MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 5, 878–897

OPEN ACCESS

Accounting for Measurement Invariance Violations in Careless Responding Detection in Intensive Longitudinal Data: Exploratory vs. Partially Constrained Latent Markov Factor Analysis

Leonie V. D. E. Vogelsmeier[a], Joran Jongerling[a], and Esther Ulitzsch[b,c]

[a]Department of Methodology and Statistics, Tilburg University, Tilburg, Netherlands; [b]Centre for Educational Measurement, University of Oslo, Oslo, Norway; [c]Centre for Research on Equality in Education, University of Oslo, Oslo, Norway

ABSTRACT

Intensive longitudinal data (ILD) collection methods like experience sampling methodology can place significant burdens on participants, potentially resulting in careless responding, such as random responding. Such behavior can undermine the validity of any inferences drawn from the data if not properly identified and addressed. Recently, a confirmatory mixture model (here referred to as fully constrained latent Markov factor analysis, LMFA) has been introduced as a promising solution to detect careless responding in ILD. However, this method relies on the key assumption of measurement invariance of the attentive responses, which is easily violated due to shifts in how participants interpret items. If the assumption is violated, the ability of the fully constrained LMFA to accurately identify careless responding is compromised. In this study, we evaluated two more flexible variants of LMFA—fully exploratory LMFA and partially constrained LMFA—to distinguish between careless and attentive responding in the presence of non-invariant attentive responses. Simulation results indicated that the fully exploratory LMFA model is an effective tool for reliably detecting and interpreting different types of careless responding while accounting for violations of measurement invariance. Conversely, the partially constrained model struggled to accurately detect careless responses. We end by discussing potential reasons for this.

KEYWORDS
Careless responding; ecological momentary assessment; mixture modeling; latent Markov modeling; three-step approach; time series; intensive longitudinal data

Introduction

Experience sampling methodology (ESM) or related methods like ecological momentary assessment (Scollon et al., 2003) are the go-to designs to gather intensive longitudinal data (ILD) for investigating (between-person variation in) within-person dynamics in psychological constructs. However, requiring individuals to complete self-report questionnaires multiple times a day over several days or weeks, ESM sampling schemes can be perceived as burdensome. This can diminish individuals' willingness or ability to respond attentively to the questionnaire items, leading them to complete the questionnaires without careful consideration of the content (Hasselhorn et al., 2023; Huang et al., 2015; Jaso et al., 2022; Ulitzsch, Nestler, et al., 2024) at one or more occasions. Such careless and insufficient effort responding (C/IER, in the following also shortened to "careless" or "inattentive" responding) may manifest as random responses or selections based on a preference for certain response categories or scale locations (e.g., choosing the middle or lower end of the scale). Detecting and addressing careless responding is crucial for ensuring accurate inferences about the dynamics of psychological constructs. Failure to identify careless responding can lead to biases in psychometric properties, such as factor structure and reliabilities, as well as distortions in correlations of interest (McGrath et al., 2010).

The most promising methods for detecting careless responding in ILD (and also cross-sectional data)[1] are latent mixture modeling approaches, where latent classes (i.e., unobserved groups), also referred to as mixture components, are predefined as attentive or careless, drawing from theoretical expectations regarding the assumed data-generating processes associated with these response patterns. More precisely, attentive responses are modeled to gauge the construct(s) of interest using latent trait models such as confirmatory factor analysis models (Arias et al., 2020; Kam & Cheung, 2024) or item response theory (IRT) models (Ulitzsch, Pohl, et al., 2022; Ulitzsch, Pohl, et al., 2023; van Laar and Braeken, 2022), whereas careless responses are conceptualized as independent of item content and the underlying construct(s) of interest. Instead, they are assumed and modeled to be influenced solely by scale preferences and/or random selection (Kam & Cheung, 2024). Note that, to enhance classification accuracy, external information (e.g., response times or item characteristics such as position or readability) can be incorporated (Meade & Craig, 2012; Ulitzsch, Pohl, et al., 2022; Ulitzsch, Yildirim-Erbasli, et al., 2022; Zhang et al., 2024). Mixture models employ probabilistic assignments (e.g., an observation can have 97% certainty of being careless). This (un)certainty can be taken into account in subsequent analyses, for instance, by weighting observations; for a discussion, see Ulitzsch, Domingue, et al. (2023; Ulitzsch, Shin, et al., 2024).

Mixture models for detecting careless responding in ILD have gained popularity only recently. The most promising one is an approach that leverages respondents' responses to items measuring latent constructs (Vogelsmeier, Uglanova, et al., 2024) because no external information like screen times is required. However, it comes with the assumption of measurement invariance, which is easily violated in ILD. In this study, we address this issue and evaluate two model adjustments to account for non-invariance. In the following sections, we begin by reviewing the existing mixture model designed for ILD. Subsequently, we explain its assumption of measurement invariance and why this is problematic. Finally, we detail the objectives of this study.

C/IER detection in ESM data

Existing mixture modeling approach and its assumption of invariance

Compared to cross-sectional data, ILD pose specific challenges for mixture modeling approaches for detecting careless responding because individuals can change between attentive and careless responding over time because of, for example, momentary changes in fatigue or (lack of) motivation (Eisele et al., 2023). Consequently, to accurately capture these fluctuations over time, classification should be conducted at a person-by-occasion level. Considering the demands of ILD, Vogelsmeier, Uglanova, et al. (2024) introduced a mixture model for constructs assessed with multiple-indicator scales, which are well-established in ESM studies (Vogelsmeier, Jongerling, et al., 2024). The proposed model allows for unveiling the moments when individuals switch between careless and attentive responding and can identify correlates of transition patterns with individual and situational characteristics. The model offers the distinct advantage that it classifies observations based on response behavior, eliminating the need for relying on additional information such as screen times.[2]

The approach classifies observations into an attentive or careless mixture component based on a mixture IRT model, and individuals are allowed to transition between these mixture components over time, which is modeled using a latent Markov chain. Note that the confirmatory mixture IRT model was tailored to the use of ordinal data from Likert scales, but a conceptually similar confirmatory mixture factor analysis model (Kam & Cheung, 2024) can be used for continuous data from a visual analog scale (VAS). The approach extends the method latent Markov factor analysis (LMFA), which was previously proposed to detect changes in measurement models, such as shifts in item interpretation (Vogelsmeier, Vermunt, van Roekel, et al., 2019). While the original LMFA was a completely exploratory method (using exploratory factor analysis or IRT to obtain the measurement models), the extension relies on theory-based confirmatory specification of these models with constraints tailored to capture attentive versus careless responding. Because of the confirmatory nature, in the following, the extension is referred to as *fully constrained* LMFA.

One limitation of this fully constrained LMFA, however, is the assumption of measurement invariance across respondents and time for the attentive and C/IER components of the mixture model, respectively. Specifically, it is assumed that all attentive observations have the same underlying model, as do all inattentive observations. Non-invariance of the attentive model was shown to become problematic if it extends beyond purely gradual changes in how well

[2]In the Discussion Section, we provide a reference to a mixture modeling approach that relies on screen times and detail advantages and disadvantages.

indicators measure the underlying constructs. Specifically, if qualitatively different attentive structures underlie the data (i.e., configural invariance is violated), attentive responses can be incorrectly flagged as careless (Vogelsmeier, Uglanova, et al., 2024). This is concerning because changes in the structure of attentive models are likely to occur in real ILD (Adolf et al., 2014; McNeish et al., 2021; Vogelsmeier, Vermunt, Bülow, et al., 2023). For example, Schmitt et al. (2024) showed that individuals switched between multiple structurally different measurement models that reflected distinct granularity patterns between specific emotions during participation. Invariance violations of the inattentive observations have yet to be studied in detail. However, initial results showed that at least a mix of random responding and scale preferences is well captured in one C/IER state (Vogelsmeier, Uglanova, et al., 2024; Uglanova et al., 2025). It is therefore possible that other mixed types are also captured well in one state. Nevertheless, the two extensions that we propose in this article have the potential to account for invariance violations in both attentive and careless responding.

Two possible ways to account for non-invariance

Employing a fully exploratory LMFA

The first possible approach is the traditional, *fully exploratory* LMFA (Vogelsmeier, Vermunt, van Roekel, et al., 2019), in which neither the number of measurement models (i.e., the number of mixture model components) nor the number of constructs and the presence or absence of item-construct relationships within these class-specific measurement models is known. Instead of confirmatory factor analysis or IRT modeling, the approach uses exploratory versions of each framework and relies on model selection procedures (e.g., the Bayesian information criterion, BIC, Schwarz, 1978) to identify how many mixture components and factors within these components underlie the data. Therefore, the method detects changes in all response patterns that manifest as differences in measurement models. Since careless responding affects relationships between items and constructs (e.g., random responding weakens the relationships between all items), it should also constitute a change in the measurement model (e.g., loadings should become lower). Therefore, fully exploratory LMFA should be able to detect careless responding in addition to different types of attentive responding. However, the fully exploratory LMFA has never been evaluated for its performance in detecting careless responding. It has only been assessed for detecting changes in the attentive model, such as changes in the interpretation of items (Vogelsmeier, 2022; Vogelsmeier, Vermunt, Böing-Messing, et al., 2019; Vogelsmeier, Vermunt, van Roekel, et al., 2019; Vogelsmeier et al., 2023). The results cannot simply be generalized to studying careless responding. On the one hand, the response patterns qualitatively differ with regard to their distributions. While attentive responses entail (approximately) multivariate normally distributed data, careless responses can have various distributions. For example, random responding manifests in a uniform distribution, while a lower-scale preference entails rather skewed data. On the other hand, LMFA was previously shown to work particularly well in distinguishing different mixture components when the measurement models are strong; that is, with high loadings and low unique variances (Vogelsmeier, Vermunt, Böing-Messing, et al., 2019; Vogelsmeier, Vermunt, van Roekel, et al., 2019), which is not given for all types of careless responding, like random responding. The strong differences between attentive and careless responses may compensate for that, however, because strong measurement model differences were shown to be detected more easily than small differences.

One limitation of the fully exploratory approach may be that the careless responding patterns can be manifold, each potentially resulting in a different component.[3] For instance, scale preferences may result in very high factor loadings, while random responding may entail factor loadings of essentially zero. Generally, careless responses do not exhibit as clear patterns as those for attentive responses, making them harder to recognize. When all patterns are subsumed into a single mixture component by the fully exploratory LMFA, model parameters may be reflective of the composite of behavioral patterns (e.g., loadings may be neither extremely high nor extremely low), such that the careless class may not be recognizable in model interpretation. When each type of behavior is captured by a different mixture component, interpretation is likely to be easier. However, whether the model selection procedures suggest one or multiple mixture components for careless responses of different types remains to be investigated.

[3]Note that the fully constrained LMFA does not distinguish between different C/IER types but aims to classify all types into one mixture component.

Employing a partially constrained LMFA

Given the possible complexity of the post-hoc interpretation of careless mixture components in the fully exploratory LMFA, a partially constrained approach may be considered. In such a *partially constrained* LMFA, a confirmatory approach would be used for the careless responding model, applying constraints like those in the fully constrained LMFA, while leaving changes in the number and nature of the attentive model(s) to exploration. This way, the careless mixture component incorporates researchers' theoretical considerations on the likely presence of careless response behavior and, if the specified component model is indeed capable of absorbing the careless responding patterns present in the data, may exhibit higher power in unveiling instances of careless responding than a fully exploratory approach. Further, this approach does not require post-hoc interpretation of the careless class(es). Instead of specifying only one careless component, one could also specify more and let model selection decide if one or more careless components are required for mixed types of careless responding. The performance of model selection and the degree to which the partially constrained LMFA distinguishes between attentive and careless responses remains to be investigated.

The present study

In this study, we compare fully exploratory and partially confirmatory LMFA as two competing approaches to distinguish between attentive and careless responding while acknowledging that both attentive and careless responding behavior may change over time. Here, we specifically focus on continuous data and therefore employ mixture factor analyses in both variants. Both LMFA variants are promising candidate approaches, but it remains unclear which is (more) recommendable. Therefore, we examine the performances in a simulation study with regard to (1) *model selection*, (2) *classification accuracy*, and, only for the fully exploratory LMFA, (3) *interpretability*. First, model selection pertains to how well two previously evaluated model selection criteria perform in selecting the correct number of mixture components and factors within the components, which is crucial for any exploratory modeling approach. Furthermore, it will be explored whether 1 or more careless components are selected for mixed types. Second, classification accuracy pertains to the sensitivity and specificity of identifying observations as careless, which can be poor even if the correct model complexity is identified. Finally, interpretability pertains to how well the careless responding component(s) can be identified as such, which is important when the careless component is not explicitly defined as such through model constraints.

In the following sections, we begin by describing the data structure. Subsequently, we introduce LMFA and detail its three variants: fully exploratory, fully confirmatory, and partially confirmatory. Afterward, we present the simulation study to investigate how well the exploratory and partially confirmatory LMFA detect careless responding while accounting for violations of measurement invariances in attentive responses. Finally, we conclude with recommendations, discuss limitations, and propose avenues for future research.

Method

Data structure

We consider ILD with multiple indicators that are assumed to measure one or more underlying psychological constructs on a continuous scale. The observations are nested within subjects and denoted by y_{ijt} with $i \in \{1, \dots, N\}$ referring to subjects, $j \in \{1, \dots, J\}$ referring to items, and $t \in \{1, \dots, T_i\}$ to timepoints. These y_{ijt} are collected in the $J \times 1$ vectors $\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \dots, y_{iJt})'$, which themselves are collected in the $T_i \times J$ data matrix $\mathbf{Y}_i = (\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \dots, \mathbf{y}'_{iT_i})'$ for subject i. The data matrices are concatenated in the dataset $\mathbf{Y} = (\mathbf{Y}'_1, \dots, \mathbf{Y}'_N)'$ with $\sum_{i=1}^{N} T_i$ rows. As the notation indicates, the number of timepoints may differ across subjects. However, for simplicity, we omit the index i in T_i in the following.

Latent Markov factor analysis

The LMFA model can be conceptually divided into the measurement part and the transition part. The transition part is the same regardless of the LMFA variant. However, the measurement part is different for the fully constrained, completely exploratory, and partially constrained approaches. In the following, we first describe the measurement part and how it differs across the LMFA types. Subsequently, we explain the transition part.

Measurement part

The measurement part determines how many mixture components (also referred to as latent states instead of latent classes, as will be explained in the next section) underlie the data, where the mixture components

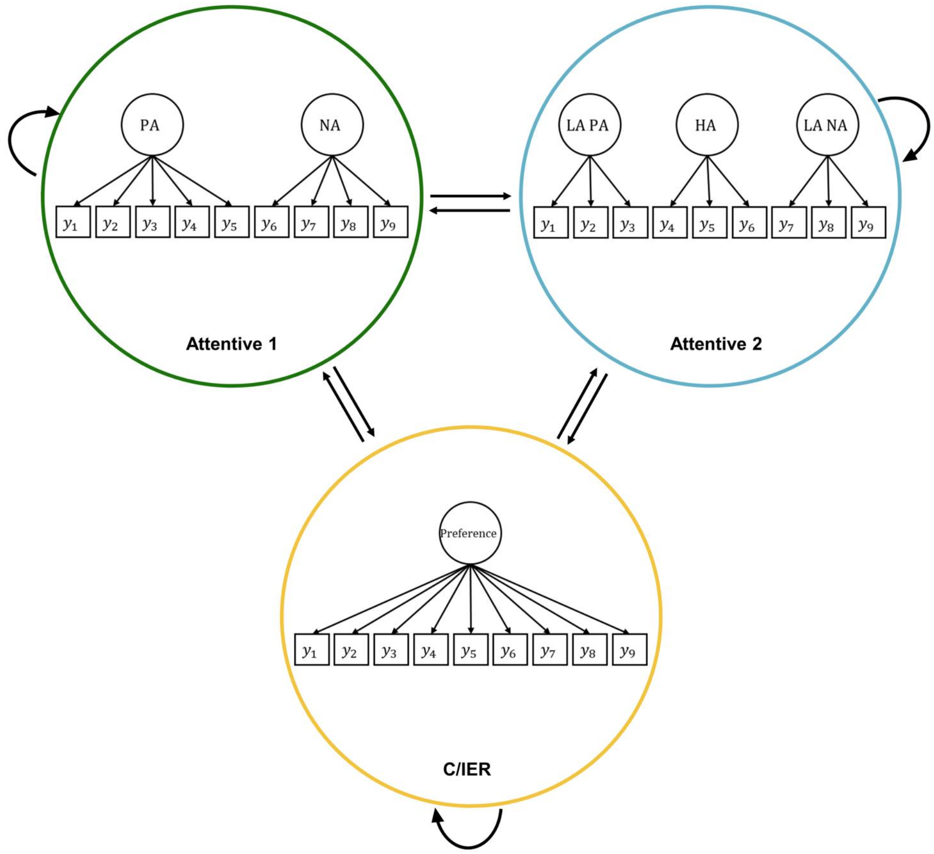

Figure 1. Graphical illustration of an LMFA model with three components: two attentive components, differing in the number and nature of the latent factors, and one C/IER state where items measure a preference factor rather than a content factor. The three components and thus measurement models are depicted inside the three big circles. The straight and curved arrows around these circles indicate the possibility of transitioning between the components at two subsequent measurement occasions and the possibility of staying in a component, respectively. In each circle, the observed items are indicated as squared boxes and the latent factors as circles. The presence of the arrows in between the items and factors indicates that items measure the latent factors. The first five items $y_1 - y_5$ are positive affect (PA) items (of which the first three are considered low-arousal (LA) emotions) and the last four items $y_6 - y_9$ are negative affect (NA) items (of which the first one is considered a high-arousal (HA) emotion). Individuals in the "Attentive 1" measurement model distinguish only between the valence of the emotions. Individuals in "Attentive 2" measurement model distinguish between the valence of the low-arousal emotions and between low and high-arousal emotions.

differ, broadly speaking, in which items measure which factors and how well. Figure 1 depicts an artificial model with three components: two attentive ones—where nine items measure two and three constructs of interest, respectively (positive affect [PA] and negative affect [NA] in component 1 and low arousal [LA] PA, LA NA, and high arousal [HA] in component 2)—and one careless one—which differs entirely from the attentive models because it measures scale preference rather than the constructs of interest. The factor models are component-specific and defined as (Lawley & Maxwell, 1962):

$$\left[\mathbf{y}_{it}|s_{itk} = 1\right] = \boldsymbol{\nu}_k + \boldsymbol{\Lambda}_k \mathbf{f}_{itk} + \mathbf{e}_{itk}, \tag{1}$$

with $k \in \{1, ..., K\}$ referring to the component. The component memberships are indicated via the binary indicators s_{itk}. These are equal to 1 for component k and equal to zero for the other components. Specifically, $s_{it1} = 1$ denotes that individual i belongs to component 1 at time point t. In turn, $[\mathbf{y}_{it}|s_{itk} = 1]$ implies that the responses $\mathbf{y}_{it}$ depend on the state-membership at time-point t. Moreover, $\boldsymbol{\nu}_k$ and $\boldsymbol{\Lambda}_k$ denote the state-specific $J \times 1$ intercept vector and the $J \times F_k$ loading matrix, respectively. The subject-specific $F_k \times 1$ vector $\mathbf{f}_{itk} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Psi}_k)$ stores individual i's factor scores at timepoint t (where F_k is the state-specific number of factors and $\boldsymbol{\Psi}_k$ the state-specific factor (co-)variances), and $\mathbf{e}_{itk} \sim \text{MVN}(\mathbf{0}, \mathbf{D}_k)$ is the subject-specific $J \times 1$ vector of residuals at timepoint t, where $\mathbf{D}_k$ contains the unique variances d_{kj} on the diagonal and zeros on the off-diagonal. The mixture components can thus differ regarding their loadings $\boldsymbol{\Lambda}_k$, intercepts $\boldsymbol{\nu}_\mathbf{k}$, unique variances $\mathbf{D}_k$, and factor covariances $\boldsymbol{\Psi}_k$. The three LMFA variants differ in how these parameters are constrained (including the number of components and factors) and, in turn, whether exploratory or confirmatory factor analysis is used within the mixture components. Below, we first introduce the fully exploratory and constrained LMFAs before explaining the partially constrained LMFA. A summary of the three versions is provided in Table 1.

Fully exploratory LMFA

The fully exploratory LMFA employs exploratory factor analysis within each mixture component. All parameters in Equation 1 are freely estimated. Only the factor variances in $\mathbf{\Psi}_k$ are restricted to one to set a scale of the latent factors. Rotational freedom is dealt with using criteria to optimize the simple structure of the factor loadings (e.g., oblimin, Clarkson & Jennrich, 1988). Model selection should determine the number of components and factors within these components that best fit the data. This entails that researchers define and estimate various plausible models and choose the best one based on model selection criteria and interpretability (details will be provided in the model selection paragraph in the Measurement Part Section). For example, considering the model in Figure 1, the exploratory analysis would indicate that there are three mixture components and thus measurement models that differ in the number of the factors. Component 1 has two factors with the first five items having considerable loadings on factor 1 and the last four having considerable loadings on the second factor. If the researcher knows that the items 1–5 are positive emotions and items 6–9 negative ones, they could give the labels PA and NA. Likewise, researchers would draw on their subject-matter expertise to interpret components 2 and 3. Note that whether one or more components refer to careless responding also has to be determined using post-hoc interpretation. The ease of this will be explored as part of the simulation study.

Fully constrained LMFA

For the fully constrained LMFA, the number of components and the number of factors per component are defined by the researcher in advance, based on theory, and the measurement models are obtained using confirmatory factor analysis. In the following, we consider the typical case with a single attentive and a single careless model. However, as explained in the introduction, it is possible to specify more attentive and/or C/IER components to account for known dynamics in item interpretation and/or for more nuances of careless responding. The attentive component requires the specification of a design matrix, indicating which items measure which constructs. There is no more rotational freedom in confirmatory factor analysis. However, the scale still has to be set, for example by fixing the factor variances in $\mathbf{\Psi}_k$ equal to one.

In the C/IER component, it is assumed that observed scores are not reflective of the to-be-measured substantive constructs for careless individuals. Instead, it is assumed that item scores are driven by mere scale preferences that are equal regardless of the item. This translates into a confirmatory factor model with a single preference factor where the loadings are all restricted to 1, even if some items are negatively worded as individuals are assumed to pay no attention to the content. Additionally, intercepts and unique variances are constrained. There are multiple suggestions on how to constrain them, ranging from more restrictive to more flexible ones. Restricting the intercepts and variances to equality across items, respectively, correspond most closely to the constraints in the conceptually similar IRT model proposed and evaluated by Uglanova et al. (2025) for cross-sectional data in general and by Vogelsmeier, Uglanova, et al. (2024) for ESM data within LMFA in particular. Although specifically tailored to scale preferences, these constraints were shown to be flexible enough to distinguish between attentive and careless responding when some individuals have scale preferences and others exhibit random responding; that is, when the C/IER component is misspecified. For other, (mostly) more restrictive constraints, see Kam and Cheung (2024).

Partially constrained LMFA

In the partially constrained LMFA, the researcher distinguishes in advance between attentive and C/IER components. The parameters of the attentive component(s) are still determined using exploratory factor analysis. This means that only the factor variances in $\mathbf{\Psi}_k$ are fixed to one and rotation criteria are applied, as in all components of the fully exploratory LMFA. For the C/IER component(s), however, confirmatory factor analysis is performed with constraints identical to those of the C/IER component(s) in the fully constrained LMFA. It should be noted that model selection is still required to choose the best model, as it is not known in advance how many attentive and C/IER components underlie the data. As already mentioned before, it is also interesting to see if model selection would choose more than one C/IER component for mixed C/IER types and wether using one or more

Table 1. Model specifications for the three LMFA versions.

Parameters/model(s)	Number of factors	Loadings	Intercepts	Unique variances	Factor mean(s)	Factor variance(s)	Factor covariance(s)
			Fully exploratory LMFA				
attentive	1 or more (based on model *selection*).	freely estimated	freely estimated	freely estimated	fixed to 0	fixed to 1	0, but can be changed with oblique rotation
C/IER	1	freely estimated	freely estimated	freely estimated	fixed to 0	fixed to 1	not applicable for one factor
			Partially constrained LMFA				
attentive	1 or more (*based* on model selection).	freely estimated	freely estimated	freely estimated	fixed to 0	fixed to 1	0, but can be changed with oblique rotation
C/IER	1	fixed to 1	fixed to equality	fixed to equality	fixed to 0	freely estimated	not applicable for one factor
			Fully constrained LMFA				
attentive	1 or more (based on theory).	based on design matrix, *some* are freely estimated, others are equal to zero.	freely estimated	freely estimated	fixed to 0	fixed to 1	can be freely estimated if desired (based on theory)
C/IER	1	fixed to 1	fixed to equality	fixed to equality	fixed to 0	freely estimated	not applicable for one factor

Note: The partially constrained LMFA is a combination of the fully exploratory and fully constrained LMFA: For the attentive model(s), the restrictions for the fully exploratory and the partially constrained versions have the same specifications. For the C/IER model(s), the partially and fully constrained versions have the same specifications. Whether there is more than one attentive and more than one C/IER component is determined based on model selection for both the fully exploratory and partially constrained models. For the fully constrained LMFA, typically, one attentive and one careless component (thought to absorb all types of C/IER) is considered, but more components could be specified based on theory.

components matters for how well the method distinguishes between attentive and careless responding.

Model selection

Two selection criteria were proposed for mixture factor analysis in general (Bulteel et al., 2013) and LMFA in particular (Vogelsmeier, Vermunt, & De Roover, 2023; Vogelsmeier, Vermunt, van Roekel, et al., 2019): the BIC (Schwarz, 1978), which balances model fit and parsimony by penalizing models with more parameters, and the convex hull (CHull) criterion (Ceulemans & Kiers, 2006; Wilderjans et al., 2013), which is an automated scree test in which models at the higher bound of the CHull in a "loglikelihood versus number of parameters" plot are identified. Subsequently, the best model is selected by pinpointing where the fit improvement levels off with increasing numbers of parameters. For more details, we refer to Vogelsmeier, Vermunt, and De Roover (2023). In this study, both criteria will be evaluated for the fully exploratory and the partially constrained LMFA. It is important to note that, for empirical data analyses, the final decision should always consider the interpretability of the model.

Transition part

The transition part determines the probabilities of transitioning between different mixture components. The possibility to either transition to another state or stay in the same state are depicted using the black arrows in Figure 1. Because of the possibility of switching components, the components are called "states" in LMFA (and in Markov modeling in general). However, they are just latent classes through which individuals transition over time. It is important to note that not all individuals have to go through all states. Some may be in the first attentive state throughout participation, others may transition only between the attentive states, and others between the second attentive state and the careless one. Individual- and time-point-specific covariates can be incorporated into the transition model, allowing to investigate correlates of transition patterns.

The transition model is obtained via a latent Markov model (LMM, e.g., Bartolucci et al., 2014). Specifically, it provides the probabilities of starting in a state (i.e., the initial probabilities) and the probabilities of transitioning to (or staying in) a state. The transition model for subject i is:

$$p(\mathbf{Y}_i, \mathbf{S}_i | \mathbf{Z}_i) = p(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}, \mathbf{s}_{i1}, \ldots, \mathbf{s}_{iT} | \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT})$$
$$= \overbrace{p(\mathbf{s}_{i1} | \mathbf{z}_{i1})}^{\text{initial state probabilities}} \prod_{t=2}^{T} \overbrace{p_{\kappa_{ti}}(\mathbf{s}_{it} | \mathbf{s}_{it-1}, \mathbf{z}_{it})}^{\text{transition probabilities}} \prod_{t=1}^{T} \overbrace{f(\mathbf{y}_{it} | \mathbf{s}_{it})}^{\text{response probabilities}} . \quad (2)$$

The $K \times 1$ vectors $\mathbf{s}_{it} = (s_{it1}, \ldots, s_{itK})'$ contain the binary indicators s_{itk} and the $U \times 1$ vectors $\mathbf{z}_{it} = (z_{it1}, \ldots, z_{itU})'$ comprise the covariate values z_{itu}, where $u = 1, \ldots, U$ denotes the subject- and timepoint-specific covariates. The latter can influence the initial or transition probabilities as described below. The state-specific response probabilities $f(\mathbf{y}_{it}|s_{itk} = 1)$ indicate the probabilities for the response patterns at timepoint t given the state membership at that timepoint, $s_{i1k} = 1$. These probabilities depend on the K state-specific models.

The initial state probabilities indicate the probabilities of starting in state k at timepoint $t = 1$ and can depend on covariate values at the first timepoint, $\mathbf{z}_{i1}$. The probabilities $\pi_k = p(s_{i1k} = 1|\mathbf{z}_{i1})$ with $\sum_{k=1}^{K} \pi_k = 1$ are collected in a $K \times 1$ vector $\boldsymbol{\pi}$. The initial state probabilities are typically modeled via a logit model to prevent parameter range restrictions:

$$\log\left(\frac{p(s_{i1k} = 1|\mathbf{z}_{i1})}{p(s_{i11} = 1|\mathbf{z}_{i1})}\right) = \beta_{0k} + \boldsymbol{\beta}_k'\mathbf{z}_{it=1}. \tag{3}$$

Here, β_{0k} are the initial state intercepts and the vectors $\boldsymbol{\beta}_k' = (\beta_{k,Z_{i11}}, \ldots, \beta_{k,Z_{i1U}})'$ are the initial state slopes that quantify the effect of the covariates on the initial state memberships for $k > 1$ because $k = 1$ is the reference category.

The transition probabilities indicate the probabilities of being in state k at timepoint $t > 1$ conditional on state $l \in \{1, \ldots, K\}$ at $t - 1$, and, thus, the probabilities of transitioning between states (or staying in the same one). There are two types of LMMs: The regular so-called discrete-time (DT-)LMM assumes the intervals between measurements, δ_{ti}, to be equal, while these intervals are allowed to differ across timepoints and individuals in the so-called continuous-time (CT-) LMM (Böckenholt, 2005; Jackson & Sharples, 2002; Vogelsmeier, Vermunt, Böing-Messing, et al., 2019). In this article, only the CT-LMM is used and described because differences in intervals are more realistic in ESM. Also note that the CT-LMM generalizes to the DT-LMM if intervals are equal (for a detailed description of the DT-LMM, we refer to Vogelsmeier, Vermunt, van Roekel, et al. (2019).

The transition probabilities in the CT-LMM, $p_{\delta_{ti},lk} = p_{\delta_{ti}}(s_{itk} = 1|s_{it-1,l} = 1, \mathbf{z}_{it})$, are collected in the $K \times K$ matrix $\mathbf{P}_{\delta_{ti}}$ with row sums $\sum_{k=1}^{K} p_{\delta_{ti},lk} = 1$. The transition probabilities $\mathbf{P}_{\delta_{ti}}$ depend on the interval δ_{ti} and the "transition intensity matrix" $\mathbf{Q}$. The transition intensities (or rates) q_{lk} define the transitions from the origin state l to the destination state k per a very small time unit and are collected in the $K \times K$ matrix Q. The intensities for the off-diagonal elements in the matrix $\mathbf{Q}$ (i.e., $k \neq l$) are

$$q_{lk} = \lim_{\delta \to 0}\left(\frac{p(s_{itk} = 1|s_{it-\delta,l} = 1, \mathbf{z}_{it})}{\delta}\right). \tag{4}$$

The diagonal elements are equal to $-\sum_{k \neq l} q_{lk}$ (Cox & Miller, 1965). Taking the matrix exponential of $\mathbf{Q} \times \delta_{ti}$ generates the transition probabilities $\mathbf{P}_{\delta_{ti}}$. This implies that the probability of transitioning to another state instead of staying in a state on two consecutive measurement occasions (i.e., $k \neq l$) increases for longer intervals. As can be seen from Equation (4), the transition intensities (and, hence, the transition probabilities) can depend on covariates $\mathbf{z}_{it}$. Typically, a log-linear model for the transition intensities is employed (again for $k \neq l$):

$$\log q_{lk} = \gamma_{0lk} + \boldsymbol{\gamma}_{lk}'\mathbf{z}_{it}. \tag{5}$$

Here, γ_{0lk} are the transition intercepts and $\boldsymbol{\gamma}_{lk}' = (\gamma_{lk,Z_{i11}}, \ldots, \gamma_{lk,Z_{i1U}})'$ the transition slopes that quantify the covariate effects on transitioning compared to staying.

It is important to note that each individual- and timepoint-specific observation is assigned to the states with probabilities that sum to 1 across all states. These probabilistic assignments reflect the certainty of an observation belonging to a particular state, such as a C/IER state. These posterior state-membership probabilities can be utilized in subsequent analyses to reduce the influence of careless responses on the results and thus conclusions. For instance, observations with low probabilities of belonging to an attentive state can be given less weight (see Ulitzsch, Domingue, et al., 2023; Ulitzsch, Shin, et al., 2024).

Observations can also be assigned to the state with the highest posterior state-membership probability (called modal assignment). This allows researchers to further explore (individual) dynamics in state memberships. Consider, for instance, the illustration in Figure 2 (again, using the example with two attentive and one C/IER state from before). The figure depicts how modal state assignments change over time (in terms of overall state proportions) and how individuals change between the three states. In this example, it can be seen that the probability of staying in any of the states is rather large. Additionally, looking at the C/IER state memberships, it is apparent that careless responding is less prominent at the beginning of participation but becomes more pronounced toward the end. The covariate participation length could be interesting to include here. Note that one may also look at individual transition plots for more nuanced insights into transitions of specific participants (for examples, see Vogelsmeier, Vermunt, & De Roover, 2023).

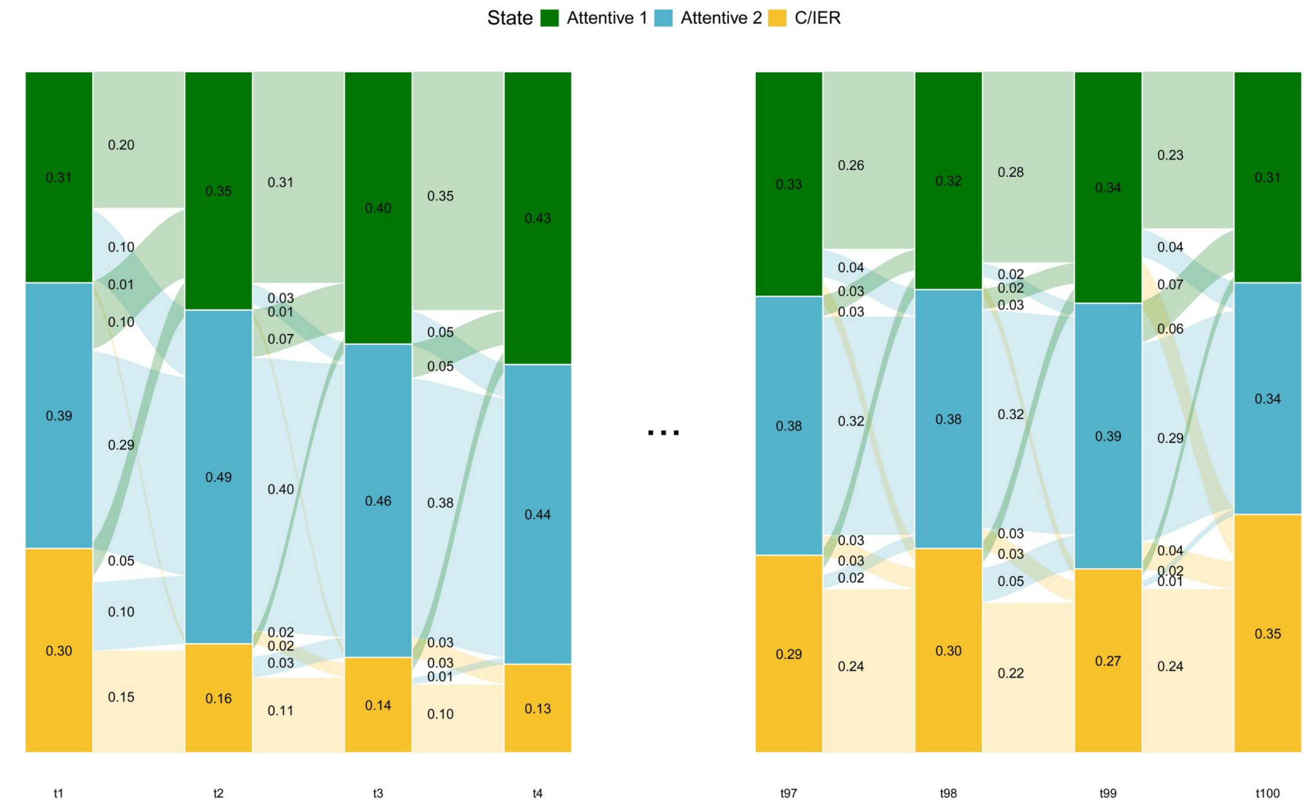

Figure 2. Example of transitions between states for 100 individuals and 100 timepoints. The left part depicts how state memberships change over the first four occasions and the right part depicts how state memberships change over the last four occasions. The membership distributions for the 8 visible time points are indicated by proportions. The slightly transparent colors between the time points show how the membership proportions shift from one timepoint to the next. For example, of 30% in the C/IER state at timepoint 1, half the individuals (i.e.15 of the total sample) stay in that state.

Note that the LMM can be easily extended by allowing for person-specific initial and transition probabilities (Vogelsmeier, Vermunt, van Roekel, et al., 2019), or by allowing for heterogeneity across groups of individuals. In the latter, the probabilities would depend on unobserved group memberships estimated using a mixture LMM (Crayen et al., 2017; Vogelsmeier, Vermunt, Bülow, et al., 2023). This extension would be interesting to detect between-person differences in transition patterns. Some individuals may be in an attentive state most of the time, others may often switch between the attentive and C/IER states, and others may start in an attentive state and, once they have moved toward a C/IER state, stay there until the end of their participation (this would make the C/IER state an "absorbing" state because individuals do not leave this state anymore). However, although such extensions are substantively interesting and possibly necessary for real data, they are not relevant for showing whether the fully exploratory and partially constrained LMFA versions can correctly disentangle attentive from C/IER responses because C/IER only affect the measurement part and not the transition part. For simplicity, in this study, we therefore focus on the regular LMM to capture transitions.

Estimation

Estimation can be performed using a one-step full information maximum likelihood estimation (Vogelsmeier, Vermunt, Böing-Messing, et al., 2019; Vogelsmeier, Vermunt, van Roekel, et al., 2019) or a step-wise approach that splits the estimation into the measurement and transition part. The step-wise approach is more efficient when including covariates in the transition model (for details and a comparison between the two approaches, see Vogelsmeier et al., 2023). Both estimations can be performed using Latent GOLD syntax (Vermunt, 2008). The three-step approach is also available in R (see package *lmfa* Vogelsmeier and De Roover, 2021), but estimation is considerably slower than in Latent GOLD. Both Latent GOLD and the R package use a multi-start procedure to reduce the chance of finding local optima (for details, see Vogelsmeier, Vermunt, van

Roekel, et al., 2019; Vogelsmeier, Vermunt, & De Roover, 2023). For the simulation study, we use the one-step estimation as covariates are not part of this study and we use Latent GOLD for computational efficiency. A technical description of the specific algorithm (an Expectation Maximization algorithm combined with a forward-backward algorithm), including information about the starting procedure and convergence criteria, can be found in the Appendix of Vogelsmeier, Vermunt, Böing-Messing, et al. (2019).

Simulation study

We conducted a simulation study to investigate how reliably the fully exploratory and the partially constrained LMFA capture careless responding in the presence of non-invariance of the attentive model. Throughout the simulation, we employed a data-generating model with two attentive states that differed in their loading pattern and one C/IER state (see Design and Procedure Section).

The simulation study had two aims in particular. The first aim was to evaluate how well LMFA, with and without constraints, can detect the true model among many candidate models using two previously proposed model selection criteria. More specifically, we investigated how often the BIC and the CHull method chose the true model without constraints among a set of candidate models without constraints (when using fully exploratory LMFA) and how often the two criteria chose the true model with constraints among a set of candidate models with constraints (when using partially constrained LMFA).[4] The second aim was to inspect the sensitivity (i.e., the proportion of correctly identifying C/IER observations) and specificity (i.e., the proportion of correctly identifying attentive observations), given the modal assignment from the correctly specified models.[5]

We integrated both aims into one simulation study, in which we manipulated three factors (in a full factorial design) that possibly influence the model selection as well as sensitivity and specificity: (1) noise in the attentive responses (with the two levels low and high), (2) type of C/IER (with the four levels: *middle scale preference*, *lower-end scale preference*, *random*, and *mixed*), and (3) frequency of C/IER (with the two levels low and high). To keep the computation time of the simulation study feasible, we omitted the manipulation of factors that have been studied in previous articles (e.g., sample size, different complexities of the attentive models, and differences across attentive states). Next to the two primary purposes of the simulation study, we also explored the interpretability of the C/IER states for the four different types of simulated C/IER.

Design and procedure

Data were generated according to the LMFA model (see Equation 2). Concerning the sampling protocol, the number of individuals was 45, and the number of measurement occasions 84 (mimicking 14 participation days with six occasions per day), which is a typical ESM setup according to a recent meta-analysis (Wrzus & Neubauer, 2023), resulting in a total of 3,780 observations.[6] To generate data with realistic intervals between measurement occasions, a participation day went from 9 am to 9 pm, with the night interval thus being 12 h and the day interval (i.e., the interval between two measurement occasions within a day) being 2.4 h.

Next, concerning the state-specific measurement models in the LMFA model, we generated data from two attentive states and one C/IER state. The number of items was equal to 20, and the scores on these items had a continuous range that fell between the score 1 and 10 (with small differences in the exact range, depending on different types of response patterns generated), representing continuous responses on a VAS. The two attentive states contained two latent factors, respectively. In both of these states, ten items had loadings on the first factor and ten on the second. Thus, both attentive states contained factor models with binary simple structures, which were also employed in previous LMFA simulation studies (Vogelsmeier, Vermunt, Böing-Messing, et al., 2019; Vogelsmeier, Vermunt, van Roekel, et al., 2019; Vogelsmeier et al., 2023)

[4]Note that we looked at the models with and without constraints separately, because investigating model selection on all models combined would only make sense if model selection works well for both fully exploratory and partially constrained LMFA separately, which was not the case, as will be described in the result section.

[5]Note that we do not examine parameter recovery of the attentive models, as prior studies have shown that it is very good if the observations are correctly classified (Vogelsmeier, Vermunt, Böing-Messing, et al., 2019; Vogelsmeier, Vermunt, van Roekel, et al., 2019; Vogelsmeier et al., 2023). Repeating this investigation would add length and distract from our novel focus on distinguishing attentive from careless observations.

[6]To retain feasibility of the simulation study design, we did not vary the sample size because it has been studied extensively for LMFA before (Vogelsmeier, Vermunt, van Roekel, et al., 2019), and it was shown that sample size affects classification performance only up to a certain degree, for example, for a model with three states, performance would not further improve when increasing sample size beyond 1200 observations, which is lower than typically observed in ESM data.

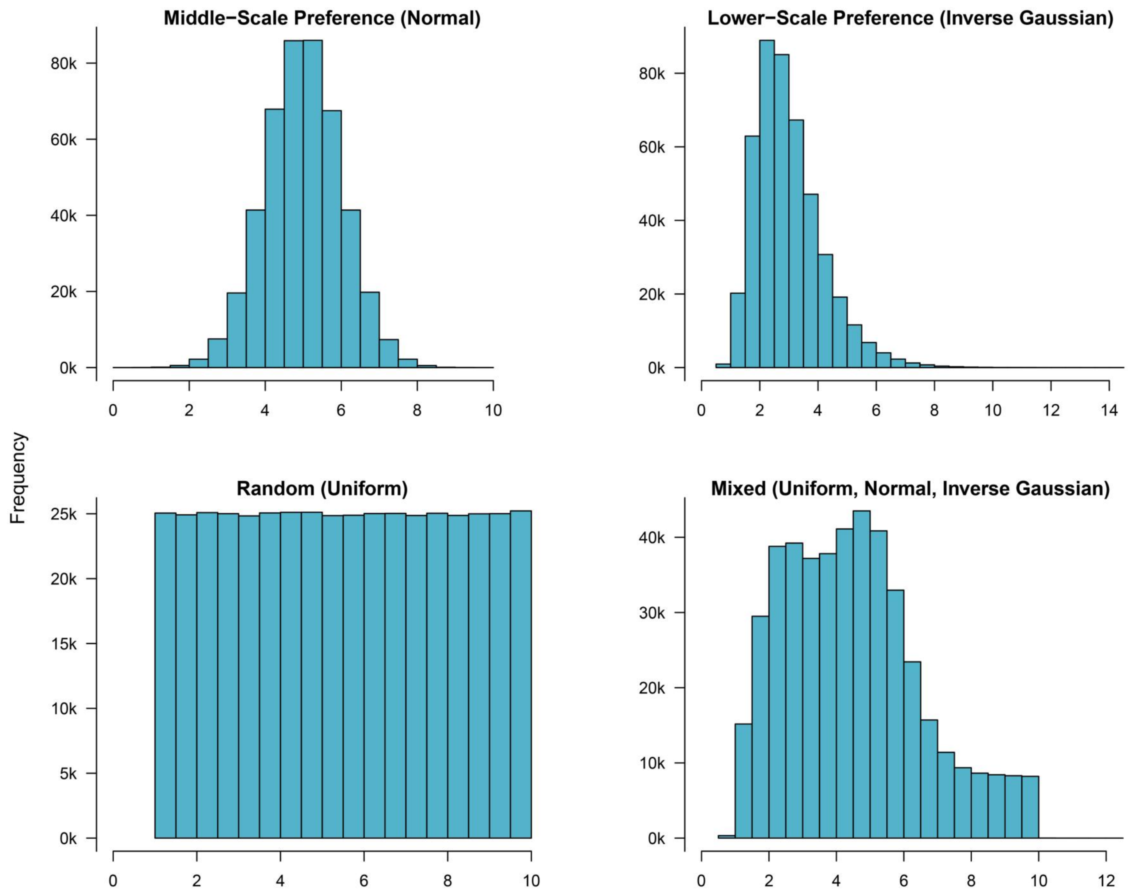

Figure 3. The four types of C/IER employed in the simulation study.

because they are very common in psychological research, for example, when studying positive and/or negative affect. The value of all nonzero loadings and all unique variances depended on the noise condition. For the low noise condition, the loadings were equal to 0.89, and the unique variances were equal to 0.2. For the high noise condition, these values were equal to 0.77 and 0.4, respectively. As a result, all item variances were equal to 1 in both conditions. We used the same conditions as employed in previous LMFA simulation studies (Vogelsmeier, Vermunt, Böing-Messing, et al., 2019; Vogelsmeier, Vermunt, van Roekel, et al., 2019; Vogelsmeier et al., 2023) because the difference in these studies was big enough to show for how much noise LMFA breaks down in correctly recovering state memberships. The two attentive states differed regarding four items: Two of the items with loadings on factor 1 in the first state loaded on factor 2 in the second state and vice versa. The resulting loading matrices for the two attentive states were thus:

$$\Lambda_1 = \begin{pmatrix} {\color{gray}\blacksquare} & {\color{gray}0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & {\color{gray}0} & {\color{gray}\blacksquare} & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare \\ {\color{gray}0} & {\color{gray}\blacksquare} & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & {\color{gray}\blacksquare} & {\color{gray}0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}'$$

$$\Lambda_2 = \begin{pmatrix} {\color{gray}0} & {\color{gray}\blacksquare} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & {\color{gray}\blacksquare} & {\color{gray}0} & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare \\ {\color{gray}\blacksquare} & {\color{gray}0} & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & \blacksquare & {\color{gray}0} & {\color{gray}\blacksquare} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}',$$

with “■” indicating loadings of 0.89 (for the low noise condition) and 0.77 (for the high noise condition) and the grey color highlighting the loadings that differ across the two attentive states. For both attentive states and all items, the intercepts were equal to

5, which is the middle value of the score range. Furthermore, the factor means for both attentive states were zero, and the factor variances per attentive state were 1.[7] The observations in the C/IER state were not drawn from factor models but from three distributions that mimic different C/IER behaviors on a continuous VAS. The C/IER types are illustrated in Figure 3. In the "random" condition, individuals randomly select a score between 1 and 10. The observations were drawn from a uniform distribution ranging from 1 to 10. For a "middle scale preference," individuals tend to select the middle score with some variation around it. The observations were drawn from a normal distribution with a mean of 5 and a standard deviation of 1. In the "lower-end scale preference," individuals tend to select a score on the lower end with some variation around it. The observations were drawn from an inverse Gaussian distribution with a mean of 3 and a shape parameter of 20. For the mixed condition, all three types of C/IER were equally often applied: The observations for respectively one-third of the 45 individuals were drawn from each of the three distributions described above.[8]

The transitions between the three latent states were manipulated via the LMM, which also determined the frequency of the two attentive and the C/IER states. In the "high frequency" conditions, the latent Markov chain was drawn from transition probabilities and initial state probabilities that lead to an equal amount of the three states. In the "low frequency" conditions, the latent Markov chain was drawn from probabilities that resulted in 10% C/IER states and, respectively, 45% of the two attentive states. Equal state sizes were chosen as a baseline, but unequal sizes are more realistic, as in the few studies that have investigated C/IER, only 5 to 10% of the observations were flagged as C/IER (e.g., Ulitzsch, Viechtbauer, et al., 2024; Vogelsmeier, Uglanova, et al., 2024)[9] Specifically, the probabilities for a 2.4-h interval (i.e., the interval between two measurement occasions within a day) were equal to

$$\mathbf{P}_{\text{highfrequency}} = \begin{pmatrix} 0.80 & 0.10 & 0.10 \\ 0.10 & 0.80 & 0.10 \\ 0.10 & 0.10 & 0.80 \end{pmatrix} \text{and}$$

$$\mathbf{P}_{\text{lowfrequency}} = \begin{pmatrix} 0.79 & 0.17 & 0.04 \\ 0.17 & 0.79 & 0.04 \\ 0.17 & 0.17 & 0.66 \end{pmatrix}.$$

The initial state probabilities were equal to $\boldsymbol{\pi}_{\text{highfrequency}} = \begin{pmatrix} 0.33 & 0.33 & 0.33 \end{pmatrix}$ and $\boldsymbol{\pi}_{\text{lowfrequency}} = \begin{pmatrix} 0.45 & 0.45 & 0.10 \end{pmatrix}$ for all individuals. To check how the manipulation played out, we examined the distribution of states across all generated datasets. The numbers precisely matched the intended proportions. Note that we disregard between-person differences for both initial and transition probabilities because making the transition model more complex unnecessarily complicates the model for the purpose of this simulation study (i.e., investigating different ways of estimating the measurement model), as explained before in the Transition Part Section.

For each condition, we sampled 100 datasets in the open-source program R (R Core Team, 2021) using all specifications and parameter values explained above. First, per person, we sampled the initial state membership using the initial-state probabilities that differed across the two frequency conditions (i.e., low vs. high). Subsequently, for each person, we sampled a random sequence of states using the transition probabilities that also depended on the frequency condition. Based on the resulting state memberships in the latent Markov chains, for the two attentive states, we drew observations from factor models with a fixed number of factors, fixed patterns of nonzero loadings, intercepts, factor means, and factor variances, but with the values for nonzero loadings and unique variances being different for the two noise conditions (i.e., low vs. high). For the C/IER state, we drew observations based on the four condition-specific C/IER distributions (i.e., uniform, normal, inverse Gaussian, or mixed distributions for random, middle-scale preference, lower-scale preference, and mixed types,

[7]To retain feasibility of the simulation study design, regarding the attentive models, only the noise in the data was varied because this likely affects how well some types of careless responding are distinguished from attentive responding. The number of states, factor overdetermination (manipulated by changing the number of factors for a fixed number of items), and between-state differences of the attentive models (manipulated by employing different degrees of non-invariance) were previously studied: Only the between-state difference affected classification performance, but even the most challenging one, where only loadings differ across states (which is also used for the current study), still showed good classification performance (Vogelsmeier, Vermunt, Böing-Messing, et al., 2019; Vogelsmeier, Vermunt, van Roekel, et al., 2019). Note that the number of items is kept constant for the following reason: When holding the number of factors and the number of items that differ across states constant, changing the number of items would simultaneously affect factor overdetermination and between-state differences, which is undesirable. Specifically, by increasing the number of items, factor determination would improve because more items provide a stronger measurement of the factor, while between-state differences would decrease as the proportion of items that differ between groups becomes smaller.

[8]Note that both LMFA versions cluster at the occasion level. Therefore, for the state composition, it is not relevant whether one-third of the individuals have mixed types or whether all individuals switch types after one-third of their participation. Both are possible ways to achieve the mixed C/IER type.

[9]The state proportions can be calculated from the transition matrices by raising the matrices to large powers (e.g., 50 or 100) until an equilibrium is reached, meaning that further matrix multiplication by itself no longer changes the probabilities.

respectively). In total, the simulation study encompassed 2 (noise) × 4 (C/IER types) × 2 (frequency) × 100 (replications) = 1600 datasets.

The analyses were conducted in Latent GOLD (Vermunt & Magidson, 2021). All datasets were analyzed using the two true models. A subset was also analyzed using both true and competing models to evaluate model selection, which is further explained below. The first true model (i.e., the fully exploratory LMFA model) had three states,[10] with two factors in the two attentive states and a single factor in the C/IER state, respectively. In the following, the model is referred to as "[221]," where the number of elements refers to the number of states and the value of each element to the number of factors for that state). Note that the states were not labeled beforehand; the only restriction in this model is the number of states and the number of factors per state. The second true model (i.e., the partially constrained LMFA model) had the same specification as the first one but with constraints in the C/IER state, as described above. In the following, this model is referred to as "[22C]" (where the "C" in state 3 indicates "constrained"). The sensitivity and specificity of the observations' classification were determined based on all 800 analyses.

Of the 1600 datasets, 240 (i.e., 15 of the 100 replications for each condition) were furthermore used to evaluate the performance of the model selection.[11] Specifically, we investigated whether in the case of single types of C/IER, the [221] and [22C] models were selected among various candidate models of differing complexities and whether in the case of mixed C/IER types, either model [221] or [22111] and [22C] or [22CCC] were selected among various candidate models. The latter is important because, strictly speaking, there are three distinct C/IER distributions underlying the observations in the mixed C/IER type condition rather than just one. Consequently, it is crucial to ascertain whether the model selection identifies one or three C/IER states for these conditions (together with the two attentive states resulting in a total of three or five states). The candidate models were chosen as follows: The true models comprised three states, each accommodating a maximum of two factors. Previous model selection studies suggest that both the BIC and CHull criteria tend to select the true model or favor simpler models in terms of factors and states (Bulteel et al., 2013; Vogelsmeier, Vermunt, van Roekel, et al., 2019). Therefore, models with one fewer state and factor than the true models were included in our investigation. Furthermore, the model complexity was extended up to five states, as the model selection has previously only been investigated for differences in attentive models and not between attentive models and different types of C/IER. Thus, we intended to ensure a sufficiently large range of models. Furthermore, five states may be necessary for the mixed C/IER conditions. For the fully exploratory LMFA without constraints, this resulted in 18 candidate models:[12] [21], [221], [2221], [22], [11], [222], [211], [111], [2222], [2211], [2111], [1111], [22222], [22221], [22211], [22111], [21111], [11111]. For the partially constrained LMFA, the candidate models also permitted one to three constrained states, which resulted in 28 candidate models: [2C], [22C], [222C], [1C], [21C], [11C], [221C], [211C], [111C], [2222C], [2221C], [2211C], [2111C], [1111C], [2CC], [1CC], [22CC], [21CC], [11CC], [222CC], [221CC], [211CC], [111CC], [2CCC], [1CCC], [22CCC], [21CCC], [11CCC]. Note that the imposed constraints were the same across the different C/IER states (i.e., equal to the constraints explained in the Method Section). They did thus not reflect the different types of C/IER in particular. Consequently, observations with different underlying C/IER types are distinguished only through differences in the intercepts and unique variances that were free to vary across the different C/IER states. This was decided for two reasons. Firstly, as previously explained, the constraints are tailored to scale preferences, but they were shown to also capture random responding (Kam & Cheung, 2024; Vogelsmeier, Uglanova, et al., 2024). Secondly, employing varying constraints for different C/IER states (i.e., different constraints for random responding and scale preferences) would introduce additional complexity to the already intricate model selection. For example, all models with a single C/IER state (i.e., models [2C], [22C], [222C], [1C], etc.) would have two versions, one with constraints for random responding and one for scale preferences. If constraints tailored to the C/IER type was needed in the partially constrained LMFA, it would raise the question of whether this LMFA variant is a practical one.

Results

In the following, we first report the results for the model selection and then for sensitivity and specificity.

[10]Note that this may not hold for the mixed C/IER type conditions, as will be explained in the next paragraph.

[11]Note that using more datasets per condition was computationally not feasible because the model selection part took half a day per dataset on a supercomputer with 32 cores.

[12]Note that the permutation of states is arbitrary. Thus, for example, model [21] is the same as model [12] and thus listed only once.

Table 2. Proportions of identifying the true models among their respective sets of candidate models using the BIC and the CHull method.

		BIC		CHull		BIC		CHull	
Frequency	Type	low noise	high noise	low noise	high noise	low noise	high noise	low noise	high noise
		[221]				[22C]			
low	middle-scale preference	1	1	1	0.80	0.93	1	0.87	1
	lower-scale preference	1	1	1	0.53	0.13	0.33	0.20	0.40
	random	1	0.93	1	0.8	0	0.07	0	0.07
	mixed type	0	0	0	0	0	0	0	0
high	middle-scale preference	1	1	0.87	0.87	1	1	0.80	0.80
	lower-scale preference	0	0	0	0	0	0	0	0
	random	0.80	0.47	0.80	0.13	0	0.07	0.07	0.07
	mixed type	0	0	0	0	0	0	0	0
		[22111]				[22CCC]			
low	mix	0.9	0.8	0	0	0.2	0	0	0
high	mix	0.7	0.6	0	0	0	0.07	0	0

Table 3. Sensitivity and specificity of classifying true C/IER state-membership observations into the C/IER states.

		Sensitivity		Specificity		Sensitivity		Specificity	
Frequency	Type	low noise	high noise	low noise	high noise	low noise	high noise	low noise	high noise
		[221]				[22C]			
low	middle-scale preference	0.99	0.90	1	0.99	0.99	0.90	1	0.99
	lower-scale preference	1	1	1	1	1	1	1	1
	random	1	0.97	1	1	0.03	0.04	0.79	0.85
	mixed type	0.96	0.65	1	1	0.67	0.55	1	1
high	middle-scale preference	0.99	0.95	1	0.98	0.99	0.95	1	0.98
	lower-scale preference	0.99	1	1	1	1	1	1	1
	random	0.70	0.63	0.96	0.95	0.02	0.01	0.79	0.84
	mixed type	0.46	0.39	1	1	0.66	0.57	1	1
		[22111]				[22CCC]			
low	mixed type	1	0.95	1	1	0.79	0.64	0.86	0.85
High	mixed type	0.99	0.94	0.99	0.99	0.68	0.65	0.80	0.85

Note: To determine sensitivity and specificity, attentive state memberships were combined in the three-state models and, in the five-state models, attentive and C/IER state memberships were combined, respectively.

Model selection

The upper part of Table 2 presents the proportions of times the models [221] and [22C] were selected from their respective sets of candidate models using both the BIC and CHull method. The lower part of Table 2 displays the proportions of times the models [22111] and [22CCC] were selected for the mixed C/IER type.

Fully exploratory LMFA

When noise and frequency of C/IER were realistically low, the BIC and the CHull selected the model [221] with a proportion of 1 for all C/IER types except the mixed type. For the mixed type, the model with five states was chosen with a proportion of 0.9 by the BIC but not at all by the CHull. The overall performance decreased a bit for high noise and a lot for the combination of high noise and high C/IER frequency, especially for the non-normally distributed response styles. The most challenging was the lower-scale preference style, where neither the BIC nor the CHull correctly identified the model [221] in any of the analyses.

Partially constrained LMFA

Regardless of the noise and frequency conditions, the BIC and the CHull method identified model [22C] only for a middle-scale preference with proportions of 0.8 or higher. For all other C/IER types, the model selection proportion was zero or at least lower than .4. Our exploration showed that mainly models with too many states were chosen. Additionally, neither the BIC nor the CHull adequately detected the model [22CCC] for the mixed C/IER type.

Sensitivity and specificity

The results for sensitivity and specificity are shown in Table 3. Based on the model selection results for the mixed-type conditions, we also show sensitivity and specificity using the five-state models [22111] and [22CCC].

Fully exploratory LMFA

When the noise level and frequency of C/IER were realistically low, sensitivity and specificity were nearly 1 for all types except the mixed type, which had a specificity of 0.96. Sensitivity remained only slightly lower under conditions of low frequency and high noise. For specificity, noise had no considerable effect.

At high frequencies of C/IER, specificity remained similarly high, even with high noise in the data. However, sensitivity was considerably lower for the random and mixed types, even under low noise conditions, with values of 0.7 and 0.46, respectively. This indicates that not all observations expected to be classified as C/IER were correctly identified. Under high noise, sensitivity values for these types dropped even further.

Examining the sensitivity and specificity for the mixed C/IER type using the [22111] model (and thus the model preferred by the model selection) reveals that both sensitivity and specificity are high for low noise and for both low and high C/IER frequencies, with values ranging from 0.99 to 1. For high noise, specificity remained stable, but sensitivity declined slightly to around 0.95.

Partially constrained LMFA

With a realistically low level of noise in the data and a low frequency of C/IER, specificity was nearly 1 for the middle- and lower-scale preferences, as well as for the mixed type. However, for random responding, specificity dropped to 0.79. Sensitivity was similarly high (nearly 1) for the middle- and lower-scale preference types. In contrast, sensitivity was only 0.03 for random responding and 0.67 for the mixed type, indicating that a considerable portion of C/IER observations remained unidentified.

When the noise level was high and the frequency of C/IER low, sensitivity and specificity showed minor changes without a clear trend. For high C/IER frequency, regardless of the noise level, sensitivity and specificity were only slightly affected, differing only by the second decimal place compared to the results for low C/IER frequency.

Examining the sensitivity and specificity for the mixed C/IER type using the [22CCC] model revealed consistently low performance across all noise levels and C/IER frequencies. The highest sensitivity and specificity values were 0.79 and 0.86, respectively, under conditions of low noise and low C/IER frequency.

Interpretability fully exploratory LMFA

Figure 4 shows the loadings in proportion to the unique variances as well as the intercepts for example data of each of the three C/IER type conditions. It can be seen that all parameter estimates are very comparable in size across the 20 items within one C/IER type. For all three types, the loadings fall around zero. For the middle and lower-scale preference types, unique variances are small compared to the random type. This occurs because individuals predominantly respond within a narrow range on the scale rather than utilizing its full range with scale preferences. The intercepts are approximately 5 for both middle-scale preference and random types. However, for the lower-scale preference type, the intercepts are lower due to most ratings being concentrated toward the lower end of the scale.

In attentive states, where items clearly measure the underlying construct(s), loadings would be considerably higher than zero and unique variances would be relatively small. The most obvious differences in the parameters of careless responses are, therefore, the small loadings across all three types and the high unique variances in the random type.

Conclusions and recommendations

For the fully exploratory LMFA, both model selection and classification of observations can be reliably performed for low and thus realistic amounts of noise and C/IER frequency. However, it should be noted that the CHull method is unreliable for mixed-type C/IER, indicating that the CHull method should complement rather than replace the BIC. For high C/IER amounts and high noise, model selection performance varies depending on the C/IER type, and if the correct model is chosen, for random C/IER, not all observations that should be flagged as C/IER are caught as such. However, this is not too concerning, since the high amount of noise and the high C/IER frequency employed in this study are unrealistic in practice.

For the partially constrained LMFA, neither model selection nor classification can be reliably performed, even when noise and C/IER frequency are realistically low. Thus, even though partially constrained models may seem appealing from a theoretical point of view, we advise avoiding them. It is important to note that this advice pertains specifically to the scenario explored in this paper, where accommodating potential non-invariance in attentive models involves permitting multiple attentive states that are obtained using exploratory factor analyses. In cases where a fully confirmatory model is employed, featuring one

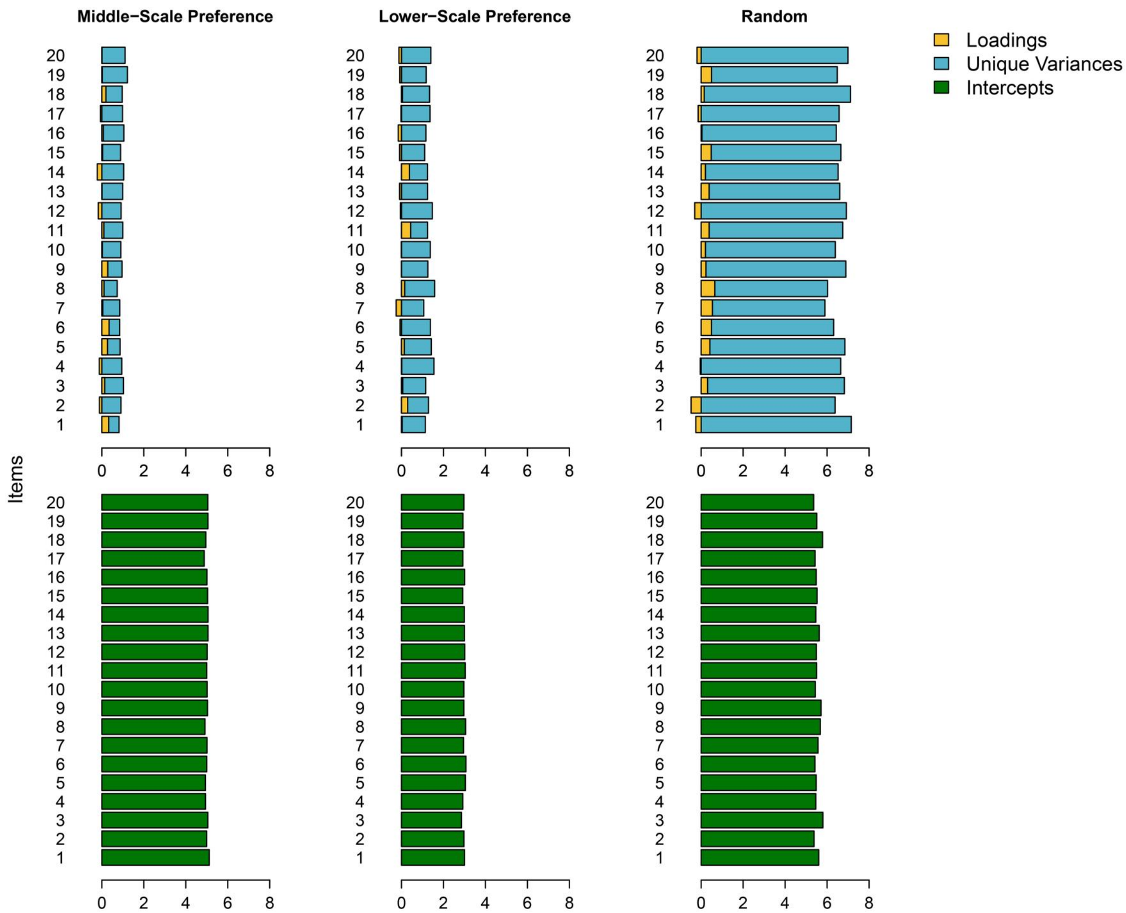

Figure 4. Distribution of loadings and unique variances (upper row) and intercepts (lower row) for three example datasets with generated middle-scale preference, lower-scale preference, and random responding, respectively. The results for the mixed C/IER type were omitted because the BIC would select one state per type and not just one in which all types are collapsed. Note that the bar charts were indistinguishable for the low and high frequency conditions and for the low and high noise conditions. Therefore, only the values for low frequency and low noise conditions are depicted.

attentive and one C/IER state, model selection is unnecessary, and studies demonstrated robust performance of the constraints in terms of sensitivity and specificity (Kam & Cheung, 2024; Vogelsmeier, Uglanova, et al., 2024).

Regarding interpretability, the C/IER states in the fully exploratory LMFA are clearly distinguishable due to their low loadings under the types of C/IER considered in the simulation study. Additionally, the random style is characterized by high unique variances. For real data, the patterns might be less obvious than in this simulation study. Therefore, in addition to visual inspection, we recommend including external variables, such as inattentiveness checks or response time, to further validate these states (for a detailed discussion on this, see Vogelsmeier, 2022).

Discussion

Intensive longitudinal data collection can be burdensome for the participants, potentially resulting in careless responding such as random responding. Besides changes in attentiveness, changes in how items are interpreted can occur over time, leading to violations of measurement invariance. Recently, a promising mixture model (fully confirmatory latent Markov factor analysis, Vogelsmeier, Uglanova, et al., 2024) was presented to identify careless responding in ILD. However, measurement invariance is a key assumption for this new method. If this assumption is violated, its effectiveness in identifying careless responding is compromised. In this study, we evaluated two variants of the mixture modeling approach LMFA for distinguishing between careless and attentive responding in the presence of non-invariance of the attentive responses.

The first variant is a fully exploratory LMFA model (Vogelsmeier, Vermunt, van Roekel, et al., 2019), which was originally developed to detect any changes in measurement models and evaluated for detecting changes in item interpretation. Our simulation study indicated that the model is also a reliable tool for detecting changes in different types of careless responding in addition to changes in attentive responding. Identifying whether components capture attentive or careless responding can be easily done using visual inspection of the detected mixture components.

The second variant differs from the fully exploratory LMFA model in that only the attentive mixture components are evaluated in an exploratory fashion, while the careless components are determined in advance using model constraints. These constraints are conceptually similar to those in the fully confirmatory LMFA model (Vogelsmeier, Uglanova, et al., 2024). Theoretically, the constraints in this partially confirmatory LMFA should aid in identifying careless components without the need for visual inspection. However, the model proved ineffective. It failed to correctly identify the complexities regarding the number of mixture components and factors, and it did not accurately classify observations even when the correct model was used. In the following, we outline a possible reason that may have contributed to the poor performance.

Interestingly, when using visual inspection to investigate the different types of careless responding after applying the fully exploratory LMFA, it was observed that the loadings for all types of careless responding employed in the study (random responding and scale preferences) had values around zero. In contrast, based on previous research (Arias et al., 2020; Kam & Cheung, 2024; Vogelsmeier, Uglanova, et al., 2024), the constraints for the same parameters were set to 1 in the partially confirmatory LMFA model. The use of 1s as constraints was motivated by the idea that scale preferences, if present, apply equally to all items and can be captured using a single latent preference factor. However, the discrepancy between the loadings in the fully exploratory LMFA and the constraints applied in the partially constrained LMFA raises the question of whether the constraints used are the most appropriate. One reason for this discrepancy is likely that the factor model constraints were tested for Likert-type data (which were assumed to be continuous but are strictly ordinal) rather than for continuous data as measured with a VAS, as employed in this study. With a Likert scale (e.g., 1–7), it is reasonable to assume that an individual rates all items with 5 if this is the category preference. With a VAS scale, it is practically very unlikely to always give the same rating (unless participants intentionally put effort into sliding the scale handle to the exact same position, which contradicts the definition of C/IER). It is more likely that individuals will score, for example, somewhere "around" 75: 89 on the first item, 69 on the second, and 90 on the third. This explains why the loadings are closer to zero rather than one in our continuous data compared to what might be observed with Likert-type data. Therefore, in future research, other constraints should be investigated for continuous data and clear guidelines on model constraints for both Likert type data and continuous data should formulated. Nevertheless, also note that the scale preference condition in our simulation study may not have been as skewed as one would expect with strong scale preferences. Choosing a slightly more extreme distribution might have led to better performance results, at least for this type of careless responding.

It is also plausible that employing varying constraints for different latent C/IER states tailored to different response styles could optimize performance in the partially constrained LMFA. Nevertheless, this introduces added complexity to the already intricate model selection and raises the question of whether this method would still be a viable alternative to the simpler and well-performing fully exploratory LMFA.

A related limitation of this study is that the C/IER types (i.e., scale preferences and random responding) were informed by prior studies for cross-sectional data (Meade & Craig, 2012; Roman et al., 2024; Schroeders et al., 2020; van Laar & Braeken, 2022). However, there remains uncertainty regarding the specific behaviors of participants who respond carelessly in ESM studies. To address this gap, further qualitative research or mixed-methods approaches are necessary to gain a deeper understanding of the nature of careless responses in ESM studies in the context of both Likert-type scales and continuous slider scales.

Another limitation is that none of the LMFA versions is applicable if the ESM data contain only single-indicator measures. For these data, researchers could use the approach by Ulitzsch, Nestler, et al. (2024), which leverages times spent on screens of electronically administered ESM studies and is based on theoretical assumptions about how much time respondents spend on items if they pay attention and how this evolves over time. For attentive screen times, the model assumes an exponential decay process that accounts for a potential decrease in screen times stemming from individuals who have become accustomed to the ESM assessment procedure and the items. In contrast, inattentive screen times are assumed to fluctuate randomly and typically exhibit shorter durations on average compared to the screen times of attentive respondents. The model permits individuals to switch between attentive and careless responding over time. The model can be enhanced by incorporating covariates at both the individual and occasion levels. Because the model solely relies on screen time information, it is well-suited for single-indicator scales. Note, however, that the model formulation is based on strict requirements for ESM data collection. For example, the approach cannot be used if the questionnaires in the morning and evening are longer than the other questionnaires of the day. Likewise, it is not yet clear how to handle commonly employed branching designs (Ulitzsch, Viechtbauer, et al., 2024). In addition, the approach relied on strong (yet not validated) assumptions and requirements concerning screen time distributions, violations of which are likely to cause estimation issues (Ulitzsch, Nestler, et al., 2024; Ulitzsch, Viechtbauer, et al., 2024).

Furthermore, all LMFA versions studied assume C/IER to be stable at a given measurement occasion, allowing for respondents to transition between different response behaviors only across occasions. Especially for lengthy questionnaires, this assumption may not hold, and one can easily imagine a respondent starting attentively when responding to the first couple of questions while being inattentive on those administered at the end of the questionnaire. While mixture IRT and factor models exist that accommodate such behavior (Roman et al., 2024; Ulitzsch, Yildirim-Erbasli, et al., 2022), their integration with any of the LMFA approaches is not straightforward and would result in a highly complex model comprising mixtures of mixtures.

In addition to the limitations discussed, an important contribution of this study is the demonstration of the flexibility of the original exploratory LMFA. This study illustrates that LMFA is well suited to detect changes in measurement models, even in complex scenarios with response behavior that manifests in nonnormal distributions inherent to some types of careless responding. Although assumptions about the (multivariate) normal distribution are made in LMFA, their violations have not been investigated. The results of this study suggest that such violations at least do not seriously affect the classification performance of LMFA, although this should be further investigated in future studies tailored to distribution violations of the attentive models.

As a final note, we would like to highlight that if researchers have strong assumptions about the type of measurement models between which individuals change, researchers can of course also specify two or more distinct confirmatory models in line with these assumptions instead of applying the fully exploratory LMFA. For instance, the item "being concerned" is known to measure negative as well as positive interpersonal emotions depending on the context (Vogelsmeier, Vermunt, Bülow, et al., 2023), which can be translated into two measurement models with the item having different item-construct relationships (i.e., loadings in a factor analysis or IRT model). Specifically, "being concerned" may load either on a positive interpersonal emotions factor or on a negative interpersonal emotions factor. However, it is often not known in advance what the attentive measurement model(s) look like and how many there are because the scales in ILD have not received the same degree of research attention as in cross-sectional research yet (Vogelsmeier, Jongerling, et al., 2024), and because the measurement of constructs over time is inherently more variable than the measurement used for collecting cross-sectional data. A fully confirmatory approach is, therefore, usually not feasible and was therefore not further considered in this paper.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Adolf, J., Schuurman, N. K., Borkenau, P., Borsboom, D., & Dolan, C. V. (2014). Measurement invariance within and between individuals: A distinct problem in testing the equivalence of intra-and inter-individual model structures. *Frontiers in Psychology*, *5*, 883. https://doi.org/10.3389/fpsyg.2014.00883

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*(6), 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014). Comments on: Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates. *Test*, *23*, 473–477. https://doi.org/10.1007/s11749-014-0387-1

Böckenholt, U. (2005). A latent Markov model for the analysis of longitudinal data collected in continuous time: States, durations, and transitions. *Psychological Methods*, *10*(1), 65–83. https://doi.org/10.1037/1082-989X.10.1.65

Bulteel, K., Wilderjans, T. F., Tuerlinckx, F., & Ceulemans, E. (2013). Chull as an alternative to AIC and BIC in the context of mixtures of factor analyzers. *Behavior Research Methods*, *45*(3), 782–791. https://doi.org/10.3758/s13428-012-0293-y

Ceulemans, E., & Kiers, H. A. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *The British Journal of Mathematical and Statistical Psychology*, *59*(Pt 1), 133–150. https://doi.org/10.1348/000711005X64817

Clarkson, D., & Jennrich, R. (1988). Quartic rotation criteria and algorithms. *Psychometrika*, *53*(2), 251–259. https://doi.org/10.1007/BF02294136

Cox, D. R., & Miller, H. D. (1965). *The theory of stochastic processes*. Chapman & Hall.

Crayen, C., Eid, M., Lischetzke, T., & Vermunt, J. K. (2017). A continuous-time mixture latent-state-trait Markov model for experience sampling data. *European Journal of Psychological Assessment*, *33*(4), 296–311. https://doi.org/10.1027/1015-5759/a000418

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Eisele, G., Vachon, H., Lafit, G., Tuyaerts, D., Houben, M., Kuppens, P., Myin-Germeys, I., & Viechtbauer, W. (2023). A mixed-method investigation into measurement reactivity to the experience sampling method: The role of sampling protocol and individual characteristics. *Psychological Assessment*, *35*(1), 68–81. https://doi.org/10.1037/pas0001177

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2023). Modeling careless responding in ambulatory assessment studies using multilevel latent class analysis: Factors influencing careless responding. *Psychological Methods*, *30*(2), 374–392. https://doi.org/10.1037/met0000580

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *The Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Jackson, C. H., & Sharples, L. D. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, *21*(1), 113–128. https://doi.org/10.1002/sim.886

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2022). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychological Methods*, *27*(6), 958–981. https://doi.org/10.1037/met0000312

Kam, C. C. S., & Cheung, S. F. (2024). A constrained factor mixture model for detecting careless responses that is simple to implement. *Organizational Research Methods*, *27*(3), 443–476. https://doi.org/10.1177/10944281231195298

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *The Statistician*, *12*(3), 209–229. https://doi.org/10.2307/2986915

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. https://doi.org/10.1037/a0019216

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 807–822. https://doi.org/10.1080/10705511.2021.1915788

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roman, Z. J., Schmidt, P., Miller, J. M., & Brandt, H. (2024). Identifying dynamic shifts to careless and insufficient effort behavior in questionnaire responses: A novel approach and experimental validation. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 775–793. https://doi.org/10.1080/10705511.2024.2304816

Schmitt, M. C., Vogelsmeier, L. V. D. E., Erbas, Y., Stuber, S., & Lischetzke, T. (2024). Exploring within-person variability in qualitative negative and positive emotional granularity by means of latent Markov factor analysis. *Multivariate Behavioral Research*, *59*(4), 781–800. https://doi.org/10.1080/00273171.2024.2328381

Schroeders, U., Schmidt, C., & Gnambs, T. (2020). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, *82*(1), 29–56. https://doi.org/10.1177/00131644211004708

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. In E. Diener (Ed.), *Assessing well-being. The collected works of Ed Diener*. Springer. https://doi.org/10.1007/978-90-481-2354-4

Uglanova, I., Nagy, G., & Ulitzsch, E. (2025). A mixture IRT model for handling different types of careless respondents. *OSF Preprints*. https://doi.org/10.31219/osf.io/tgys3_v1

Ulitzsch, E., Domingue, B. W., Kapoor, R., Kanopka, K., & Rios, J. (2023). A probabilistic filtering approach to non-effortful responding. *Educational Measurement: Issues and Practice*, *42*(3), 50–64. https://doi.org/10.1111/emip.12567

Ulitzsch, E., Nestler, S., Lüdtke, O., & Nagy, G. (2024). A screen-time-based mixture model for identifying and monitoring careless and insufficient effort responding in ecological momentary assessment data. *Psychological Methods*. https://doi.org/10.1037/met0000636

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2023). Using response times for joint modeling of careless responding and attentive response styles. *Journal of Educational and Behavioral Statistics*, *49*(2), 173–206. https://doi.org/10.3102/10769986231173607

Ulitzsch, E., Shin, H.-J., & Lüdtke, O. (2024). Accounting for careless and insufficient effort responding in large-scale survey data—Development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods*, *56*(2), 804–825. https://doi.org/10.3758/s13428-022-02053-6

Ulitzsch, E., Viechtbauer, W., Lüdtke, O., Myin-Germeys, I., Nagy, G., Nestler, S., & Eisele, G. (2024). Investigating the effect of experience sampling study design on careless and insufficient effort responding identified with a screen-time-based mixture model. https://doi.org/10.31234/osf.io/2s38a

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in survey data. *The British Journal of Mathematical and Statistical Psychology*, *75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501. https://doi.org/10.1111/jedm.12317

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*(1), 33–51. https://doi.org/10.1177/0962280207081238

Vermunt, J. K., & Magidson, J. (2021). *Upgrade manual for latent gold basic, advanced, syntax, and choice version 6.0*. Statistical Innovations Inc.

Vogelsmeier, L. V. D. E. (2022). *Latent Markov factor analysis: A mixture modeling approach for evaluating within- and between-person measurement model differences in intensive longitudinal data* [Doctoral Thesis]. Tilburg University. Ridderprint. https://pure.uvt.nl/ws/portalfiles/portal/59625665/Vogelsmeier_Latent_14_01_2021.pdf

Vogelsmeier, L. V. D. E., De Roover, K. (2021). *'lmfa': An r-package for exploring measurement invariance in intensive longitudinal data with continuous-time latent Markov factor analysis*. https://github.com/LeonieVm/lmfa

Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2024). Assessing and accounting for measurement in intensive longitudinal studies. *Quality of Life Research*, *33*(8), 2107–2118. https://doi.org/10.1007/s11136-024-03678-0

Vogelsmeier, L. V. D. E., Uglanova, I., Rein, M., & Ulitzsch, E. (2024). Investigating dynamics in attentive and inattentive responding together with their contextual correlates using a novel mixture IRT model for intensive longitudinal data. *The British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12373

Vogelsmeier, L. V. D. E., Vermunt, J. K., Böing-Messing, F., & De Roover, K. (2019). Continuous-time latent Markov factor analysis for exploring measurement model changes across time. *Methodology*, *15*(Supplement 1), 29–42. https://doi.org/10.1027/1614-2241/a000176

Vogelsmeier, L. V. D. E., Vermunt, J. K., Bülow, A., & De Roover, K. (2023). Evaluating covariate effects on ESM measurement model changes with latent Markov factor analysis: A three-step approach. *Multivariate Behavioral Research*, *58*(2), 262–291. https://doi.org/10.1080/00273171.2021.1967715

Vogelsmeier, L. V. D. E., Vermunt, J. K., & De Roover, K. (2023). How to explore within-person and between-person measurement model differences in intensive longitudinal data with the r package lmfa. *Behavior Research Methods*, *55*(5), 2387–2422. https://doi.org/10.3758/s13428-022-01898-1

Vogelsmeier, L. V. D. E., Cloos, L., Kuppens, P., & Ceulemans, E. (2024). Evaluating dynamics in affect structure with latent Markov factor analysis. *Emotion*, *24*(3), 782–794. https://doi.org/10.1037/emo0001307

Vogelsmeier, L. V. D. E., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent Markov factor analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 557–575. https://doi.org/10.1177/0163278720976762

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). Chull: A generic convex-hull-based model selection method. *Behavior Research Methods*, *45*(1), 1–15. https://doi.org/10.3758/s13428-012-0238-5

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538

Zhang, L., Ulitzsch, E., & Domingue, B. (2024). Bayesian factor mixture modeling with response time for detecting careless respondents. https://doi.org/10.31234/osf.io/qc9jb