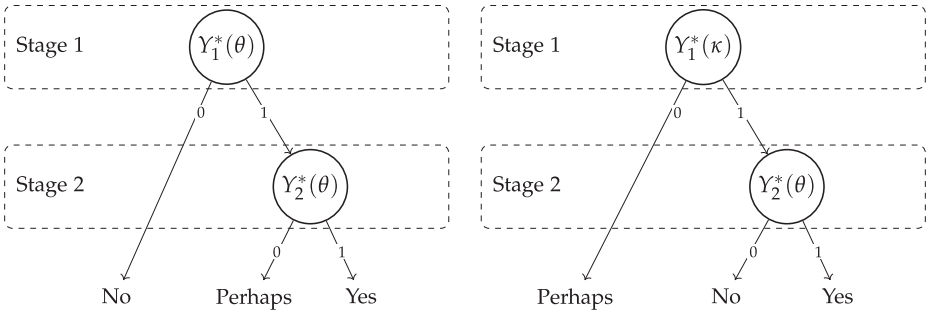
The second possible response process can be described by the tree structure in Figure 1(b). This nested process tree (NPT) also has two stages, but the two internal nodes depend on different latent traits. In Stage 1, respondents decide whether they want to use the response option *Perhaps*. This decision relies on a latent trait, denoted by κ, that is separate from the target latent trait. κ represents a tendency to select the *Perhaps*, although the psychological meaning of this tendency may vary by context. For sensitive items in the verbal aggression questionnaire, κ may reflect respondents' willingness to reveal

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Table 1. Correspondence between observed responses and internal node codes

	LPT		NPT	
	$Y_1^*(\theta)$	$Y_2^*(\theta)$	$Y_1^*(\kappa)$	$Y_2^*(\theta)$
0 = No	0	NA	1	0
1 = Perhaps	1	0	0	NA
2 = Yes	1	1	1	1

Note: NA indicates that the corresponding internal node is not applicable on that branch of the response tree.



(a) Linear process tree (LPT). (b) Nested process tree (NPT).

Figure 1. Item response trees for processes with a three-response item.

their attitudes. In other settings, it may capture a middle response style or non-effortful responding. As shown in the last two columns of Table 1, if $Y_1^*(\kappa) = 0$, the observed response is 1 = *Perhaps*. If $Y_1^*(\kappa) = 1$, the respondents move to the next stage. In Stage 2, the internal node $Y_2^*(\theta)$ represents the direction of the attitude and depends on the target trait θ. Therefore, $Y_2^*(\theta) = 0$ leads to 0 = *No*, and $Y_2^*(\theta) = 1$ leads to 2 = *Yes*.

1.2. Between-person variation in response process

Although the IRTree model provides information regarding how respondents choose from options of Likert-type items, standard IRTree models usually assume that all respondents follow the same response process. This assumption may not hold in practice because the same item may trigger more than one response process. For the three-category verbal aggression items discussed above, some respondents may follow the LPT, whereas others may follow the NPT. To account for such between-person variation, mixture IRTree models have been proposed (e.g., Alagöz & Meiser, 2024; De Boeck et al., 2011; Kim & Bolt, 2021; Tijmstra et al., 2018). These models typically assume that respondents belong to two or more mutually exclusive latent classes (LCs). For each LC, an IRT or IRTree model is specified to describe the response process used by respondents in that class across all items in the scale.

For instance, Tijmstra et al. (2018) introduced a two-class mixture IRTree model for Likert-type items with a neutral middle category (e.g., *Neutral*). The model assumes that respondents in one LC treat the middle category as the center of an ordinal response scale, while respondents in the other LC use the middle category to pass or avoid responding. For the first LC, a generalized partial credit model (gPCM; Muraki, 1992) is used to describe the relation between the target latent trait and the observed responses. For the second LC, an IRTree model similar to the NPT in Figure 1(b) is applied to model the response process.

The mixture IRTree model proposed by Kim and Bolt (2021) focuses on four-category Likert-type items (e.g., *Strongly disagree*, *Disagree*, *Agree*, and *Strongly agree*) and modeled two response processes

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

using different IRTree models. Accordingly, two LCs of respondents were considered: one class relies only on the target latent trait and does not show an extreme response style (ERS) when responding to items; the other class incorporates an additional ERS latent trait that encourages "extreme" responses (e.g., *Strongly disagree* and *Strongly agree*). For the non-ERS class, a two-stage IRTree model is used, in which both internal nodes depend on the target trait. The IRTree for the ERS class also consists of two stages. In Stage 1, respondents determine the direction of their attitude based on their target trait. In Stage 2, they further determine the extremity of their response (e.g., selecting *Strongly agree* over *Agree*) based on their ERS trait.

1.3. Within-person variation in response process

The above mixture IRT/IRTree models capture between-person variation but assume that a respondent follows the same response process for all items. However, this assumption can be overly restrictive. As discussed in Tourangeau et al. (2000), some self-reported surveys may include items that are perceived as intrusive or sensitive by some respondents. One important source of item sensitivity is the social desirability of response options. When answering sensitive items, respondents may use response options differently from how they use them for nonsensitive items. Other factors, such as limited understanding of the item or insufficient background information, may also affect how respondents use response options (Tijmstra & Bolsinova, 2025).

To account for within-person variation in response processes, Tijmstra and Bolsinova (2025) introduced a mixture IRTree model that focuses on the use of the middle response category (e.g., *Neutral*). Using a multinomial processing tree framework (Klauer, 2010), their model assumes that, across different items, the same respondent may reach the *Neutral* response category through two cognitive pathways: either by using it as an informative option or as a nonresponse option. When the middle category is used as an informative option for target trait inference, its probability can be modeled using standard polytomous IRT models, such as Samejima's (1969) graded response model (GRM). When the response is intended as a nonresponse option, the probability of choosing it is fixed at 1. Consequently, the probability of a middle response is a weighted sum of the middle response probabilities from these two cognitive pathways, with weights representing the proportions of informative and noninformative responses.

Other models also allow response processes to vary within respondents across items. Huang (2020) proposed a mixture IRTree model for performance decline, which includes a normal class, where respondents exert full effort across the test, and an aberrant class, where respondents may drop out, skip items, or reduce effort through a sequential decision process. Merhof and Meiser (2023) developed a dynamic response strategy model to model changes in response strategies over item positions. Their model reformulates IRTree models as multidimensional IRT models and allows the loadings on latent traits to vary as a function of item position.

1.4. This study

In this study, we introduce a new response-level mixture IRTree modeling framework, namely, MIX-R, to account for both within- and between-person variation in the response processes. MIX-R is motivated by the response-mixture IRT model introduced by Huang et al. (2025), which models person- and item-specific LCs for item response strategies and uses auxiliary person-level, item-level, and/or response-level information (e.g., response time), when available, to inform LC memberships. Broadly speaking, MIX-R can be viewed as a response-level generalization of the generalized mixture IRT framework of Tijmstra et al. (2018). Rather than assigning each respondent to a single response structure across all items, MIX-R treats process membership as response-specific and models it jointly through person and item effects. This allows the same respondent to use different complete response trees across items and permits process prevalence to be examined at the response, person, and item levels.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

MIX-R differs from existing within-person mixture IRTree models in how it represents response process switching. Tijmstra and Bolsinova (2025) specify a single IRTree with two latent routes to the middle category, allowing the same middle response to reflect either an informative response or a nonresponse. In contrast, MIX-R uses a response-specific latent process indicator to select one of two or more complete IRTrees for each response. The indicator is mutually exclusive, meaning that each response is assumed to arise from exactly one process. Thus, MIX-R places the mixture at the level of item responses, so that the same respondent may use different response processes for different items. MIX-R also does not require auxiliary information, unlike models that use response-level information to capture performance decline or dynamic response strategies (e.g., Huang, 2020; Merhof & Meiser, 2023). It can incorporate person-, item-, and response-level information to predict response-level LC memberships and can be generalized to items with different numbers of response categories and to settings with more than two possible response processes.

The remainder of this article is organized as follows. In Section 2, we illustrate the MIX-R model in the context of three-category items with two possible response processes. We then outline its generalization to items with more than three response options. In Section 3, we demonstrate the empirical utility of the MIX-R model using the verbal aggression data (De Boeck & Wilson, 2004). In Section 4, through three simulation studies, we systematically evaluated the performance of the proposed MIX-R model. Finally, in Section 5, we discuss limitations of this study and future research directions.

2. Proposed approach

Without loss of generality, we present the MIX-R using the verbal aggression items with three response options (i.e., 0 = *No*, 1 = *Perhaps*, and 2 = *Yes*). We assume that respondent j $(j = 1, \ldots, J)$ follows one of the two response processes to answer item i $(i = 1, \ldots, I)$. Let C_{ij} denote the response-specific LC for response Y_{ij}. C_{ij} = LPT if the linear process tree in Figure 1(a) is used, and C_{ij} = NPT if the NPT in Figure 1(b) is used.

LPT model: Recall that, when C_{ij} = LPT, the respondent perceives only quantitative differences between the three response options and proceeds stepwise from the lowest to the highest category via the two internal nodes. The two pseudo-items in the LPT process, denoted as $A1$ and $A2$, rely on only the target latent trait θ_j. For each pseudo-item, we use a Rasch model to specify the probability of endorsement:

$$p_{ij}^{A1} = \text{logit}^{-1}(\theta_j - b_i^{A1}), \qquad \text{and} \qquad p_{ij}^{A2} = \text{logit}^{-1}(\theta_j - b_i^{A2}),$$

where b_i^{A1} and b_i^{A2} are the difficulty parameters for the Stage-1 and Stage-2 pseudo-items, respectively. Because the LPT represents a linear ordinal process, we impose $b_i^{A2} > b_i^{A1}$, so that endorsing *Yes* requires a higher θ level than moving beyond *No*. The target latent trait θ is assumed to follow a normal distribution. Then we have the conditional probabilities for the observed response Y_{ij}:

$$\begin{aligned} P(Y_{ij} = \text{No} \mid C_{ij} = \text{LPT}) &= 1 - p_{ij}^{A1}, \\ P(Y_{ij} = \text{Perhaps} \mid C_{ij} = \text{LPT}) &= p_{ij}^{A1}(1 - p_{ij}^{A2}), \\ P(Y_{ij} = \text{Yes} \mid C_{ij} = \text{LPT}) &= p_{ij}^{A1} p_{ij}^{A2}. \end{aligned} \tag{1}$$

NPT model: When C_{ij} = NPT, the first internal node, $B1$, depends on κ, a trait capturing the extent to which a respondent prefers using the *Yes* and *No*. The second node, $B2$, relies on θ. The endorsement probabilities of two NPT pseudo-items are

$$p_{ij}^{B1} = \text{logit}^{-1}(\kappa_j - b_i^{B1}), \qquad p_{ij}^{B2} = \text{logit}^{-1}(\theta_j - b_i^{B2}),$$

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

where b_i^{B1} and b_i^{B2} are the difficulty parameters for κ-based and θ-based nodes, respectively. κ is assumed to follow a normal distribution. The conditional probabilities are

$$\begin{aligned} P(Y_{ij} = \text{Perhaps} \mid C_{ij} = \text{NPT}) &= 1 - p_{ij}^{B1}, \\ P(Y_{ij} = \text{No} \mid C_{ij} = \text{NPT}) &= p_{ij}^{B1}(1 - p_{ij}^{B2}), \\ P(Y_{ij} = \text{Yes} \mid C_{ij} = \text{NPT}) &= p_{ij}^{B1}\, p_{ij}^{B2}. \end{aligned} \tag{2}$$

In the proposed model, the target trait θ_j is shared across the direction-relevant nodes of both trees and is identified on a common scale. Respondents may therefore be compared with respect to their θ_j. In contrast, κ_j characterizes the decision specific to NPT and should not be interpreted as another measure of the target trait or compared numerically with θ_j. The validity of cross-respondent comparisons of θ_j remains conditional on the assumption that the trait has the same substantive interpretation across the alternative trees.

Prevalence model: We model, π_{ij}, the probability that respondent j follows LPT to answer item i, using a logistic prevalence model,

$$\pi_{ij} = \text{logit}^{-1}\left(\alpha_0 + \mathbf{z}_j^\top \boldsymbol{\gamma} + \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{h}_{ij}^\top \boldsymbol{\delta} + u_j + v_i\right), \tag{3}$$

where $\mathbf{z}_j$, $\mathbf{x}_i$, and $\mathbf{h}_{ij}$ are vectors of person-, item-, and response-level covariates, respectively. The coefficient vectors $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$ describe how these covariates are associated with the log-odds of following LPT rather than NPT. Unless stronger design assumptions are justified, the covariate effects should be interpreted as associational. The terms u_j and v_i are person and item random effects. Larger values of u_j indicate that respondent j is more likely to follow LPT across items, whereas larger values of v_i indicate that item i is more likely to elicit LPT across respondents. Instead of estimating the effects of covariates simultaneously with parameters of items, a stepwise estimation approach that separates the measurement and structural components of MIX-R can be employed (Alagöz et al., 2025; Vermunt, 2010). Under this approach, the measurement model is first estimated using only item responses, without incorporating any covariates, to obtain posterior class membership probabilities of respondents. Subsequently, the effects of covariates are estimated by regressing the LC membership probabilities on covariates of interest.

If no covariate is included or when all covariates are zero, $\text{logit}^{-1}(\alpha_0)$ can be interpreted as the baseline LPT probability when both person and item random effects are zero, $u_j = 0$ and $v_i = 0$ (i.e., the probability that a person with an average tendency to use LPT will follow LPT when responding to an item with an average tendency to elicit it). More generally, the response-level prevalence of LPT is the model-implied average $\bar{\pi}$; and $\bar{\pi}_j$ summarizes respondent j's tendency to follow LPT across items, and $\bar{\pi}_i$ summarizes how likely item i is to elicit LPT across respondents:

$$\bar{\pi} = \frac{1}{JI}\sum_{j=1}^{J}\sum_{i=1}^{I}\pi_{ij}, \qquad \bar{\pi}_j = \frac{1}{I}\sum_{i=1}^{I}\pi_{ij}, \qquad \text{and} \quad \bar{\pi}_i = \frac{1}{J}\sum_{j=1}^{J}\pi_{ij}. \tag{4}$$

Combining Equations (1)–(3), the probability of observing a specific response option, such as *No*, is

$$P(Y_{ij} = \text{No}) = \pi_{ij} P(Y_{ij} = \text{No} \mid C_{ij} = \text{LPT}) + (1 - \pi_{ij}) P(Y_{ij} = \text{No} \mid C_{ij} = \text{NPT}). \tag{5}$$

2.1. Model estimation and identification

Estimation: The MIX-R model was estimated in a Bayesian framework using the R package NIMBLE (de Valpine et al., 2017) with the marginalized mixture likelihood $P(Y_{ij}) = \pi_{ij}P(Y_{ij} \mid C_{ij} = \text{LPT}) + (1 - \pi_{ij})P(Y_{ij} \mid C_{ij} = \text{NPT})$, integrating out the discrete LC indicator. All unknowns are continuous and sampled using NIMBLE's adaptive random-walk Metropolis samplers. We run multiple chains, monitor rank-normalized split $\hat{R}$ (Vehtari et al., 2021) and bulk/tail ESS, and assess out-of-sample fit via LOOIC (Vehtari et al., 2017). Note that a data-augmentation version that samples discrete LCs C_{ij}

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Table 2. Priors used in the main MIX-R specification

Parameter	Prior	Notes
θ_j	$\mathcal{N}(0,1)$	Target trait
κ_j	$\mathcal{N}(0,1)$	Secondary trait
$b^{A1}_{i,\,\text{raw}}$	$\mathcal{N}(0,5^2)$	LPT Stage 1 threshold
$b^{B1}_{i,\,\text{raw}}$	$\mathcal{N}(0,5^2)$	NPT Stage 1 threshold
$g^A_i = b^{A2}_i - b^{A1}_i$	$\text{Exp}(1)$	Enforces $b^{A2}_i > b^{A1}_i$
τ_i	$\text{Beta}(3,3)$	Defines $b^{B2}_i = b^{A1}_i + \tau_i g^A_i$
α_0	$\mathcal{N}(0,5^2)$	Prevalence intercept
u_j	$\mathcal{N}(0,1)$	Person effect in prevalence model
$v_{i,\,\text{raw}}$	$\mathcal{N}(0,\sigma_v^2)$	Item effect in prevalence model
σ_v	half-t_3(scale = 1)	Item-effect SD
$\gamma_p, \beta_q, \delta_r$	$\mathcal{N}(0,2.5^2)$	Covariate effects, when included

using Pólya–Gamma updates (Polson et al., 2013) for the prevalence model is also possible with enough information (e.g., repeated measures in a longitudinal design and the availability of auxiliary variables). For simplicity, the results reported in this article use the marginalized form.

Priors: All internal nodes use Rasch models. Table 2 summarizes the priors used for the MIX-R specification. The latent traits are assigned independent standard normal priors, $\theta_j \sim \mathcal{N}(0,1)$ and $\kappa_j \sim \mathcal{N}(0,1)$. This fixes the scale of each trait, provides mild regularization, and reflects our choice not to model the latent-trait association as a primary estimand in the main analysis. The raw Stage 1 thresholds and the item effects in the prevalence model are centered across items, with $v_i = v_{i,\,\text{raw}} - \bar{v}_{\text{raw}}$. The fixed prior $u_j \sim \mathcal{N}(0,1)$ regularizes person-level heterogeneity on the logit scale and improves computational stability in the present application, where each respondent contributes a modest number of item responses.

When the association between θ_j and κ_j is of substantive interest, the independent priors can be replaced by a bivariate normal prior, $(\theta_j, \kappa_j)^\top \sim \mathcal{N}_2\{(0,0)^\top, \Sigma_\rho\}$, where $\Sigma_\rho = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The correlation can be parameterized as $z_\rho \sim \mathcal{N}(0,\sigma_\rho^2)$ and $\rho = \tanh(z_\rho)$. The scale σ_ρ should reflect substantive knowledge about the plausible association. Because the independent- and correlated-trait versions make different assumptions about the joint distribution of the latent traits, they should be treated as distinct model specifications. For completeness, we examine this prior choice in Appendix A of the Supplementary Material.

Identification: Each binary Rasch node contributes one free intercept. With slopes fixed to 1, the node likelihood is invariant to a common location shift. We identify location and set the latent scale by centering item difficulties within each tree; standard normal priors on the person traits provide the scale reference. In the correlated-trait specification, the marginal variances of θ_j and κ_j remain fixed to one, while ρ controls only their latent association. For three-category items, we enforce the LPT ordering $b^{A2}_i > b^{A1}_i$ by setting $g^A_i = b^{A2}_i - b^{A1}_i > 0$, with $g^A_i \sim \text{Exp}(1)$. This prior mildly regularizes threshold separation on the logit scale; for example, $P(g^A_i < 3) = 0.95$. The ordering is specific to the LPT as a linear ordinal tree and does not limit the generalizability of MIX-R, because other response processes can be represented by IRTrees with their own process-appropriate constraints. NPT and LPT parameterizations are aligned by constraining $b^{B2}_i = b^{A1}_i + \tau_i(b^{A2}_i - b^{A1}_i)$ with $\tau_i \sim \text{Beta}(3,3)$. This places the NPT θ-threshold between the two LPT θ-thresholds for the same item. As a result, the *No* versus *Yes* distinction under NPT remains comparable to the ordered meaning of *No*, *Perhaps*, and *Yes* under LPT. The $\text{Beta}(3,3)$ prior is symmetric around 0.5 and mildly favors interior placements while allowing

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

the full interval $(0, 1)$. The prevalence model is identified by taking NPT as the reference process and defining π_{ij} as the probability of following LPT, centering the item-specific effects v_i, and assigning u_j a zero-centered regularizing prior. If a 2PL is used, one slope per latent dimension is fixed (e.g., $\alpha = 1$ on a reference item-node) and the remaining slopes receive diffuse log-normal priors.

2.2. Generalization to R response options and K response processes

MIX-R can be generalized to items with R response options and K possible response processes through modeling K LCs. Specifically, for the kth $(1 \le k \le K)$ LC, we adopt an IRTree with $\mathcal{M}^{(k)}$ binary internal nodes to describe the corresponding response process. Different combinations of the internal node outcomes then lead to R observed response options. Similar to Table 1, the one-to-one mapping between the R options and $\mathcal{M}^{(k)}$ internal nodes can be summarized by an $R \times \mathcal{M}^{(k)}$ matrix. Let $p_{ijm}^{(k)}$ denote the probability that respondent j endorses the mth $(1 \le m \le \mathcal{M}^{(k)})$ binary internal node in the kth IRTree of item i. The marginal probability of observing response option r $(1 \le r \le R)$ can be written as

$$P(Y_{ij} = r) = \sum_{k=1}^{K} \pi_{ijk} \prod_{m \in \mathcal{M}_r^{(k)}} \left[p_{ijm}^{(k)}\right]^{I_{r,m}^{(k)}} \left[1 - p_{ijm}^{(k)}\right]^{1 - I_{r,m}^{(k)}}, \qquad \sum_{k=1}^{K} \pi_{ijk} = 1, \tag{6}$$

where π_{ijk} is the probability that respondent j follows the kth IRTree to answer item i, and $\sum_{k=1}^{K} \pi_{ijk} = 1$. The indicator $I_{r,m}^{(k)}$ represents the element in the rth row and mth column of the mapping matrix, with $I_{r,m}^{(k)} \in \{0, 1\}$.

To model process prevalence, let the Kth LC be the reference class. The log-odds of respondent j following the kth IRTree for item i are modeled as

$$\log\left(\frac{\pi_{ijk}}{\pi_{ijK}}\right) = \eta_{ijk} = \alpha_k + u_{jk} + v_{ik}, \quad \text{for } k = 1, \ldots, K-1, \tag{7}$$

where u_{jk} and v_{ik} are LC-specific person and item random intercepts, and $u_{jk} \sim \mathcal{N}(0, \sigma_{u,k}^2)$ and $v_{ik} \sim \mathcal{N}(0, \sigma_{v,k}^2)$. For the reference LC, we let $\alpha_K = u_{jK} = v_{iK} = 0$. When observed information is available, the linear predictor in Equation (7) can be expanded to include person-, item-, and response-level predictors, $\mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}$, $\mathbf{z}_j^\top \boldsymbol{\gamma}^{(k)}$, and $\mathbf{h}_{ij}^\top \boldsymbol{\delta}^{(k)}$. These terms may be useful when the goal is to improve prediction or describe associations between observed features and response-process prevalence.

These generalizations do not require an odd number of response categories or a unique middle option. Appendix G of the Supplementary Material illustrates, respectively, a four-category specification contrasting linear and nested response processes and a five-category specification that jointly represents response-process and response-style heterogeneity.

3. Empirical demonstration

To demonstrate the MIX-R model, we use the verbal aggression self-report data (De Boeck & Wilson, 2004). The data contain responses to 24 items from $N = 316$ respondents (243 women and 73 men) with their Trait Anger scores from the State-Trait Anger Expression Inventory (STAXI; Spielberger, 1988; Spielberger & Sydeman, 1994). The items were formed by crossing four frustrating situations (two *Self-to-blame* and two *Other-to-blame*), three verbally aggressive behaviors (*Curse*, *Scold*, and *Shout*), and two behavior modes (*Want* and *Do*). Each item has three response options: 0 = *No*, 1 = *Perhaps*, and 2 = *Yes*. For details of the data, we refer readers to De Boeck and Wilson (2004). In the original administration, these items were mixed with additional items of a similar type, and the order of items was randomized. Thus, any detected process differences are likely to be a reflection of substantive item and person differences rather than item locations.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Table 3. Model fit indices of five competing models

	LOOIC	pLOO	Pareto $k > 0.7$	$\Delta LOOIC$	SE $\Delta LOOIC$
MIX-R	**11,879.22**	**512.34**	**0**	**0.0**	**/**
MIX-R+X	**11,883.80**	**512.52**	**0**	**4.6**	**8.1**
LPT	12,374.64	311.41	0	495.4	36.6
NPT	12,167.93	442.85	0	288.8	34.2
MIX-P	11,923.62	531.74	79	44.4	32.8

3.1. Analysis

In the analysis, we consider five competing model specifications: an IRTree model for the linear response process (hereafter, the LPT model), an IRTree model for the nested response process (hereafter, the NPT model), a person-level mixture IRTree model (hereafter, the MIX-P model), a response-level mixture IRTree model (MIX-R model), and a MIX-R model with observed covariates (MIX-R+X). In all mixture models, one LC follows the LPT in Figure 1(a), and the other follows the NPT in Figure 1(b). In MIX-P, each respondent follows one process across all items. In MIX-R and MIX-R+X, the process may vary across person–item responses.

In this application, θ represents respondents' propensity to react with verbal aggression when provoked, and κ represents their tendency to reveal a directed attitude rather than select *Perhaps*. MIX-R includes only person and item random effects in the prevalence model. MIX-R+X adds observed predictors of response-process prevalence. Person-level covariates include gender and Trait Anger scores; item-level covariates include behavior mode, situation type, and behavior types.

We compare the five specifications using out-of-sample predictive fit and Pareto-k reliability. We then examine item properties and process-specific item functioning under the selected model. Finally, we summarize the overall response-level, item-level, and person-level prevalence of LPT, as defined in Equation (4). For MIX-R+X, covariate effects are interpreted as associations with the log-odds of following LPT rather than NPT.

3.2. Estimation, priors, and convergence diagnostics

All five models were estimated in a Bayesian framework via MCMC using NIMBLE (de Valpine et al., 2017). To support model comparison, we used the same prior families and scales for parameters with comparable roles across models as stated in Section 2.1. We used four parallel chains, each with 24, 000 iterations; of these, 4, 000 are warm-up, and thinning =10. We assessed MCMC behavior using visual diagnostics (e.g., trace and autocorrelation plots) and the rank-normalized split $\hat{R}$ (Vehtari et al., 2021). For the primary quantities of interest (e.g., the overall prevalence parameter α_0), we required that $\hat{R} \leq 1.01$ and bulk/tail ESS ≥ 400. For person and item parameters, we targeted $\hat{R} \leq 1.05$ and ESS ≥ 100.

3.3. Results

3.3.1. Model comparison

All models show satisfactory convergence based on $\hat{R}$. For MIX-R specifications, no monitored parameter had $\hat{R} > 1.05$. Because conventional criteria such as BIC and DIC can be unreliable for mixture models (Drton & Plummer, 2017), we assessed predictive fit using PSIS-LOO, reported as LOOIC, with smaller values indicating better fit. We also report the Pareto-k diagnostic, where values above 0.7 indicate unreliable LOO approximations (Vehtari et al., 2017). Table 3 summarizes the model comparison results.

As shown in Table 3, both response-level mixture models fit the data better than the two single-process models. This suggests that the verbal aggression responses are unlikely to arise from a

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

single response process. MIX-R also outperformed MIX-P by LOOIC and showed better diagnostic reliability. MIX-P had 79 left-out units with Pareto-$k > 0.7$, whereas both MIX-R specifications had none. Thus, the person-level mixture captured some process heterogeneity, but its predictive performance relied on a less stable approximation. Substantively, MIX-R is also better aligned with the response-process question in this study. MIX-P assumes that each respondent follows one response process across all items, whereas MIX-R allows the same respondent to switch processes across items. This distinction is important for the verbal aggression data because item content may affect how respondents use *Perhaps*; for example, *Want* items and *Do* items may invite different uses of the middle category. Therefore, the better LOOIC, stronger Pareto-k reliability, and closer match to the within-person process-switching hypothesis together support MIX-R as the primary model.

Adding covariates to MIX-R did not improve predictive fit. As shown in the table, MIX-R+X had a slightly larger LOOIC than MIX-R, and the difference was small relative to its uncertainty ($\Delta LOOIC = 4.6$, SE = 8.1). We therefore use MIX-R as the primary model for interpreting process-specific item functioning and response-process prevalence. MIX-R+X is used only to describe whether process prevalence is associated with respondent characteristics and item features.

3.3.2. Item properties under LPT and NPT

Item parameter estimates: The posterior means and 95% credible intervals (CrIs) of item difficulty parameters ($\hat{b}^{A1}, \hat{b}^{A2}, \hat{b}^{B1}, \hat{b}^{B2}$), along with the item random effects $\hat{v}_i$ are summarized in Table S2 in the Supplementary Material. The estimates are broadly well identified, with several MIX-R item parameters, especially for the NPT first node, showed wider CrIs than those from the single-process models, reflecting additional uncertainty from estimating response-level process membership. Table S3 in the Supplementary Material provides a descriptive comparison of item-parameter posterior means across LPT, NPT, MIX-P, and MIX-R. Figure 2 illustrates how the two response processes differ at the item level through category response curves for three representative items: `S1DoCurse`, `S3DoScold`, and `S3WantCurse`.

Figure 2 illustrates how LPT and NPT can produce different category probabilities for the same item. Under LPT, both internal decisions depend on θ. As θ increases, *No* generally decreases, *Yes* increases, and *Perhaps* is highest in the middle range of θ. This pattern appears in the first row of Figure 2. Under NPT, κ determines whether respondents select *Perhaps* or proceed to the θ-based *No*/*Yes* decision. When κ is low, *Perhaps* remains likely across θ, so changes in θ have limited effect on the observed response. As κ increases, respondents are more likely to reach the second node, and the *No* and *Yes* curves become more responsive to θ.

The three items show different forms of process-specific item functioning. For the item `S1DoCurse`, LPT and NPT differ mainly in *Perhaps*. Under LPT, *Perhaps* is likely only within a limited middle range of θ, and response probabilities are mostly divided between *No* and *Yes*. Under NPT, *Perhaps* is mainly determined by κ. When κ is high, *Perhaps* has low probability, and the *No* and *Yes* curves resemble those under LPT for the same θ range. For `S3DoScold`, the main difference is that low to moderate θ mostly leads to *No* under LPT, but to *Perhaps* under NPT when κ is low. Thus, across much of the θ range, predicted category probabilities under NPT depend strongly on κ. For `S3WantCurse`, LPT assigns more probability to the middle category, whereas NPT can assign more probability to *No* at lower θ. This occurs because the first NPT node is easier to pass for this item, allowing more respondents to reach the θ-based *No*/*Yes* decision. These curve-level differences help explain the expected-score differences reported in Table 4.

Expected item score differences: To describe the observed-score implications of the two response processes, we compare the expected item scores under LPT and NPT, denoted by $\mathrm{E}[Y_i^{\mathrm{LPT}}]$ and $\mathrm{E}[Y_i^{\mathrm{NPT}}]$. These expected scores are computed using the conventional item scoring 0 = *No*, 1 = *Perhaps*, and 2 = *Yes*. For each MCMC draw s, we first compute $p_{ijk}^{\mathrm{LPT}(s)}$, the probability that respondent j chooses

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

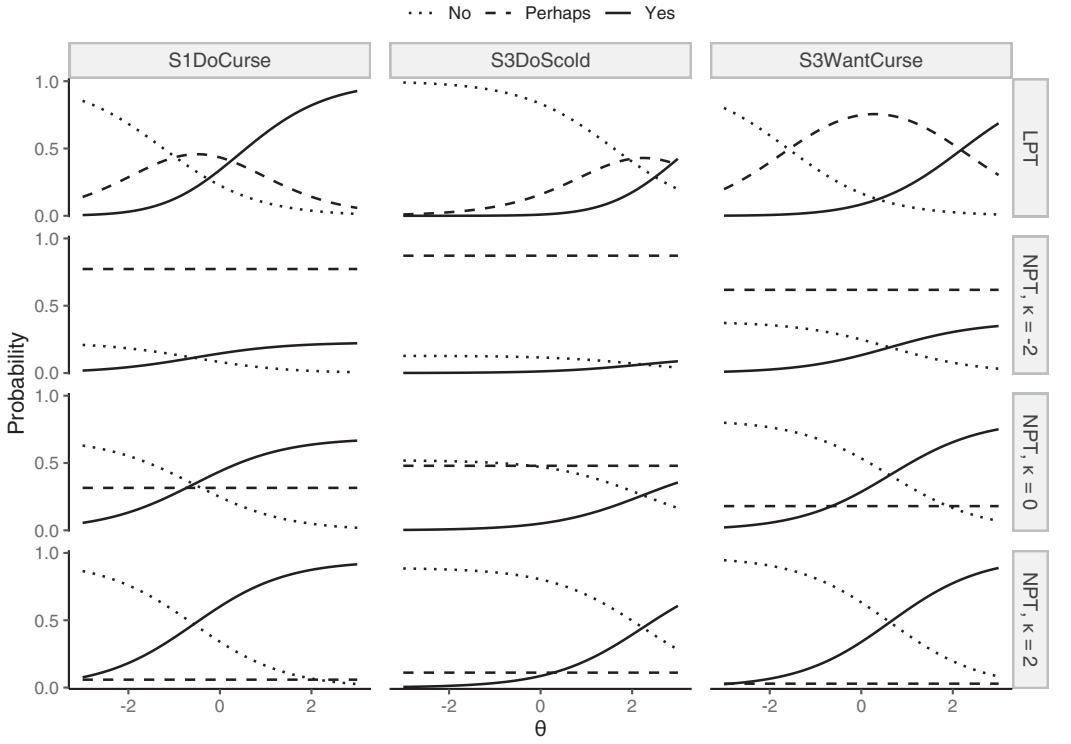


Figure 2. Category response curves under LPT and NPT for selected items.
Note: Columns show selected items, and rows show LPT and NPT at $\kappa = -2, 0, 2$. Dotted, dashed, and solid lines denote *No*, *Perhaps*, and *Yes*, respectively. For `S1DoCurse`, LPT and high-κ NPT give similar *No*/*Yes* curves because *Perhaps* has low probability. For `S3DoScold`, low to moderate θ mainly leads to *No* under LPT but to *Perhaps* under NPT when κ is low. For `S3WantCurse`, LPT assigns more probability to *Perhaps*, whereas NPT can assign more probability to *No* at lower θ.

category $k = (0, 1, 2)$ of item i under LPT. We then obtain the draw-specific average expected scores across respondents:

$$\mathrm{E}^{(s)}\left[Y_i^{\mathrm{LPT}}\right] = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=0}^{2} k p_{ijk}^{\mathrm{LPT}(s)}. \tag{8}$$

We summarize $\mathrm{E}^{(s)}\left[Y_i^{\mathrm{LPT}}\right]$ across draws to obtain the mean of $\mathrm{E}\left[Y_i^{\mathrm{LPT}}\right]$. The expected item score under NPT, $\mathrm{E}\left[Y_i^{\mathrm{NPT}}\right]$, is calculated similarly. The difference $\Delta\mathrm{E}\left[Y_i\right] = \mathrm{E}\left[Y_i^{\mathrm{NPT}}\right] - \mathrm{E}\left[Y_i^{\mathrm{LPT}}\right]$ is computed within each draw and then summarized across draws.

Table 4 summarizes expected observed scores under LPT and NPT. This comparison shows what can happen when responses from different processes are scored using the same conventional $0, 1, 2$ rule. Under this rule, the same item can contribute different expected observed scores depending on the response process. The differences are not uniform across items. For most items, the expected score under NPT is higher than under LPT. Several *Want* items show the opposite pattern, including `S1WantCurse`, `S1WantScold`, `S2WantCurse`, `S2WantScold`, `S3WantCurse`, and `S4WantCurse`. Thus, ignoring response-process heterogeneity can change how item responses contribute to a sum score, and the direction depends on the item.

Item-level prevalence and predictors: Recall that the item random effects v_i index the prevalence of the LPT process for item i, with a larger $\hat{v}_i$ value that item i is more likely to elicit LPT, after accounting for the model terms included in the prevalence model. As shown in Figure 3, `S3DoShout`, `S3WantShout`, and `S3DoScold` showed the largest positive item effects, while `S2WantScold`, `S1WantScold`, and

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Table 4. Expected observed scores under LPT and NPT by item mode

Do items				*Want* items			
Item	$E[Y^{LPT}]$	$E[Y^{NPT}]$	$\Delta E[Y]$	Item	$E[Y^{LPT}]$	$E[Y^{NPT}]$	$\Delta E[Y]$
S1DoCurse	1.06	1.11	0.05	S1WantCurse	1.32	1.08	−0.24
S1DoScold	0.74	0.89	0.15	S1WantScold	1.06	0.91	−0.15
S1DoShout	0.36	0.57	0.21	S1WantShout	0.60	0.77	0.18
S2DoCurse	0.97	1.05	0.08	S2WantCurse	1.40	1.16	−0.24
S2DoScold	0.61	0.82	0.22	S2WantScold	1.05	0.96	−0.09
S2DoShout	0.25	0.49	0.23	S2WantShout	0.59	0.79	0.20
S3DoCurse	0.51	0.66	0.15	S3WantCurse	0.89	0.77	−0.12
S3DoScold	0.24	0.61	0.37	S3WantScold	0.42	0.52	0.10
S3DoShout	0.07	1.00	0.93	S3WantShout	0.21	1.00	0.79
S4DoCurse	0.78	0.96	0.17	S4WantCurse	1.11	0.97	−0.14
S4DoScold	0.45	0.70	0.25	S4WantScold	0.47	0.68	0.20
S4DoShout	0.13	0.33	0.20	S4WantShout	0.29	0.53	0.24

Note: Expected scores are computed using the conventional coding 0 = *No*, 1 = *Perhaps*, and 2 = *Yes*. $\Delta E[Y] = E[Y^{NPT}] - E[Y^{LPT}]$. The comparison applies the same observed-score rule to both processes. It should be read as the expected observed-score consequence of using conventional sum scoring when responses may arise from different response processes.

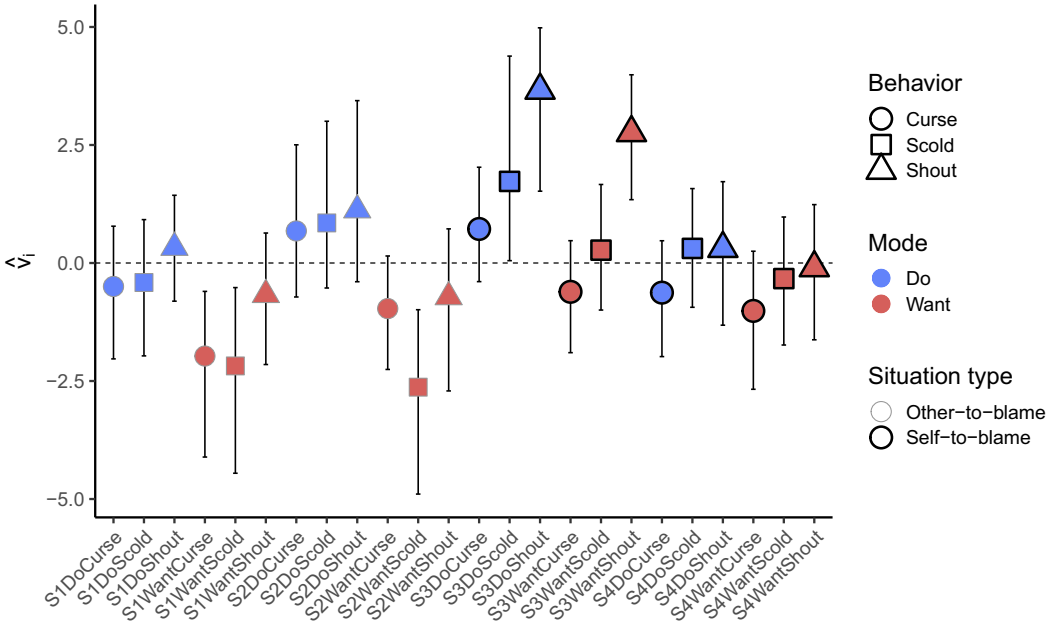


Figure 3. Item random effects $\hat{v}_i$ with 95% credible intervals.

`S1WantCurse` showed the most negative item effects. These estimates indicate that some items were more likely than others to elicit LPT.

To examine whether item differences relate to item features, Table 5 reports summaries from MIX-R and MIX-R+X. In the MIX-R column, item-level entries are post hoc regressions using posterior

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Table 5. Model summaries and associations with response-process prevalence under MIX-R and MIX-R+X

Quantity	MIX-R	MIX-R+X
Latent associations		
$\mathrm{cov}(\hat{\theta}, \hat{\kappa})$	−0.100 [−0.208, 0.010]	−0.100 [−0.208, 0.010]
$\mathrm{cov}(\hat{\theta}, \hat{u})$	−0.199 [−0.303, −0.091]	−0.185 [−0.290, −0.077]
Person-level predictors		
Female vs. male	−0.078 ($p = 0.25$)	0.510 [−0.173, 1.205]
Trait Anger score	−0.013 ($p < 0.05$)	−0.071 [−0.131, −0.011]
Item-level predictors		
Want vs. Do	−1.361 ($p < 0.01$)	−1.528 [−2.809, −0.386]
Self-to-blame vs. other-to-blame	**1.179** ($p < 0.01$)	**1.602 [0.414, 2.780]**
Blaming behavior	−1.371 ($p < 0.05$)	−2.010 [−3.402, −0.612]
Expressive behavior	−0.237 ($p = 0.64$)	−0.450 [−1.771, 0.855]

Note: Values in brackets are 95% credible intervals. Bold values indicate that the interval excludes zero or, for the MIX-R posterior-estimate associations, that the corresponding test is statistically significant. Latent associations were computed as correlations among posterior-mean latent score estimates obtained under the independent-prior specification; a sensitivity analysis using a bivariate prior yielded the same substantive conclusions, with the covariate-adjusted θ-κ association remaining weak and its 95% credible interval including zero (see Table S1 in the Supplementary Material). In the MIX-R column, person-level predictor entries are regression coefficients from models using posterior mean $\hat{u}_j$ as the outcome, and item-level predictor entries are regression coefficients from models using posterior mean $\hat{\nu}_i$ as the outcome. In the MIX-R+X column, predictor entries are coefficients estimated jointly in the prevalence model and are on the log-odds scale for following LPT rather than NPT. *Scold* and *Curse* are coded as blaming behavior, and *Curse* and *Shout* are coded as expressive behavior.

mean $\hat{\nu}_i$ as the outcome. In the MIX-R+X column, the same item features are included directly in the prevalence model. Both approaches give consistent results: *Want* items and blaming behaviors are associated with lower LPT prevalence, whereas self-to-blame items are associated with higher LPT prevalence. Expressive behavior shows no clear association. Together, these results suggest that item features help explain why some items are more likely than others to elicit LPT.

3.3.3. Person-level prevalence and latent scores

Table 5 also summarizes person-side results. Under both MIX-R and MIX-R+X, the posterior covariance between $\hat{\theta}_j$ and $\hat{\kappa}_j$ was weakly negative, −0.100, with a 95% CrI of $[-0.208, 0.010]$ including 0. This result should be interpreted cautiously because the main specification assigns independent priors to the two latent traits and does not treat their association as a primary estimand. As a sensitivity check, analyses using a bivariate normal prior for θ_j, κ_j produced the same general pattern of a weak negative association; these results are reported in Appendix A of the Supplementary Material.

The person-level covariate results describe whether respondent characteristics are associated with response-process prevalence. In MIX-R, post hoc regressions using posterior mean $\hat{u}_j$ as the outcome showed no clear gender association, but Trait Anger was negatively associated with $\hat{u}_j$. MIX-R+X showed the same pattern when these covariates were included directly in the prevalence model. Conditional on the other terms in the prevalence model, respondents with higher Trait Anger scores were somewhat less likely to follow LPT and more likely to follow NPT.

3.4. Response-level prevalence and patterns

The response-level prevalence of LPT is $\mathrm{E}[P_{\mathrm{LPT}}] = 0.488$ (95% CrI $[0.434, 0.550]$). In other words, about half of responses were expected to arise from the LPT process. As shown in Figure 4, this average masks

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

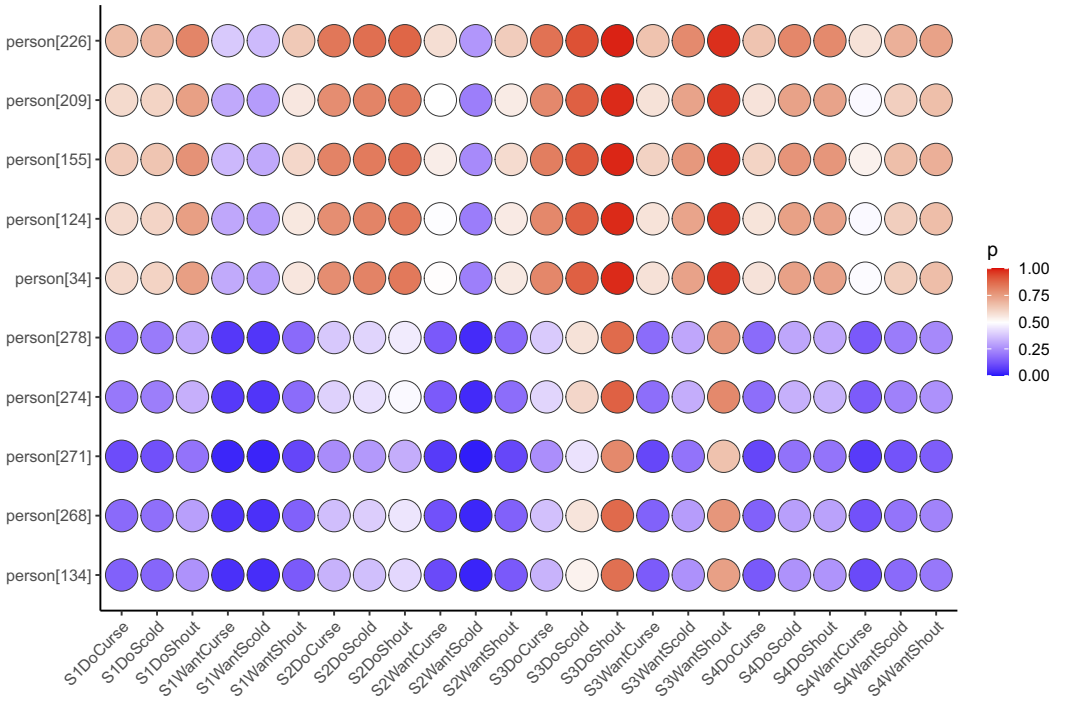


Figure 4. Response-level prevalence of selected persons.
Note: Each row is a person, and each column is an item. Color encodes the response-level probability of using LPT: darker red indicates higher LPT probability, and darker blue indicates lower LPT probability. The figure highlights representative individuals with low overall LPT prevalence ($P_{\text{LPT}} < 0.29$) and high overall LPT prevalence ($P_{\text{LPT}} > 0.66$).

substantial variation across both persons and items. Some respondents were consistently more likely to use LPT, whereas others were more likely to use NPT. Within the same respondent, LPT probability also varied by item. Some items such as `S3DoShout` tend to elicit LPT processes, while some items such as `S2WantScold` tend to show higher NPT prevalence.

4. Simulation

We conduct three simulation studies to evaluate MIX-R. Study 1 examines whether predictive model comparison selects MIX-R when the data contain response-level process switching and favors simpler models when the data-generating process is simpler. Study 2 evaluates recovery of the overall prevalence of the target response process across a wide range of true values. Study 3 examines recovery of item parameters, person parameters, and person- and item-level prevalence under varying test length, sample size, and class balance. Across studies, we report bias, RMSE, 95% CrI coverage, and rank-order fidelity when applicable. All studies use a three-category response scale ($R = 3$) matching the LPT and NPT structures. Complete design details are provided in Appendix C of the Supplementary Material.

4.1. Simulation study 1: Model fit and model selection

4.1.1. Design

We evaluate whether the model selection criterion used in Section 3.3.1 favors the data-generating structure. We consider four generators: LPT, NPT, MIX-P, and MIX-R. For each generator, we simulate 25 datasets ($N = 500$, $I = 20$, and $R = 3$) and fit all four models under common priors, parameterizations,

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Table 6. Predictive fit comparisons

Generator model	Fitted model	*LOOIC* (mean)	*p*LOO (mean)	Δ*LOOIC*	Selected rate
LPT	**LPT**	**12,674.88**	**423.41**	**0.00**	**0.48**
	NPT	13,720.96	607.78	1,046.08	0.00
	MIX-P	12,675.11	437.34	0.24	0.32
	MIX-R	12,675.48	427.94	0.61	0.20
NPT	**NPT**	**11,836.66**	**628.79**	**0.00**	**0.92**
	LPT	13,230.39	427.65	1,393.73	0.00
	MIX-P	11,851.02	633.73	14.36	0.00
	MIX-R	11,850.68	630.98	14.02	0.08
MIX-P	**MIX-P**	**12,329.48**	**599.14**	**0.00**	**1.00**
	LPT	14,562.08	448.03	2,232.60	0.00
	NPT	13,675.35	635.30	1,345.87	0.00
	MIX-R	12,916.37	655.10	586.89	0.00
MIX-R	**MIX-R**	**13,289.12**	**653.03**	**0.00**	**0.68**
	LPT	13,976.23	441.24	687.11	0.04
	NPT	13,693.53	648.15	404.41	0.12
	MIX-P	13,365.39	649.09	76.27	0.16

Note: Within each generator block, Δ*LOOIC* is computed relative to the generator-matched model. The selected rate is the proportion of replications in which each fitted model had the smallest *LOOIC* among the four candidates. Selected rates sum to one within each generator block.

and MCMC controls. Predictive performance is summarized by LOOIC and *p*LOO. Within each generator condition, we report ΔLOOIC=LOOIC−$\text{LOOIC}_{\text{ref}}$, where the generator-matched fit is the reference; negative values favor the comparison model. We also report the selected rate, the proportion of replications in which each fitted model had the smallest LOOIC among the four candidates. Additional design details can be found in Appendix C.1 of the Supplementary Material. To describe the realized degree of process separation, we computed response-level separation summaries for the MIX-R generated datasets using the true generating parameters; details are reported in Appendix D of the Supplementary Material. Across the 25 MIX-R datasets, average classification certainty had a mean of 0.93 and ranged from 0.79 to 1.00, with 0.50 indicating no separation. Thus, the MIX-R condition was informative but not artificially easy, because some response-level process memberships remained ambiguous.

4.1.2. Results

Across generator blocks, the true model typically attains the lowest mean LOOIC as shown in Table 6. When data were generated from a single process, the misspecified single-process alternative fit poorly. For example, when LPT generated the data, NPT had a mean Δ*LOOIC* of 1,046.08. When NPT generated the data, LPT had a mean Δ*LOOIC* of 1,393.73. When mixture structure is present, simpler models with lower *p*LOO do not compensate for misspecification: under MIX-P, all competitors are far worse.

The results also show that MIX-R did not spuriously dominate when the true data-generating process was simpler. Under the LPT generator, LPT had the lowest mean *LOOIC*, although MIX-P and MIX-R were very close; the selected rates were 0.48 for LPT, 0.32 for MIX-P, and 0.20 for MIX-R. Under the NPT generator, NPT was selected in 92% of replications, and MIX-R was selected in only 8%. Under the MIX-P generator, MIX-P was selected in all replications, whereas MIX-R was substantially worse, with

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

a mean Δ*LOOIC* of 586.89. These findings indicate that when process heterogeneity is absent or purely between persons, the additional response-level flexibility of MIX-R is not favored by out-of-sample prediction.

When data were generated from MIX-R, the proposed model had the lowest mean *LOOIC* and the highest selected rate. MIX-R was selected in 68% of replications. Appendix E of the Supplementary Material shows that competitor selections occurred when realized LPT prevalence was near 0 or 1, so the generated data were close to simpler cases. Tables S8 and S9 in the Supplementary Material show little effect on θ or dominant-process item recovery, although rare-process parameters were less stable. The competing models had larger mean *LOOIC* values, with Δ*LOOIC* values of 687.11 for LPT, 404.41 for NPT, and 76.27 for MIX-P. Thus, the model comparison was sensitive to response-level process switching while remaining reasonably conservative when simpler structures approximated the generated data.

4.2. Simulation study 2: Recovery of overall prevalence

4.2.1. Design

We evaluate recovery of the overall prevalence of the LPT process, p_{true}, which varied at nine levels from 0 to 1 ($p_{\text{true}} \in \{0, 0.05, 0.20, 0.40, 0.50, 0.60, 0.80, 0.95, 1\}$). Data are generated from MIX-R; the prevalence intercept α_0 is solved based on Equation (3), so the expected prevalence equals p_{true}. Person/item effects and item parameters follow Table S10 in the Supplementary Material. We fix $I = 20$, $N = 500$, and simulate 100 replications per level (900 datasets). Each dataset is fitted with the MIX-R model. For posterior draw s, we compute the draw-specific response-level prevalence, corresponding to Equation (4), as

$$\bar{\pi}^{(s)} = \frac{1}{NI} \sum_{i,j} \text{logit}^{-1}\left(\alpha_0^{(s)} + u_j^{(s)} + v_i^{(s)}\right).$$

The replication-level estimate is $\hat{\bar{\pi}}_r = \text{E}[\bar{\pi}^{(s)}]$, with 95% CrI obtained from the posterior quantiles of $\bar{\pi}^{(s)}$. Full details appear in Appendix C.2 of the Supplementary Material.

4.2.2. Results

Across the 100 replications per condition, we report bias, RMSE, and 95% CrI coverage for $\hat{p}_r$ in Table 7. Estimates closely track the prevalence across conditions. In mid-range true prevalence (40%~60%), bias was near zero and RMSE = 0.04 with coverage from 0.90 to 0.94. At the lower end (5%–20%), estimates showed small negative bias. At the upper end (80%–95%), estimates showed small positive bias. Coverage was stable across most of the interior but decreases at 95%, where the parameter is close to the boundary. At the hard boundaries (0% and 100%), both the point estimates and uncertainty collapse to the boundary.

Mean and median agree closely in the middle range, while small asymmetries appear near the edges: at 5%, the median equals the truth (0.05) but the mean is slightly below (0.03); at 80%–95%, the pattern reverses (mean ≥median). Figure 5 visualizes these patterns by plotting estimated prevalence against true prevalence across the full range.

4.3. Simulation study 3: Person and item parameter recovery

4.3.1. Design

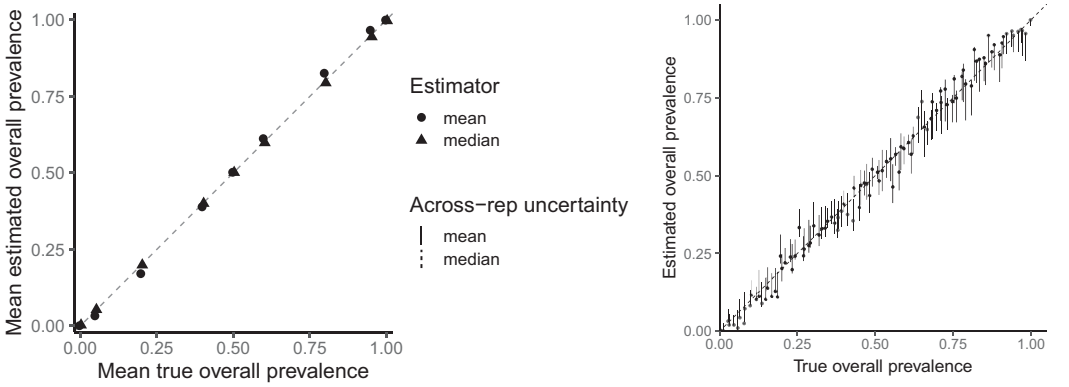
We assess recovery of item parameters for both trees (b_i^{A1}, b_i^{A2}; b_i^{B1}, b_i^{B2}), person parameters (θ_j, κ_j), and unit-level LPT prevalence under MIX-R. We use a 3×2×2 factorial design that varies $p_{\text{true}} \in \{0.20, 0.50, 0.80\}$, $N \in \{500, 2,000\}$, and $I \in \{20, 50\}$, with 25 replications per condition (300 datasets). Within each (I, N) cell, the same person and item parameters were reused across prevalence conditions, so changes across p_{true} reflect class balance rather than different parameter draws. Data were generated from MIX-R and fitted with MIX-R and a misspecified LPT model. Because LPT

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Table 7. Recovery of overall LPT prevalence

p_{true}	Mean of $\hat{\bar{\pi}}_r$	Median of $\hat{\bar{\pi}}_r$	Mean bias	RMSE	95% CrI coverage
0%	0.00	0.00	0.00	0.00	–
5%	0.03	0.05	−0.02	0.02	0.89
20%	0.17	0.20	−0.03	0.04	0.90
40%	0.39	0.40	−0.01	0.04	0.91
50%	0.50	0.50	0.00	0.04	0.90
60%	0.61	0.60	0.01	0.04	0.94
80%	0.83	0.79	0.03	0.04	0.91
95%	0.97	0.94	0.02	0.02	0.80
100%	1.00	1.00	0.00	0.00	–

Note: Bias is computed as $\hat{\bar{\pi}}_r - p_{\text{true}}$. Coverage is omitted at the hard boundaries because the posterior estimates collapse to 0 or 1.



(a) Across-rep means/medians. (b) Replication-level means.

Figure 5. Recovery of overall prevalence under MIX-R.
Note: Panel (a) shows the across-replication mean and median estimates at each true prevalence level. Panel (b) shows replication-level posterior means with 95% CrIs. The dashed diagonal is the identity line.

does not include b_i^{B1}, b_i^{B2}, or κ_j, cross-model comparisons focus on b_i^{A1}, b_i^{A2}, and θ_j. We report bias, RMSE, 95% CrI coverage, and rank-order fidelity (within-replication Pearson r, Fisher-z averaged); full design details and condition-wise summaries are provided in Appendix C.3 of the Supplementary Material.

4.3.2. Results

Item parameters: Across all conditions, MIX-R recovers the LPT thresholds (b_i^{A1}, b_i^{A2}) accurately, with near-zero bias, high rank fidelity, and close-to-nominal coverage. Recovery improved when LPT was more prevalent and when more item responses were available. In contrast, the misspecified LPT model showed large bias and severe under-coverage when the true data included substantial NPT use.

As shown in Figure 6, the LPT fit improved as p_{true} increased, but its CrIs remained narrow, with coverage far below the nominal 95% level even when LPT was dominant. These results show that ignoring response-process heterogeneity can distort item-threshold recovery. NPT threshold recovery under MIX-R depended more strongly on class balance. When NPT was sufficiently

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

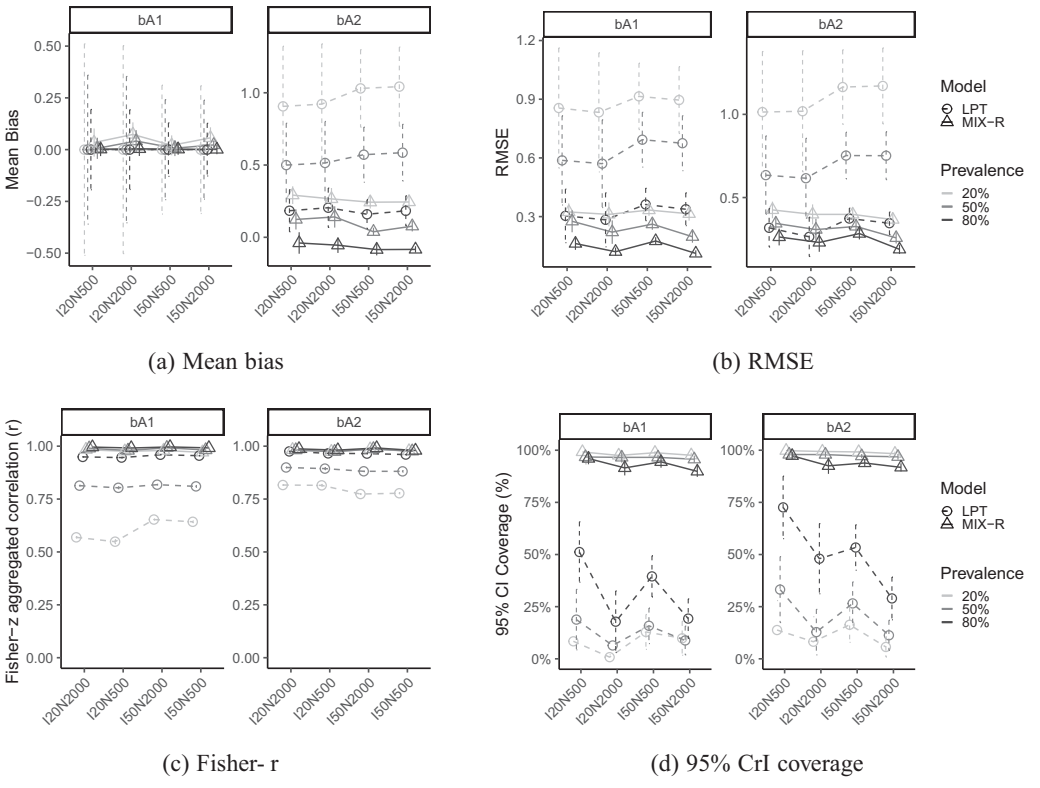


Figure 6. Recovery of LPT item thresholds b^{A1} and b^{A2} under MIX-R and the misspecified LPT model.

represented, b_i^{B1} and b_i^{B2} were recovered with good accuracy and high rank fidelity. When LPT dominated, information for the NPT first node decreased, leading to larger RMSE and weaker rank recovery for b_i^{B1}. Complete condition-wise results are reported in Table S11 in the Supplementary Material. Figure S3 in the Supplementary Material provides representative item-level recovery plots, showing that MIX-R estimates align with the true values when both processes are sufficiently represented.

Person estimates: MIX-R also recovered the target trait θ_j well as shown in Figure 7.

Recovery improved mainly with test length, as person information increases with the number of items. κ_j was more sensitive to class balance: recovery was strongest when NPT was sufficiently represented and weaker when LPT dominated. The misspecified LPT model performed worse for θ_j, especially when LPT was not the dominant process, showing that response-process misspecification can propagate to person-score estimation. Full summaries are provided in Table S12 in the Supplementary Material, and representative person-level recovery plots can be found in Figure S4 in the Supplementary Material.

Unit-level prevalence: MIX-R reproduces both person-level ($\hat{\pi}_j$) and item-level ($\hat{\pi}_i$) LPT prevalences with small bias across conditions as shown in Table 8. Item-level prevalence was recovered especially well, with strong rank fidelity that improved with larger N. Person-level prevalence showed lower rank fidelity, especially when $p_{\text{true}} = 0.80$, but bias and RMSE remained modest. As shown in Table 8, increasing I improved person-level recovery, whereas increasing N improved item-level recovery. Figure 8 illustrates the same pattern in the high-information condition. Overall, MIX-R recovered unit-level prevalence well enough to support both person- and item-level summaries of process use.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

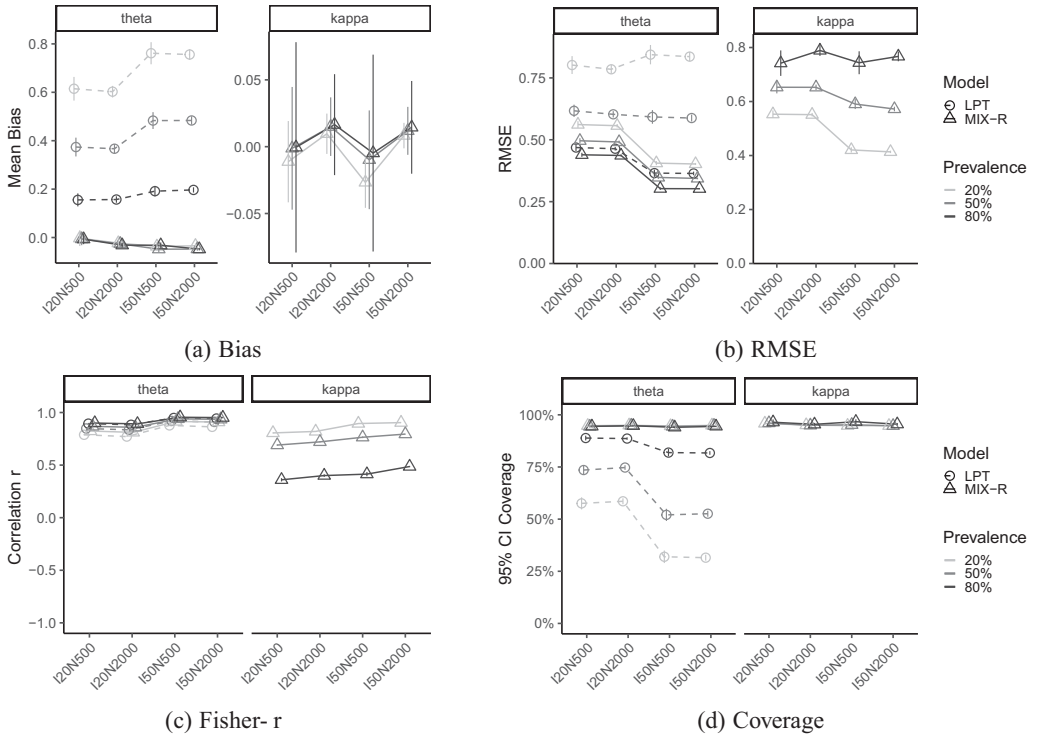


Figure 7. Recovery of person traits θ and κ under MIX-R and the misspecified LPT model.

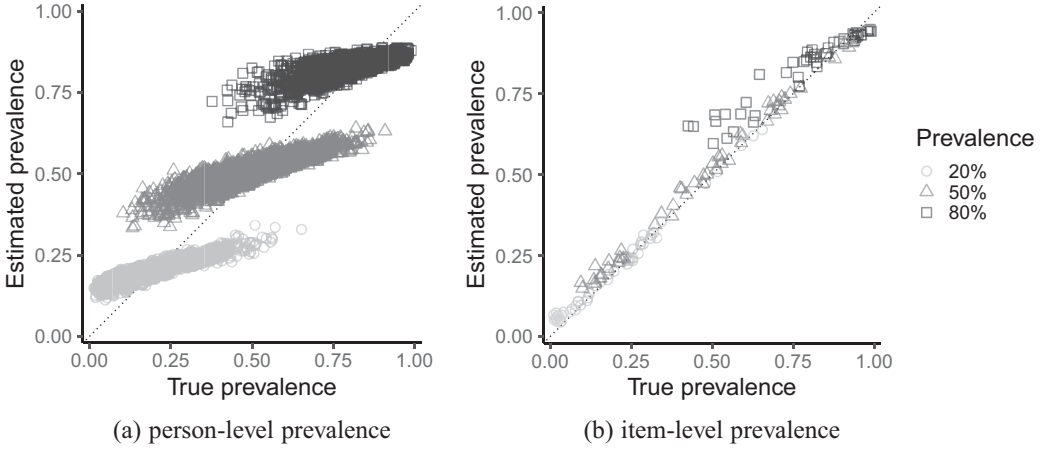


Figure 8. Recovery of unit-level LPT prevalence in the high-information condition ($I = 50$, $N = 2,000$).

5. Discussion

Understanding how respondents arrive at a response category is important for establishing validity and refining item content and design. IRTree models are well suited for this purpose because they represent response processes as sequential stages. Existing approaches often assume that all respondents follow the same process, or that a given respondent follows the same process across all items. These assumptions can be restrictive when item content and wording elicit distinct response processes. To address this

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Table 8. Recovery of person- and item-level LPT prevalence under MIX-R

Unit	Prevalence	I	N	Mean bias	RMSE	Mean 95% CrI coverage	Fisher-r (mean)
Person	20%	$I = 20$	$N = 500$	0.03	0.10	94.8%	0.46
			$N = 2,000$	0.00	0.09	95.4%	0.48
		$I = 50$	$N = 500$	0.02	0.09	93.7%	0.53
			$N = 2,000$	0.00	0.09	94.6%	0.56
	50%	$I = 20$	$N = 500$	0.05	0.13	91.9%	0.39
			$N = 2,000$	0.00	0.12	94.4%	0.44
		$I = 50$	$N = 500$	0.03	0.13	91.8%	0.60
			$N = 2,000$	0.00	0.11	93.9%	0.63
	80%	$I = 20$	$N = 500$	0.06	0.12	86.3%	0.27
			$N = 2,000$	0.02	0.10	92.8%	0.31
		$I = 50$	$N = 500$	0.05	0.11	88.4%	0.52
			$N = 2,000$	0.03	0.10	92.0%	0.57
Item	20%	$I = 20$	$N = 500$	0.03	0.09	94.2%	0.92
			$N = 2,000$	0.00	0.04	96.6%	0.98
		$I = 50$	$N = 500$	0.02	0.07	95.8%	0.93
			$N = 2,000$	0.00	0.04	95.3%	0.98
	50%	$I = 20$	$N = 500$	0.05	0.12	94.8%	0.94
			$N = 2,000$	0.00	0.06	97.0%	0.99
		$I = 50$	$N = 500$	0.03	0.10	94.4%	0.94
			$N = 2,000$	0.00	0.05	95.3%	0.98
	80%	$I = 20$	$N = 500$	0.06	0.14	86.4%	0.82
			$N = 2,000$	0.02	0.09	90.4%	0.94
		$I = 50$	$N = 500$	0.05	0.13	87.3%	0.76
			$N = 2,000$	0.03	0.07	91.0%	0.94

Note: Person-level prevalence is $\hat{\pi}_j$, and item-level prevalence is $\hat{\pi}_i$. Truths are computed from the simulated response-process indicators C_{ij}. Fisher-r summarizes rank-order recovery within each replication and is averaged using Fisher-z transformation.

limitation, in this study, we introduce a response-level mixture IRTree model, namely, MIX-R, to capture both the between- and within-person variation in response processes.

The MIX-R framework uses IRTree models to represent distinct response processes and models the probability of using each process as a function of person-, item-, and response-level covariates. In this sense, MIX-R extends mixture IRTree work that models heterogeneity across respondents or in the use of particular categories by allowing the whole response process to vary at the response level (Alagöz & Meiser, 2024; Kim & Bolt, 2021; Tijmstra & Bolsinova, 2025; Tijmstra et al., 2018). Although this study demonstrates MIX-R using three-category items with two response processes, the framework can be generalized to items with more response categories or more than two processes.

Using verbal aggression data, we demonstrated several empirical uses of the MIX-R. First, MIX-R shows how item functioning can differ across response processes, including differences in category probabilities and expected observed scores. Second, it estimates response-level process prevalence and can therefore help diagnose how response categories are used in Likert-type data (e.g., whether *Perhaps* functions as qualitatively distinct from *No* and *Yes*). Third, it links response-

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

process prevalence to item- and person-level covariates. These results show how MIX-R can support the design and validation of Likert-type measurement scales.

5.1. Simulation summary

The simulation studies supported the performance of MIX-R in three ways. First, predictive model comparison favored MIX-R when the data were generated with response-level process switching, but did not automatically favor MIX-R when the true structure was simpler. This suggests that MIX-R's added flexibility does not by itself lead to systematic over-selection. Beyond model selection, the overall prevalence of LPT was recovered accurately across a wide range of true values, especially when the true prevalence was away from the boundaries. This supports model-implied prevalence summaries as interpretable measures of process use. Furthermore, item parameters, person parameters, and unit-level prevalence were generally recovered well when both response processes were sufficiently represented. Recovery became harder when one process dominated, especially for parameters tied to the less frequent process. Thus, process-specific item and person parameters should be interpreted alongside the estimated prevalence of the corresponding process. The misspecified LPT model showed larger bias and poorer coverage when the true responses came from a response-level mixture, indicating that ignoring process heterogeneity can distort both item and person inference.

5.2. Empirical findings

In the verbal aggression data, MIX-R provided an interpretable account of response-process variation. MIX-R and MIX-P yielded better model fit than the single-process models, supporting a mixture of LPT and NPT processes in the data. The MIX-R results also showed between-item variation in LPT prevalence, suggesting that respondents may switch processes across items and that response-process variation can occur both between and within persons.

The empirical results showed that process choice matters for item functioning and scoring. Category response curves showed that the same item can produce different probabilities of *No*, *Perhaps*, and *Yes* under LPT and NPT. When the same $0, 1, 2$ scoring rule was applied, expected observed scores also differed across processes, showing how conventional scores can change when response-process heterogeneity is ignored. We also found that response-process prevalence was related to item content and respondent characteristics. *Want* items and blaming behaviors were associated with lower LPT prevalence, whereas self-to-blame items were associated with higher LPT prevalence. Trait Anger was associated with lower LPT prevalence, suggesting that process use may vary with respondent characteristics as well as item features.

5.3. Implications for modeling and practice

Methodologically, the results caution against treating response categories as having single process meaning. The response-style literature has long shown that observed categories can reflect systematic response tendencies rather than only the target trait (Baumgartner & Steenkamp, 2001). MIX-R extends this concern from category use to response-process use (Plieninger & Meiser, 2014): a category such as *Perhaps* may not simply reflect a global midpoint preference; it may arise from distinct processes that are engaged differently across items and respondents. If process switching is ignored, item parameters and person scores can be process-confounded, and model comparisons may favor an overly simple structure when mixtures are present. MIX-R provides a principled way to examine this possibility through predictive model comparison and person- and item-level prevalence diagnostics.

Substantively, our findings align with the idea that item content can shape response processes (De Boeck & Wilson, 2004; Tourangeau et al., 2000). The distinction between wanting to acting aggressively and actually doing so is central to the verbal aggression framework, where inhibition is reflected in the gap between aggressive impulses and expressed behavior (Smits et al., 2004). In our results,

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

contexts that emphasize immediate outward expression, such as *Do–Shout*, tended to lead respondents toward decisions that accentuate the *No*/*Yes* distinction, whereas more reflective *Want* contexts often kept the *Perhaps* path available. These process differences can change expected scores measured by the instrument and, in turn, conclusions about item severity, person ordering, and associations with covariates such as Trait Anger. It is worth noting that, in our empirical data analysis, the item parameters and covariate effects were estimated jointly. This modeling strategy was motivated by the substantive assumptions regarding the influence of item and person characteristics on response processes. In applications where the number of covariates is relatively large or the analysis is primarily exploratory, however, the stepwise estimation approach (Alagöz et al., 2025; Vermunt, 2010), in which the item parameters are estimated first, followed by the estimation of covariate effects, can be more desired.

More broadly, these findings are consistent with survey-methodology work showing that respondent answers are shaped by both question features and respondent characteristics. Olson et al. (2019) showed that question and respondent characteristics are associated with respondent answering behaviors, including behaviors related to comprehension and response mapping. From this perspective, different items in the same scale may invite different response processes, and respondents may differ in which process they are more likely to use. MIX-R represents this variation statistically by allowing response-process membership to vary across responses and by relating this variation to item and respondent characteristics.

A complementary use of MIX-R is to support substantive analyses of existing data without assigning each response definitively to a single process. Person-level estimates and derived scores average over uncertainty in process membership, enabling process-adjusted comparisons while retaining uncertainty about how responses were generated. Researchers may compare person rankings, group differences, or associations with external criteria across a conventional single-process model and MIX-R to assess whether conclusions depend on assumptions of process homogeneity. MIX-R may yield wider posterior intervals when some processes provide less information about the target construct, but this appropriately reflects uncertainty that a misspecified single-process model would conceal. Evaluating the consequences of such uncertainty for statistical power and downstream analyses is an important direction for future research.

5.4. Limitations and future research

Like the IRTree model, the proposed MIX-R model implicitly assumes that within-item subprocesses are conditionally independent. In the IRTree model, violation of this assumption may result in lower item slope estimates for later decision nodes (Jeon, 2023; Lyu et al., 2023). Accordingly, an important limitation of the present study is that the source of the class heterogeneity cannot be definitively determined. Specifically, it remains unclear whether the identified class heterogeneity reflects substantive differences in the underlying psychological response processes or instead captures residual conditional dependence across decision nodes. Although the conditional independence assumption cannot be directly evaluated, one can use the recently developed latent space item response model (LSIRM; Jeon et al., 2021) to explore the potential person-by-item interactions.

Several other limitations also remain and suggest directions for future research. First, our simulations covered only part of the possible design space. We varied test length, sample size, and process prevalence, but did not fully vary class separation, item-parameter extremity, or the number and structure of response processes. These factors are likely to affect model selection, parameter recovery, and prevalence estimation. Future simulation work should therefore examine how much information is needed to distinguish response processes under different levels of class balance, process separation, and test length.

Second, interpreting latent traits associated with response processes requires external validation. In this study, κ was modeled as distinct from the target trait and was interpreted in relation to respondents' willingness to give a definitive response. However, the empirical analysis relied on predictive fit and model diagnostics rather than observed ground-truth process labels. Future studies could combine MIX-R with additional evidence, such as cognitive interviews, think-aloud protocols, response times,

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

mouse tracking, or other process data, to evaluate whether the modeled processes correspond to respondents' actual response behavior. This is especially important because model-implied process labels may reflect several mechanisms, including effortful response behavior and satisficing (Krosnick, 1991), sensitivity (Tourangeau & Yan, 2007), or response style (Van Vaerenbergh & Thomas, 2013). Conceptually distinct response processes and response styles may produce overlapping category-response patterns. Appendix G.2 of the Supplementary Material illustrates how these sources of heterogeneity can be represented jointly in a richer five-category scale, where response-level process membership determines the organization of the response decisions, whereas a separate extremity factor governs endpoint use within either process. In the present three-category NPT, however, an omitted midpoint tendency affecting the directed-versus-middle decision may be absorbed by κ. More generally, omitted response styles may affect process-specific traits, prevalence estimates, or item parameters. Although this potential confounding does not imply that κ is itself a response-style parameter because it affects responses only when NPT is engaged, it warrants caution in interpreting estimated process prevalence. Such prevalence should be viewed as evidence consistent with the specified trees rather than as evidence necessarily free of response-style influence.

Third, the present application focused on two response processes for three-category items. The choice of candidate trees should be made deliberately, based on theory, item content, and the response processes that are plausible for the assessment context. Although MIX-R can be generalized to more than two processes and to items with more response categories, these extensions raise additional questions about identifiability, prior choice, and computation. Future work could connect MIX-R with dynamic IRTree models, which study within-person changes in response strategies over item position (Merhof & Meiser, 2023), or with response-time mixture models that use auxiliary information to distinguish response strategies (Huang et al., 2025). Future work could also examine how to select and regularize covariates when person-, item-, and response-level predictors are included in the prevalence model.

Finally, future work can use MIX-R for item evaluation and scale refinement. Items that consistently elicit unintended or mixed response processes may be revised, retained as diagnostic items, or studied further to understand how wording and content shape response behavior. In this sense, MIX-R can connect response-process theory with item design. Separating response process from response style can support richer psychological inference and more defensible measurement in multicategory self-report settings.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10146.

Data availability statement. The empirical data analyzed in this study are publicly available, and the data source is described and cited in the article. The analysis code is available at https://github.com/jevanluo/MIX-R-paper-code.

Funding statement. No funding was received to assist with the preparation of this manuscript.

Competing interests. M.J. is an Associate Editor for Psychometrika, but was excluded completely from the review process. The remaining authors have no other relevant financial or non-financial interests to disclose.

Ethical standards. This study is a secondary analysis of existing data, and ethics approval is waived.

Consent to participate. This study is a secondary analysis of existing data, and informed consent is waived.

Consent to publication. This study is a secondary analysis of existing data, and consent publication is waived.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Alagöz, O. E. C., & Meiser, T. (2024). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement*, *84*(5), 957–993. https://doi.org/10.1177/00131644231206765

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Alagöz, O. E. C., Meiser, T., & Khorramdel, L. (2025). Disentangling individual differences in cognitive response mechanisms for rating scale items: A flexible-mixture multidimensional IRTree approach. *Behavior Research Methods*, *57*, 256. https://doi.org/10.3758/s13428-025-02778-0

Bao, Z., Harris, J., Lavender, V., Rafferty, A. M., & Armes, J. (2024). Enhancing validity through cognitive interviewing: A methodological example using the registered nursing forecasting nurse survey. *Measurement and Evaluations in Cancer Care*, *2*, 100008. https://doi.org/10.1016/j.ymecc.2024.100008

Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, *35*(8), 583–603. https://doi.org/10.1177/0146621611428446

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. https://doi.org/10.18637/jss.v048.c01

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer. https://doi.org/10.1007/978-1-4757-3990-9

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403–413. https://doi.org/10.1080/10618600.2016.1172487

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, *54*(3), 333–363. https://doi.org/10.1111/jedm.12150

Drton, M., & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of The Royal Statistical Society Series B-Statistical Methodology*, *79*(2), 323–380. https://doi.org/10.1111/rssb.12187

Embretson, S. E. (2023). Understanding examinees' item responses through cognitive modeling of response accuracy and response times. *Large-scale Assessments in Education*, *11*(1), Article no. 9. https://doi.org/10.1186/s40536-023-00158-8

Huang, H.-Y. (2020). A mixture IRTree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement*, *80*(6), 1168–1195. https://doi.org/10.1177/0013164420914711

Huang, S., Chung, S., & Falk, C. F. (2024). Modeling response styles in cross-classified data using a cross-classified multidimensional nominal response model. *Journal of Educational Measurement*, *61*(3), 486–510. https://doi.org/10.1111/jedm.12401

Huang, S., Luo, J., & Jeon, M. (2025). A response time-based mixture item response theory model for dynamic item-response strategies. *Behavior Research Methods*, *57*(1), 54. https://doi.org/10.3758/s13428-024-02555-5

Jeon, M. (2023). Commentary: Explore conditional dependencies in item response tree data. *Psychometrika*, *88*(3), 803–808. https://doi.org/10.1007/s11336-023-09915-8

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*(3), 1070–1085. https://doi.org/10.3758/s13428-015-0631-y

Jeon, M., Jin, I. H., Schweinberger, M., & Baugh, S. (2021). Mapping unobserved item–respondent interactions: A latent space item response model with interaction map. *Psychometrika*, *86*(2), 378–403. https://doi.org/10.1007/s11336-021-09762-5

Kim, N., & Bolt, D. M. (2021). A mixture IRTree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, *81*(1), 131–154. https://doi.org/10.1177/0013164420913915

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lenzner, T., Hadler, P., & Neuert, C. (2023). An experimental test of the effectiveness of cognitive interviewing in pretesting questionnaires. *Quality & Quantity*, *57*, 5055–5073. https://doi.org/10.1007/s11135-022-01489-4

Lyu, W., Bolt, D. M., & Westby, S. (2023). Exploring the effects of item-specific factors in sequential and IRTree models. *Psychometrika*, *88*(3), 745–775. https://doi.org/10.1007/s11336-023-09912-x

Merhof, V., & Meiser, T. (2023). Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes. *Psychometrika*, *88*(4), 1354–1380. https://doi.org/10.1007/s11336-023-09901-0

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Olson, K., Smyth, J. D., & Ganshert, A. (2019). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology*, *7*(2), 275–308. https://doi.org/10.1093/jssam/smy006

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899. https://doi.org/10.1177/0013164413514998

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt.2), 100.

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.

Smits, D. J. M., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behaviour. *European Journal of Personality*, *18*(7), 537–555. https://doi.org/10.1002/per.529

Spielberger, C. D. (1988). *State–trait anger expression inventory: Professional manual*. Psychological Assessment Resources.

Spielberger, C. D., & Sydeman, S. J. (1994). State–trait anxiety inventory and state–trait anger expression inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 292–321). Lawrence Erlbaum Associates.

Tijmstra, J., & Bolsinova, M. (2025). Modeling within-and between-person differences in the use of the middle category in Likert scales. *Applied Psychological Measurement*, 266–281. https://doi.org/10.1177/01466216251322285

Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, *50*(6), 2325–2344. https://doi.org/10.3758/s13428-017-0997-0

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

Cite this article: Luo, J., Huang, S. and Jeon, M. (2026). Investigating Within-Person Variation in Response Processes for Likert-Scale Data with a Response-Level Mixture IRTree Model. *Psychometrika*, 1–25. https://doi.org/10.1017/psy.2026.10146

Downloaded from https://www.cambridge.org/core. 18 Nov 2026 at 03:28:14, subject to the Cambridge Core terms of use.