Received: 10 February 2021 | Revised: 7 June 2021 | Accepted: 19 June 2021

DOI: 10.1002/bimj.202100046

Biometrical Journal

RESEARCH ARTICLE

A multilevel structural equation model for assessing a drug effect on a patient-reported outcome measure in on-demand medication data

Rob Kessels[1,2] | **Mirjam Moerbeek**[3] | **Jos Bloemers**[1,4] | **Peter G.M. van der Heijden**[3,5]

[1] Emotional Brain BV, Almere, The Netherlands

[2] Department of Biometrics, Netherlands Cancer Institute, Amsterdam, The Netherlands

[3] Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

[4] Utrecht Institute for Pharmaceutical Sciences and Rudolf Magnus Institute of Neuroscience, Utrecht University, Utrecht, The Netherlands

[5] Department of Social Statistics and Demography, University of Southampton, Southampton, United Kingdom

Correspondence
Rob Kessels, Emotional Brain BV, Louis Armstrongweg 88, 1311 RL Almere, The Netherlands.
Email: robkessels5@gmail.com

Funding information
Emotional Brain BV

Abstract

We analyze data from a clinical trial investigating the effect of an on-demand drug for women with low sexual desire. These data consist of a varying number of measurements/events across patients of when the drug was taken, including data on a patient-reported outcome consisting of five items measuring an unobserved construct (latent variable). Traditionally, these data are aggregated prior to analysis by composing one sum score per event and averaging this sum score over all observed events. In this paper, we explain the drawbacks of this aggregating approach. One drawback is that these averages have different standard errors because the variance of the underlying events differs between patients and because the number of events per patient differs. Another drawback is the implicit assumption that all items have equal weight in relation to the latent variable being measured. We propose a multilevel structural equation model, treating the events (level 1) as nested observations within patients (level 2), as alternative analysis method to overcome these drawbacks. The model we apply includes a factor model measuring a latent variable at the level of the event and at the level of the patient. Then, in the same model, the latent variables are regressed on covariates to assess the drug effect. We discuss the inferences obtained about the efficacy of the on-demand drug using our proposed model. We further illustrate how to test for measurement invariance across grouping covariates and levels using the same model.

KEYWORDS
latent variable modeling, measurement invariance, multilevel analysis, patient-reported outcomes, structural equation modeling

1 | INTRODUCTION

This paper discusses an application of a multilevel structural equation model (ML-SEM) to assess the effect of an active drug compared to placebo on a patient-reported outcome with measurements taken at multiple occasions/events. These data have two key characteristics. First, the patient-reported outcome is a set of questionnaire items intended to measure a construct as a latent variable. Second, the data follow a hierarchical structure with events nested within patients where the number of observed events is highly unbalanced across patients. Such a seemingly complex data structure may encourage researchers to apply aggregation prior to analysis to rely on relatively simple analysis techniques. In this paper, we explain why aggregating such data have several drawbacks and propose a model that overcomes these drawbacks. The model we will present combines a multilevel confirmatory factor analysis (ML-CFA) (Muthén, 1991, 1994) model with multilevel regression. This model can be used to compare two or more treatments on a latent variable with measurements taken at multiple (unbalanced) events, without the need for aggregating the data *a-priori*. However, the use of the proposed model also introduces important assumptions concerning measurement invariance that need to be addressed. We therefore also explain how to test these assumptions using a similar modeling procedure. Before developing the model, the data set used in this paper that motivated this research will be described.

1.1 | Motivating example

This paper was motivated by a study investigating the effect of an on-demand drug that is intended to improve sexual satisfaction in women. On-demand medication is only taken when necessary (e.g., to relief symptoms when they occur, such as pain) and trial data of on-demand medication intake are often observations of discrete events or episodes of when the medication was taken, including for example data on symptom severity for each discrete event.

The data analyzed in this paper come from a double-blind, placebo-controlled, randomized clinical trial investigating the efficacy of the on-demand use of the combined administration of testosterone and sildenafil (T+S), compared to placebo, in American women diagnosed with hypoactive sexual desire disorder (HSDD; which currently is part of the diagnosis female sexual interest/arousal disorder [FSIAD]) caused by low sensitivity of the brain for sexual cues (Trial registration: www.ClinicalTrials.gov: ID: NCT01432665) (Tuiten et al., 2018). During the trial, patients were instructed to take the medication prior to an anticipated sexual event and report about the event using the validated Sexual Event Diary (SED) (Van Nes et al., 2017). The SED is a web-based questionnaire that patients had to fill out within 24 h following a sexual event. For each sexual event, patients had to indicate the level of pleasure, inhibition ("ability to let yourself go"), sexual desire, bodily arousal, and subjective arousal on five-point Likert scale items (1 = *not at all*, 5 = *totally*). In this trial, all patients started with a 4-week baseline establishment (BLE) period where no medication was used by any patient. After the BLE period, patients were randomly assigned to an active drug (T+S) or placebo treatment and continued with an 8-week active treatment period (ATP) where patients used the drug they were randomly assigned to.

Traditionally, such data are aggregated prior to analysis in order to rely on conventional analysis techniques. For the current data, this means that each patient ends up with two scores: one score belonging to the BLE period and one score belonging to the ATP. To obtain these scores, two consecutive steps have to be carried out. First, for each sexual event, one composite outcome score is calculated by taking the sum of the Likert-scale items, resulting in one score per sexual event. Second, for each patient, this sum score is averaged over all events observed during the BLE period and over all events observed during the ATP. This results in a classical 2×2 within-between design with two repeated measurements per patient. Interest is then focused on the interaction effect between the between-factor treatment condition (T+S vs. placebo) and the within-factor study period (BLE period vs. ATP) and analyzed using a mixed between-within-subjects ANOVA. The presented ML-SEM approach in this paper will serve as a better alternative to analyze these data, which will be explained in the remainder of this Introduction section.

1.2 | Multilevel modeling and structural equation modeling

Aggregating the sexual event data of the motivating example leads to as loss of information because data values of many events are combined into fewer values. This can be overcome when the data are coded in a hierarchical data structure of events (level 1) collected in patients (level 2) and analyzed using multilevel analysis (Hox et al., 2018; Singer & Willett, 2003) (also known as mixed model analysis; Brown & Prescott, 2015). Furthermore, issues that arise when summing item

scores can be overcome by employing structural equation modeling (SEM), or latent variable modeling (Bollen & Curran, 2006; Rabe-Hesketh & Skrondal, 2008) approaches. SEM offers the possibility to relate observed items to corresponding factors (latent variable) and to assess the effect of covariates on these latent variables in one model. These alternative approaches will be discussed below.

Reducing the available repeated measurements to fit the classical 2×2 design by averaging the sum score over the events observed during the BLE period and over the events observed during the ATP for each patient has two important drawbacks. The first drawback concerns missing data. For patients without any reported events during either the BLE period or ATP, no aggregated score can be derived resulting in a missing outcome for that period. When analyzing the data using the between-within-subject analysis of variance (ANOVA), this missing data problem is generally handled by applying listwise deletion that involves deleting data of patients who report no events during the BLE period or ATP. This drawback can already be overcome by employing a multilevel analysis using the aggregated scores per patient as nested observations within patients.

The second drawback of reducing the data to fit a 2×2 design is concerned with the fact that each individual average score has a different standard error attached to it when the variance of the underlying event scores differs between patients and/or when the number of events per patient differs. In general, averages derived from more events will have a smaller standard error than averages derived from fewer events. This additional information is ignored when first averaging over all events and subsequently using two scores per patient in the analysis.

These two drawbacks can be overcome when a multilevel analysis using the individual events as nested observations is performed. This approach offers two additional possibilities. First, it enables the examination of potentially varying drug effects across patients by estimating a random drug effect. In clinical trials, the average drug effect is of main interest, but examining varying drug effects across individual patients can offer a valuable contribution to the evaluation of clinical trial results. As such, the ability to include random drug effects in the multilevel model makes the model more flexible in evaluating drug trials than, for example, the ANOVA model, where random drug effects cannot be estimated. To reliably estimate these random effects, it is crucial to include as many observations as possible for each patient, which provides another argument for preferring the multilevel model using the individual events as nested observations. In fact, estimating random drug effects in a multilevel model using the two aggregated scores (BLE period score and ATP score) is not possible because with only two observations per patient, this random drug effect becomes unidentifiable (Cnaan et al., 1997).

The second additional possibility of using the individual events rather than aggregated scores is that other relevant covariates available at the event-level can be included, such as other items of the SED.

In a previous study, it was shown that the application of a univariate multilevel approach considering the individual events as nested observations overcomes the abovementioned drawbacks when comparing this approach to the application of a between-within-subject ANOVA and a multilevel analysis using the two aggregated BLE period and ATP scores per patient as (balanced) outcomes, using the same sexual event data (Kessels et al., 2019). In that study, the sum score of the SED items pleasure, inhibition, sexual desire, bodily arousal, and subjective arousal derived for each sexual event was used as univariate outcome. In this paper, we will build on this work by Kessels et al. (2019) by including the five SED items as indicators of a factor (latent variable) in a structural equation model instead of using a sum score as outcome variable and we will apply this to the same data set.

The rationale behind creating a sum score as outcome variable in Kessels et al. (2019) was based on two studies that showed that these five SED items had excellent construct validity for measuring one factor, or latent variable (Van Nes et al., 2017, 2018). Sum scoring is a popular method for applied researchers and clinicians to approximate a latent variable, as it is straightforward and simple, it has a clear interpretation and it preserves the variability in the data (DiStefano et al., 2009). However, McNeish and Wolf (2020) argue why researchers should carefully consider the use of sum scores. When employing sum scoring, it is implicitly assumed that all items have equal weight in relation to the latent variable (factor) being measured. This causes serious issues when items are differentially related to the latent variable. If items relate differently to the latent variable being measured, it means that two patients having identical sum scores could have different true factor scores. This can lead to severe implications when the score is used as outcome in a clinical trial to assess treatment efficacy. This potential problem with sum scoring can be overcome when SEM is performed. A well-known SEM model is a confirmatory factor analysis model (Brown, 2014). The basic idea of utilizing factor analysis is that every item relates differently to the latent variable being measured and that every item has unique error variance (measurement error). The varying relationships between the items and the latent variable at hand are estimated as factor loadings and unique error variance can be modeled, which results in less biased latent variable estimates (Cole et al., 1993). SEM also offers the possibility to include regression-type structures assessing the effect of covariates on latent variables

(structural equation part) in the same model where the latent variable is measured by the observed items (measurement model). This latter possibility ensures that true factor scores rather than potentially biased sum scores are used as outcome and is the recommended approach in this paper.

1.3 | ML-SEM

The advantages of multilevel modeling and SEM compared to traditional ways for analyzing data raised interest for combining the two approaches, referred to as ML-SEM (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005; Rovine & Molenaar, 2000). In this paper, an ML-SEM model will be used for analyzing the sexual event data described above. This ML-SEM consists of two types of models: the ML-CFA model (Muthén, 1994, 1991) and the multilevel multiple indicators multiple causes (ML-MIMIC) model (Muthén, 1989). Both of these models are described below.

First, an ML-CFA (Muthén, 1994, 1991) model will be fitted to the data. This involves analyzing the covariance structure of the observed items within patients at the event level and analyzing the covariance structure of the item means between patients at the patient level by fitting a factor model on both levels in one model. As described before, in previous studies, the five SED items (pleasure, inhibition, sexual desire, bodily arousal, and subjective arousal) were factor analyzed on the event level and patient level, but this was done separately over two different analyses using the individual item scores and aggregated item mean scores, respectively (Van Nes et al., 2017, 2018). However, when only using the aggregated item mean scores, within-patient variability is ignored and when only using the individual item scores, between-patient variability is ignored. Ignoring variability within and/or between patients in a nested data structure can result in biased estimates of the factor model (Kaplan et al., 2012). In an ML-CFA model, this problem does not occur as within- and between-patient variabilities are both taken into account by specifying a factor structure at both levels simultaneously.

By analyzing the covariance between variables at both levels using multilevel factor analyses, it allows one to explain why different factor structures at the within-patient level are not necessarily replicable at the between-patient level. Also, in situations where the factor structure is equal across levels, it is possible to establish whether greater variability exists between patients or within patients. These sort of research questions have been investigated in psychology research using the ML-CFA model to study *state* and *trait* variability of latent constructs like coping, and negative affect and positive affect (Merz & Roesch, 2011; Roesch et al., 2010). A trait is considered to be a person's characteristic that remains (more or less) stable over time, but can differ between persons. A state is the adaption of a person to a particular moment in time or situation that can differ within persons and as such represents a person's deviation from his/her stable trait that fluctuates over time. Applied to the sexual event data analyses in this paper, an ML-CFA model can give additional insights in how a patient's sexual behavior fluctuates from event to event (state-like variability) in comparison to how sexual behavior differs between patients (trait-like variability).

Second, the ML-CFA model will be extended as an ML-MIMIC model (Muthén, 1989). In a MIMIC model, the latent variable is measured by the observed variables (factor model), which is the measurement model part, and is regressed on observed covariates, which is the structural model part (Jöreskog & Goldberger, 1975). In a multilevel setting, the relationships between observed variables and latent variables occur at multiple levels in the hierarchy. The ML-MIMIC model can therefore be used to study the cross-level interaction effect between treatment condition and study period on a latent variable, measured by the SED items. Part of the ML-MIMIC model is the ML-CFA model and the other part of the ML-MIMIC model is a regular multilevel regression model with covariates predicting latent variables, allowing the inclusion of fixed and random effects.

The application of SEM is very common in, for example, the behavioral, educational, and social sciences. By contrast, SEM has played a much smaller role in analyzing clinical trials. SEM has been described as an attractive alternative method in analyzing multivariate, longitudinal clinical trials where hypotheses concerning treatment effects on constructs that reflect multiple (patient reported) endpoints could be assessed (Song et al., 2008; Donaldson, 2003). Furthermore, multilevel latent variable models have been proposed as analysis technique to assess global quality of life where SEM is presented as alternative to aggregating quality of life items (Kifley et al., 2012). The ML-SEM modeling approach proposed in this paper has rarely been recognized as main option in analyzing patient reported outcomes in clinical trials. A reason for this could be the relative complexity of this model as opposed to the much easier sum score. The goal of this paper is to show how the patient reported outcome in the on-demand medication data of the motivated example can be analyzed using the ML-SEM procedure described above and to convince researchers why this model is to be preferred. An important assumption in the applicability of this approach is measurement invariance, a topic that received very little attention in the literature describing SEM-type solutions for clinical research. Measurement invariance is a statistical phenomenon

that states that the same latent variable is measured across levels of the multilevel hierarchy (event level and patient level) and across specified groups (e.g., treatment condition) (Vandenberg & Lance, 2000). Measurement invariance of the ML-SEM model presented in this paper is required to use the latent variable as outcome variable and to allow comparison of latent means. Therefore, in this paper, in addition to the application of the ML-CFA and ML-MIMIC model to analyze the on-demand medication data, we investigate measurement invariance across levels and across grouping variables at the within-patient level (study period distinguishing between two groups of events within patients) and between-patient level (treatment condition distinguishing between the two treatments between patients). Finally, we compare the results of the ML-SEM model presented in this paper and the univariate multilevel model using the sum score of the five SED items as outcome as presented in a previous paper (Kessels et al., 2019).

2 | METHODS

2.1 | Data preparations

The data analyzed in this paper are the sexual event data coming from a placebo-controlled, randomized clinical trial. The five SED items (pleasure, inhibition, sexual desire, bodily arousal, and subjective arousal) were used as indicators in the ML-CFA model. The design elements study period ($0 = BLE$, $1 = ATP$) and treatment condition ($0 = placebo$, $1 = T{+}S$) were included as covariates. Furthermore, patients had to indicate if they used study medication prior to the sexual event ($1 = yes$, $0 = no$). Note that the SED items pleasure, inhibition, sexual desire, bodily arousal and subjective arousal, and the covariate study period are variables varying over events and are therefore located at the event level within patients. On the other hand, because each patient receives either placebo or T+S, the variable treatment condition varies only between patients and is therefore located at the patient-level. For analysis, events reported during the ATP were only included if medication was taken prior to the event.

2.2 | Notation and ML-CFA model

Considering the sexual event data, let the number of patients be indexed by $i = 1, \ldots, N$, where N is the sample size and let the number of observed sexual events within patients be indexed by $e = 1, \ldots, E_i$ with E_i being the total number of observed events during the BLE period and ATP for patient i. In two-level data, the general ML-CFA model can then be expressed by the following set of equations:

$$\boldsymbol{Y}_{ei} = \boldsymbol{\mu}_i + \boldsymbol{\Lambda}_w \boldsymbol{\eta}_{ei(w)} + \boldsymbol{\epsilon}_{ei(w),} \tag{1}$$

$$\boldsymbol{\mu}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}_b \boldsymbol{\eta}_{i(b)} + \boldsymbol{\epsilon}_{i(b)}, \tag{2}$$

where the first equation models the within-patient variation (i.e., state variation) and the second equation models the between-patient variation (i.e., trait variation). Substituting Equation (2) into Equation (1) and rearranging terms yields the following single-model equation:

$$\boldsymbol{Y}_{ei} = \boldsymbol{\mu} + \boldsymbol{\Lambda}_w \boldsymbol{\eta}_{ei(w)} + \boldsymbol{\Lambda}_b \boldsymbol{\eta}_{i(b)} + \boldsymbol{\epsilon}_{ei(w)} + \boldsymbol{\epsilon}_{i(b)}, \tag{3}$$

where $\boldsymbol{Y}_{ei}$ is a vector of observed item scores of patient i on event e, defined on five-point Likert scales. These observed scores are predicted by a regression equation involving a vector of intercepts $\boldsymbol{\mu}$, the within-patient factor loading matrix $\boldsymbol{\Lambda}_w$ multiplied by the within-patient vector of latent variables (factors) $\boldsymbol{\eta}_{ei(w)}$ for event e, the between-patient factor loading matrix $\boldsymbol{\Lambda}_b$ multiplied by the between-patient vector of latent variables (factors) $\boldsymbol{\eta}_{i(b)}$ for patient i plus a within-patient vector of residual error terms $\boldsymbol{\epsilon}_{ei(w)}$ and a between-patient vector of residual error terms $\boldsymbol{\epsilon}_{i(b)}$. The factor loading matrix reflects the pattern and magnitude of the relationship between the observed items and unobserved latent variables. The presented model above (Equations (1)–(3)) is a random-intercept model where the patient item intercepts $\boldsymbol{\mu}_i$ in Equation (1) are random intercepts that vary across patients.

The factors are multivariate normally distributed with means of zero and covariance matrices $\boldsymbol{\Psi}_w$ and $\boldsymbol{\Psi}_b$ at the within-patient level and between-patient level, respectively. Furthermore, the residual error terms ($\boldsymbol{\epsilon}_{ei(w)}, \boldsymbol{\epsilon}_{i(b)}$) are also

multivariate normally distributed with means of zero and covariance matrices $\boldsymbol{\Theta}_w$ and $\boldsymbol{\Theta}_b$ for the within-patient residual terms and between-patient residual terms, respectively. Combining the within and between variation results in the following total covariance matrix:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_b \boldsymbol{\Psi}_b \boldsymbol{\Lambda}_b{}' + \boldsymbol{\Theta}_b + \boldsymbol{\Lambda}_w \boldsymbol{\Psi}_w \boldsymbol{\Lambda}_w{}' + \boldsymbol{\Theta}_w, \tag{4}$$

which means that in an ML-CFA model, the total variance–covariance matrix is expressed as a function of the factor loadings, the factor covariances, and the covariances of the residual measurement errors.

2.3 | Measurement invariance

In the ML-CFA model presented in the previous subsection, the observed dependent variables are indicators reflecting underlying latent variables at the within-patient and between-patient level. These latent variables can be used as new dependent variables in multilevel regression analyses by including covariates predicting these latent variables at both levels (MIMIC approach). When a covariate represents a grouping variable (e.g., treatment condition or study period), including this covariate allows comparing latent variable means across groups. Invariance of measurement parameters (identical factor loadings and intercepts) across groups is a prerequisite to compare latent variable means across groups (Meredith, 1993). Invariance of factor loadings and intercepts ensures that any differences between groups on the observed indicators are only attributable to differences in latent variables. This simplifies the interpretation of any latent mean differences. However, when extending the ML-CFA model in Equation (3) as ML-MIMIC model by including covariates, the model estimates only one model for the full (combined) study sample of patients and events (Muthén, 1989). This means that a MIMIC model implicitly assumes measurement invariance across study periods and across treatment conditions. This implicit assumption may be too strict and requires further examination. In this subsection, we will discuss how to test these measurement invariance assumptions. In addition to testing measurement invariance across study periods and treatment conditions, we also discuss how to impose measurement invariance constraints on the ML-CFA model across the within-patient level and between-patient level. Overall, we distinguish between configural invariance, uniform and nonuniform factorial invariance across grouping variables at the within-patient and between-patient level, and invariance of across-level factor loadings.

The first step is to ensure that the same factor structure holds across levels. This is the least restrictive constraint that can be imposed on an ML-CFA model and is generally known as configural invariance (Stapleton et al., 2016). If there exists a different factor structure at both levels, the model becomes hard to interpret and the factors cannot serve as dependent variables in a multilevel regression (Hox et al., 2018). As discussed in the Introduction section, previous factor analyses on the item scores of the SED revealed a one-factor solution on the event level and patient level (Van Nes et al., 2017, 2018). These factors were interpreted as factors measuring sexual functioning of a single event and sexual functioning of a patient. Based on these results, we expect that in our analysis, a one-factor solution at both levels fits the data. This one-factor ML-CFA model is presented in Figure 1. If this model indeed fits the data, it shows that the model is configural invariant across levels (Stapleton et al., 2016).

When including the grouping covariates study period and treatment condition as predictors of the latent variable, the ML-CFA model presented in Figure 1 becomes an ML-MIMIC model. As discussed before, in such an ML-MIMIC model, measurement invariance across study periods (BLE period and ATP) at the within-patient level and across treatment conditions (T+S and placebo) at the between-patient level is implicitly assumed. Here, we present a method to verify this strict assumption by testing invariance of intercepts and factor loadings across groups. If this assumption can be accepted, it indicates that patients with identical sexual function scores but from different groups have the same probability of getting a particular score on the sexual function scale. This means any latent mean differences can be properly interpreted. When this measurement invariance assumption is (partly) violated, ignorance of this violation can lead to biased path coefficients of the grouping covariate effect (Guenole & Brown, 2014). In such cases, caution in interpreting latent mean differences is warranted.

Usually, for testing measurement invariance across groups, multigroup factor analysis is applied (Sörbom, 1974). This method involves specifying separate models for each group and then imposing equality constraints between these group models to test the equality of factor loadings and intercepts over groups. This approach cannot be used to test equality of intercepts and factor loadings of grouping variables located at the within-level, because a within-level grouping variable is crossed with units at the between-level. In the example used in this paper, this simply means that

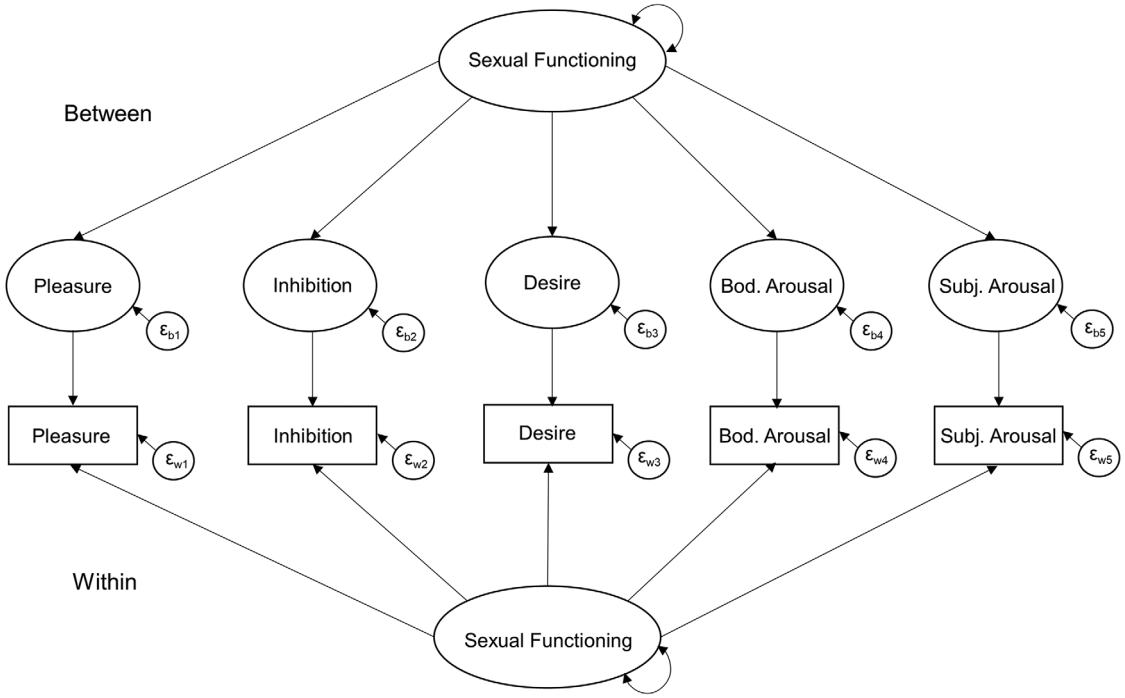

FIGURE 1 The ML-CFA model. Bod = Bodily, Subj = Subjective

the two study periods (BLE period and ATP) are observed in every patient. In other words, study period is crossed with patients.

An alternative method to test measurement invariance in SEM is using the MIMIC modeling approach described before, where a covariate is assumed to have a causal effect on a latent variable. Kim et al. (2015) showed that the MIMIC procedure can be used to test measurement invariance of within-level groups in an ML-CFA model. This can be done as follows. Let $\boldsymbol{X}_{ei}$ be a vector of observed within-patient covariates for patient i on event e. The MIMIC model equations at the within-patient level can then be written as

$$\boldsymbol{Y}_{ei} = \boldsymbol{\mu}_i + \boldsymbol{\Lambda}_w \boldsymbol{\eta}_{ei(w)} + \boldsymbol{\Gamma}_y \boldsymbol{X}_{ei} + \boldsymbol{\epsilon}_{ei(w)}, \tag{5}$$

$$\boldsymbol{\eta}_{ei(w)} = \boldsymbol{\Gamma}_{\eta_{ei(w)}} \boldsymbol{X}_{ei} + \boldsymbol{\zeta}_{ei(w)}, \tag{6}$$

where $\boldsymbol{\Gamma}_{\eta_{ei(w)}}$ represents a matrix of regression coefficients estimating the effects of the covariates $\boldsymbol{X}_{ei}$ on the within-patient latent variable vector $\boldsymbol{\eta}_{ei(w)}$ and where $\boldsymbol{\Gamma}_y$ is a matrix of regression coefficients estimating the effects of the covariates $\boldsymbol{X}_{ei}$ on the observed outcomes $\boldsymbol{Y}_{ei}$. A regression coefficient associated with a grouping covariate (e.g., study period) in Equation (6) represents the effect of group membership on the latent variable. The inclusion of a grouping covariate on the observed outcomes in Equation (5) over and above the grouping effect on the latent variable vector allows testing measurement invariance of the intercepts across groups, which is referred to as uniform factorial invariance testing by Kim et al. (2015).

Equation (5) can be expanded to test for measurement invariance of factor loadings between study periods as well by including an interaction effect between study period and the latent variables (Barendse et al., 2010, 2012). This yields

$$\boldsymbol{Y}_{ei} = \boldsymbol{\Lambda}_w \boldsymbol{\eta}_{ei(w)} + \boldsymbol{\Gamma}_y \boldsymbol{X}_{ei} + \boldsymbol{\Gamma}_{\eta_{ei(w)}y} \boldsymbol{\eta}_{ei(w)} \boldsymbol{X}_{ei} + \boldsymbol{\epsilon}_{ei(w)}, \tag{7}$$

where the regression coefficients present in $\boldsymbol{\Gamma}_{\eta_{ei(w)}y}$ allow the testing of factor loading invariance, or nonuniform factorial invariance (Kim et al., 2015). When study period is included as a covariate in Equation (7), the associated regression coefficient in $\boldsymbol{\Gamma}_y$ then represents intercept invariance or uniform factorial invariance and the associated regression coefficient in $\boldsymbol{\Gamma}_{\eta_{ei(w)}y}$ then represents factor loading invariance or nonuniform factorial invariance. Presence of uniform and

nonuniform factorial invariance (regression coefficients are equal to zero) indicates that any differences between the BLE period and the ATP on the observed items are only attributable to differences in the latent variables. As study period represents a covariate varying over the events and therefore also varying over time, a factorial invariance test for study period can also be regarded as a method to assess measurement invariance across events to inspect whether the model is consistent over time. Presence of factorial invariance is a necessary condition for substantive interpretation of differences between the observed items. If there is no uniform and/or nonuniform factorial invariance, any differences in observed items between a grouping covariate are not only attributable to differences in the latent variables, but may also be caused by other (unmeasured) confounders and complicates the interpretation (Jak et al., 2013).

It should be noted that the regression coefficients in Equation (7) are estimated as fixed effects, assuming that the status of invariant intercepts and/or factor loadings does not vary across patients (no random effects), which is, as pointed out by Kim et al. (2015), a "realistic assumption with a reasonably developed measure."

The MIMIC approach for testing factorial invariance between grouping covariates at the within-patient level can also be adopted for testing factorial invariance across grouping covariates located at the between-patient level (e.g., treatment condition). In fact, testing factorial invariance for grouping covariates located at both levels of the hierarchy can be conducted simultaneously in one model, creating an ML-MIMIC model.

Let $\boldsymbol{Z}_i$ be a vector of (grouping) covariates for patient i. Then, testing uniform and nonuniform factorial invariance for $\boldsymbol{Z}_i$ and $\boldsymbol{X}_{ei}$ expands the complete ML-CFA model in Equation (3) as follows:

$$\boldsymbol{Y}_{ei} = \boldsymbol{\mu} + \boldsymbol{\Lambda}_w\boldsymbol{\eta}_{ei(w)} + \boldsymbol{\Gamma}_y\boldsymbol{X}_{ei} + \boldsymbol{\Gamma}_{\eta_{ei(w)}y}\boldsymbol{\eta}_{ei(w)}\boldsymbol{X}_{ei} + \boldsymbol{\Lambda}_b\boldsymbol{\eta}_{i(b)} + \boldsymbol{\Gamma}_y\boldsymbol{Z}_i + \boldsymbol{\Gamma}_{\eta_{i(b)}y}\boldsymbol{\eta}_{i(b)}\boldsymbol{Z}_i + \boldsymbol{\epsilon}_{ei(w)} + \boldsymbol{\epsilon}_{i(b)}. \tag{8}$$

The segment $[\boldsymbol{\Gamma}_y\boldsymbol{X}_{ei} + \boldsymbol{\Gamma}_{\eta_{ei(w)}y}\boldsymbol{\eta}_{ei(w)}\boldsymbol{X}_{ei} + \boldsymbol{\Gamma}_y\boldsymbol{Z}_i + \boldsymbol{\Gamma}_{\eta_{i(b)}y}\boldsymbol{\eta}_{i(b)}\boldsymbol{Z}_i]$ in Equation (8) represents the part of the model that is used to test uniform and nonuniform factorial invariance across grouping covariates present in $\boldsymbol{X}_{ei}$ and $\boldsymbol{Z}_i$. When study period and treatment condition are included as covariates, the associated regression coefficients in Equation (8) indicate the status of intercept and factor loading invariance between the BLE period and ATP and between the placebo and T+S condition, respectively. Note that in the MIMIC procedure for testing invariance across groups, the causal effect of the grouping covariate on the latent variable should be included as shown for the within-patient situation in Equation (6). In the same vein, between-patient covariates must be included as predictors of the between-patient latent variables: $\boldsymbol{\eta}_{i(b)} = \boldsymbol{\Gamma}_{\eta_{i(b)}}\boldsymbol{Z}_i + \boldsymbol{\zeta}_{i(b)}$. The MIMIC approach presented here for testing invariance must be performed for each observed item separately. It therefore serves as a method to detect a particular noninvariant item.

Finally, we investigate across-level invariance of factor loadings. If configural invariance across levels cannot be rejected, a stricter form of across-level measurement invariance can be imposed by setting the factor loadings equal across levels. Imposing equal factor loadings across levels can be done by setting $\boldsymbol{\Lambda}_w = \boldsymbol{\Lambda}_b = \boldsymbol{\Lambda}$. Equation (3) then simplifies to $\boldsymbol{Y}_{ei} = \boldsymbol{\mu} + \boldsymbol{\Lambda}(\boldsymbol{\eta}_{ei(w)} + \boldsymbol{\eta}_{i(b)}) + \boldsymbol{\epsilon}_{ei(w)} + \boldsymbol{\epsilon}_{i(b)}$. Mehta and Neale (2005) show that with equal factor loadings across levels, $\eta_{ei(w)} + \eta_{i(b)} = \eta_{ei}$, where η_{ei} is a first-level latent variable that is composed of within- and between-patient deviations. This means that the within-patient latent variable now has a random intercept at the between-patient level. Equal factor loadings across levels also indicate that latent factor variances are directly comparable, because invariant factor loadings equates the scale across levels of the common latent variable (Mehta & Neale, 2005). By comparing the latent factor variances, the proportion of variance located at the between-patient level can be calculated. This measure is also known as the intraclass correlation (ICC). A large ICC value suggests higher trait variability, or larger differences between patients, but small differences within patients. A small ICC value indicates higher state variability, or small differences between patients, but larger differences within patients.

If across-level factor loadings are invariant, it implies that the common factors have the same interpretation across levels (Rabe-Hesketh et al., 2004). Furthermore, equal factor loadings result in a latent variable that itself has a random intercept. Therefore, equal factor loadings across levels are a necessary prerequisite for further extending the model with covariates and a latent random slope (which we will explain in the next section).

2.4 | ML-MIMIC model: Including covariates predicting the latent variable

The ML-MIMIC model was already introduced in Section 2.3 as a model suitable for testing uniform and nonuniform factorial invariance by including covariate effects predicting the observed items and latent variables. In this section, covariates are included as predictors of latent variables only. Again, let $\boldsymbol{X}_{ei}$ be a vector of covariates located at the within-patient level observed for patient i on event e and let $\boldsymbol{Z}_i$ be a vector of covariates located at the between-patient level observed for patient

i. The general model can then be written as follows (Cao et al., 2019):

$$\eta_{ei(w)} = \mathbf{\Gamma}_{(w)}\boldsymbol{X}_{ei} + \boldsymbol{\zeta}_{ei(w)}, \tag{9}$$

$$\eta_{i(b)} = \mathbf{\Gamma}_{(b)}\boldsymbol{Z}_{i} + \boldsymbol{\zeta}_{i(b)}, \tag{10}$$

where $\mathbf{\Gamma}_{(w)}$ and $\mathbf{\Gamma}_{(b)}$ represent the within-patient and between-patient effects of the covariates on the latent factors and where $\boldsymbol{\zeta}_{ei(w)}$ and $\boldsymbol{\zeta}_{i(b)}$ are the residual error terms. These error terms are multivariate normally distributed with means of zero and covariance matrices of $\mathbf{\Psi}_w$ and $\mathbf{\Psi}_b$ at the within-patient and between-patient level, respectively.

Considering the sexual event data, let x_{ei} be the covariate study period and let z_i be the covariate treatment condition. We further assume there exists one common latent variable at both levels with equal across-level loadings (see Figure 1). In the previous section, we explained that with invariant across-level loadings, the within-patient latent variable $\eta_{ei(w)}$ has a random intercept at the between-patient level: $\eta_{i(b)}$. Subsequently, we can rewrite the ML-MIMIC model as follows (Cao et al., 2019):

$$\boldsymbol{Y}_{ei} = \boldsymbol{\mu} + \mathbf{\Lambda}(\eta_{ei(w)} + \eta_{i(b)}) + \boldsymbol{\epsilon}_{ei(w)} + \boldsymbol{\epsilon}_{i(b)}, \tag{11}$$

$$\eta_{ei(w)} = \gamma_{(w)}x_{ei} + \zeta_{ei(w)}, \tag{12}$$

$$\eta_{i(b)} = \gamma_{(b)}z_{i} + \zeta_{i(b)}, \tag{13}$$

where $\gamma_{(w)}$ is the fixed effect of study period and $\gamma_{(b)}$ is the fixed effect of treatment condition. Equation (11) represents the measurement model part. Equations (12) and (13) represent the structural equation part. In particular, Equations (12) and (13) represent a regular multilevel model where the random intercept is the aggregate (average across second-level units) of the first-level latent dependent variable. This means that we can interpret the common factor at the between-patient level as the aggregate of the common factor located at the event level.

To include the cross-level interaction between study period and treatment condition (main treatment effect), the slope for study period, $\gamma_{(w)}$, is regressed on treatment condition. The slope for study period can further be modeled as a random effect by including a random slope term, indicating that the effect of study period is allowed to vary across patients. This leads to the following regression model equations (Cao et al., 2019):

$$\begin{aligned}
\eta_{ei(w)} &= \gamma_{i(w)}x_{ei} + \zeta_{ei(w),} \\
\eta_{i(b)} &= \gamma_{(b)}z_{i} + \zeta_{i(b)}, \\
\gamma_{i(w)} &= \gamma_{10} + \gamma_{(c)}z_{i} + \zeta_{\gamma_{i(w)}},
\end{aligned} \tag{14}$$

where $\gamma_{i(w)}$ now represents the random slope of study period. This random slope is predicted by the average slope γ_{10}, the effect of treatment condition $\gamma_{(c)}$ and a random slope term $\zeta_{\gamma_{i(w)}}$. This random slope has a mean of zero and its variance reflects the degree to which the fixed effect of study period varies across patients (after differences due to treatment condition between patients are considered). The regression effect $\gamma_{(c)}$ represents the cross-level interaction effect between study period and treatment condition.

The multilevel Model 14 allows for testing differences of study period and treatment condition on the latent variable sexual functioning. Under the assumption that there exists uniform and nonuniform factorial invariance across the two study periods and both treatment conditions, group differences on the observed SED item scores are only attributable to differences in the latent variable sexual functioning that can be examined by the model in (14). This can be illustrated when looking at Equation (11) and Model 14. Substituting the equations of Model 14 into Equation (11) and rearranging terms, we can write the expected value of $\boldsymbol{Y}_{ei}$ as

$$\mathrm{E}(\boldsymbol{Y}_{ei}) = \boldsymbol{\mu} + \mathbf{\Lambda}(\gamma_{10}x_{ei} + \gamma_{(c)}x_{ei}z_{i} + \gamma_{(b)}z_{i}). \tag{15}$$

Under the assumption of uniform and nonuniform factorial invariance, $\boldsymbol{\mu}$ and $\mathbf{\Lambda}$ in Equation (15) are assumed to be invariant across study periods and treatment conditions. Consequently, considering the placebo condition where $z_i = 0$, the

difference between both study periods on E($\boldsymbol{Y}_{ei}$) is $\boldsymbol{\mu} + \boldsymbol{\Lambda}(\gamma_{10})$. For the T+S condition, where $z_i = 1$, the difference between both study periods on E($\boldsymbol{Y}_{ei}$) is then $\boldsymbol{\mu} + \boldsymbol{\Lambda}(\gamma_{10} + \gamma_{(c)})$. For both conditions, $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are assumed to be equal indicating that the difference between both treatment conditions on the change from baseline is only present in γ_{10} and $\gamma_{(c)}$, which are the regression effects on the latent variable. In situations where $\boldsymbol{\mu}$ and/or $\boldsymbol{\Lambda}$ differ across treatment conditions, any differences in E($\boldsymbol{Y}_{ei}$) are also affected by different values in the intercepts and/or factor loadings between groups and that would complicate the interpretation of latent mean differences.

Finally, we will shortly discuss the univariate multilevel model using the sum score of the five SED items as outcome, as described in the previous work (Kessels et al., 2019). Instead of $\boldsymbol{Y}_{ei}$ being an observed vector of these SED items, Y_{ei} was denoted as the observed sum score of the five SED items of patient i on event e. The observed covariates study period ($0 = BLE, 1 = ATP$) and treatment condition ($0 = placebo, 1 = T + S$) were also denoted as x_{ei} and z_i, respectively. The multilevel regression model was written as

$$Y_{ei} = \beta_{0i} + \beta_{1i} x_{ei} + \epsilon_{ei}, \tag{16}$$

$$\beta_{0i} = \gamma_{00} + \gamma_{01} z_i + u_{0i}, \tag{17}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} z_i + u_{1i}, \tag{18}$$

where β_{0i} and β_{1i} are the random intercept and random slope for patient i, respectively. The single multilevel equation can be written as

$$Y_{ei} = \gamma_{00} + \gamma_{10} x_{ei} + \gamma_{01} z_i + \gamma_{11} z_i x_{ei} + u_{1i} x_{ei} + u_{0i} + \epsilon_{ei}, \tag{19}$$

where γ_{00} is the fixed intercept, γ_{10} is the fixed main effect of study period, γ_{01} is the fixed main effect of treatment condition and γ_{11} is the cross-level interaction. Furthermore, u_{0i} and u_{1i} represent the random intercept and random slope, respectively, and ϵ_{ei} represents the within-patient residual error term. The fixed regression coefficients γ_{10}, γ_{01}, and γ_{11} in Equation (19) correspond to $\gamma_{10}, \gamma_{(b)}$, and $\gamma_{(c)}$ in Model 14. Of primary interest in evaluating the drug effect in the clinical trial of the motivating example is $\gamma_{(c)}$: the fixed cross-level interaction effect that represents the difference in change from baseline between T+S and placebo. The other fixed regression coefficients, γ_{10} and $\gamma_{(b)}$, represent the change from baseline for placebo patients and the difference between placebo and T+S at baseline, respectively.

2.5 | Statistical analysis

Analyses in this paper were performed in `Mplus`, version 7.3 (Muthén & Muthén, 1998-2017). Data and `Mplus` code to generate the results are available as Supporting Information on the journal's web page (https://onlinelibrary.wiley.com/doi/bimj.202100046/suppinfo). We treated the five-point Likert-scale SED items as approximately continuous and used robust maximum likelihood (MLR) in `Mplus` to estimate the parameters. MLR provides standard errors that are robust for nonnormality.

To test the model fit of the ML-CFA model shown in Figure 1, the configural invariant model with one common factor, we considered the log likelihood value and its associated chi-square (χ^2) statistic. The chi-square is an absolute measure of fit that evaluates the discrepancy between the observed covariance matrix and the covariance matrix predicted by the model. The observed covariance matrix can be interpreted as a matrix under a saturated model. In practice, a nonsignificant chi-square statistic is interpreted as a good model fit. However, the chi-square statistic is often not used as a sole index for evaluating model fit as it has increasing power to detect trivial differences in large samples (Brown, 2014). We therefore also considered another absolute model fit index: the standardized root mean square residual (SRMR). The SRMR is the average discrepancy between the correlations observed in the observed matrix and the correlations predicted by the model. Because in an ML-CFA model there is an observed correlation matrix at the within- and between-patient level, an SRMR model fit index is obtained for both levels (SRMR_{within} and $\text{SRMR}_{between}$). SRMR values smaller than 0.08 are indicative of good model fit (Brown, 2014). We further inspected the root mean square error of approximation (RMSEA), an approximate fit index that evaluates to what extent a model fits reasonably well in the population by including a penalty function for the number of freely estimated parameters. An RMSEA below 0.08 is considered satisfactory and RMSEA values lower than 0.05 are indicative of good fitting models (Browne & Cudeck, 1992). All fit indices presented in this

section are automatically generated in Mplus output when employing MLR. For the RMSEA, we also reported the 90% confidence intervals, which were derived in R, version 4.0.4 (R Core Team, 2020), using the MBESS package, version 4.8.0 (Kelly, 2007b). Note that when fitting multilevel regression models in Mplus version 7.3, of the abovementioned fit indices, only the log likelihood is provided.

For testing across-level invariance of factor loadings, we considered the same abovementioned fit indices. In addition, we applied the chi-square difference test ($\Delta\chi^2$) comparing the chi-square fit of two nested models: the model with and without equal across-level factor loadings. The difference in chi-square yields a test statistic that is also chi-square distributed with degrees of freedom equal to the difference in the number of estimated parameters. A nonsignificant chi-square difference test indicates that the model with equal factor loadings across levels does not fit worse compared to the configural invariant model. When MLR is used for estimation, the Satorra–Bentler (SB)-scaled chi-square difference test is recommended (Muthén & Muthén, 1998-2017; Satorra & Bentler, 1994). In the SB chi-square difference test, a scaling correction factor (automatically estimated in Mplus) is inluded to better approximate the chi-square distribution under nonnormality. For the ML-CFA models, we reported both the scaled and unscaled chi-square statistics (as well as the scaled and unscaled chi-square differences). The unscaled chi-square statistics are derived under maximum likelihood estimation rather than MLR.

For testing uniform and nonuniform factorial invariance, the model in Equation (8) was fitted to the data. This model includes two interaction effects between a latent variable and an observed covariate that can be created in Mplus using the XWITH command and estimated using the integration algorithm. For these models, only the log likelihood as fit index is provided. Therefore, for testing uniform and nonuniform factorial invariance, we performed likelihood ratio tests comparing the log likelihood of two nested models: a model with four direct covariate effects (Equation (8)) and a constrained model where these covariate effects were fixed to be zero. Again, because MLR is used as estimation method, the Satorra–Bentler-scaled likelihood ratio (SB LR) test, using a scaled correction factor, was applied (Satorra & Bentler, 1994), in which we follow the procedure for testing uniform and nonuniform factorial invariance described by Kim et al. (2015). The SB LR test statistic is approximately chi-square distributed with degrees of freedom equal to the difference in estimated parameters. The SB LR test was applied for each SED item separately and in case the test produces a nonsignificant result, the constrained model is favored, indicating that there is uniform and nonuniform factorial invariance for one SED item. When the SB LR test yields a negative likelihood ratio test statistic, it is required to adjust the correction factor in order to ensure a positive test statistic (Satorra & Bentler, 2010). In detecting a noninvariant item, it has been shown that the likelihood ratio testing procedure of the ML-MIMIC model and the regular MIMIC model may lead to false positive results (Kim et al., 2015, 2012; Oort, 1998). Therefore, in addition to the likelihood ratio test, we also consider the direct covariate effects ($\boldsymbol{\Gamma}$ parameters in Equation (8)) to detect any noninvariant items (Barendse et al., 2010, 2012).

For testing direct covariate effects of the regression models, the estimated regression coefficients were divided by the estimated standard error to obtain a z-statistic that was used to derive a two-sided p-value.

3 | RESULTS

3.1 | Descriptive statistics

There were a total of 627 observations/sexual events (completed SEDs) for 53 patients with an average of 11.79 observations per patient. Of these 53 patients, 27 patients were randomized to placebo and 26 to the active drug T+S. The estimated sample statistics are summarized in Table 1. The ICCs for all items suggest approximately equal amounts of within-patient and between-patient variation of the observed variables.

3.2 | ML-CFA model and measurement invariance

The model fit indices of the ML-CFA model with a one-factor solution at both levels are presented in the first column of Table 2. Although the (scaled) chi-square value suggests that the model does not fit the data, the other fit indices all show that the model fits the data reasonably well. This result implies that configural invariance can be assumed: a common one-factor model for $\boldsymbol{Y}_{ei}$. The unstandardized parameter estimates of this model are shown in Table 3. The latent factor scales were identified by fixing the factor loading of pleasure to 1.00. The remaining factor loadings, latent factor variances, and residual variances were estimated without any constraints. The means of the factors were fixed to 0.00. All items

TABLE 1 Estimated sample statistics

Outcome	Pleasure	Inhibition	Desire	Bod. Arousal	Subj. Arousal
Means (SD)	3.23 (1.29)	3.20 (1.28)	2.90 (1.34)	2.89 (1.38)	2.94 (1.43)
ICC	0.46	0.52	0.48	0.51	0.51
	Between-patient covariance matrix				
Pleasure	0.79				
Inhibition	0.71	0.92			
Desire	0.74	0.69	0.91		
Bod. Arousal	0.82	0.72	0.77	0.99	
Subj. Arousal	0.85	0.77	0.87	0.98	1.06
	Within-patient covariance matrix				
Desire	0.94				
Bod. Arousal	0.65	0.85			
Subj. Arousal	0.63	0.58	0.97		
Pleasure	0.71	0.63	0.65	0.94	
Inhibition	0.74	0.69	0.73	0.79	1.02

Abbreviations: Bod, Bodily; ICC, intraclass correlation; Subj, Subjective; SD, standard deviation.

TABLE 2 Model fit indices

	Model		
Fit index	**ML-CFA**	**ML-CFA ($\mathbf{\Lambda}_w = \mathbf{\Lambda}_b$)**	**ML-CFA (sum score)**
Scaled χ^{2*}	26.407	28.187	110.454
Correction Factor	1.7838	1.9071	1.7690
Unscaled χ^2	30.559	32.451	172.719
Scaled $\Delta\chi^2$		3.002	84.218
Unscaled $\Delta\chi^2$		1.892	142.16
RMSEA (90% CI)	0.051 (0.028—0.075)	0.040 (0.017—0.062)	0.069 (0.056–0.083)
SRMR_{within}	0.011	0.011	0.040
$\text{SRMR}_{between}$	0.026	0.022	0.115

Note. The $\Delta\chi^2$ statistics are calculated relative to the ML-CFA. The correction factor is the scaling correction factor given in the `Mplus` output for the H0 model under the log likelihood estimates. The RMSEA is derived using the scaled χ^2.

*ML-CFA: df = 10, $p = .003$; ML-CFA ($\mathbf{\Lambda}_w = \mathbf{\Lambda}_b$): df = 14, $p = .013$; ML-CFA (sum score): df = 28, $p < .001$.

have approximately equal factor loadings with slightly larger factor loadings at the between-patient level compared to the within-patient level for subjective arousal, pleasure, and inhibition. Furthermore, for both levels, the same pattern in order of factor loading size is observed with inhibition having the largest factor loading and bodily arousal the smallest factor loading. When standardizing the loadings, all obtained correlations are larger than .79 that indicates a strong relationship between the items and the latent variable at both levels.

In Table 4, the results of the MIMIC procedure for testing the assumption whether there is uniform and nonuniform factorial invariance for study period and treatment condition are presented. For each item of the SED, a set of nested ML-MIMIC models was compared to test factorial invariance for study period and treatment condition simultaneously. First, four direct covariate effects were included on each observed SED item separately in one model to detect factorial noninvariance of the intercept and factor loading of that item. For example, in the first column of Table 4, the SED item pleasure was regressed on study period, treatment condition, and two interaction effects for testing intercept and factor loading invariance for study period at the within-patient level and treatment condition at the between-patient level. For each item, the model with the four covariate effects was compared to the constrained model where all four covariate effects were constrained to be zero assuming factorial invariance at both levels for the variable. The log likelihood of the constrained model is presented in the last column of Table 4. The model comparison was evaluated using the SB LR test.

All SB LR tests were not significant, except for the SED item bodily arousal. This suggests uniform and nonuniform factorial invariance for all SED items, except for bodily arousal. We subsequently tested the intercepts and factor loadings

TABLE 3 Unstandardized parameter estimates ML-CFA

	Within patients[a]		Between patients[b]	
Item	**Factor loadings**	**Residual variances**	**Factor loadings**	**Residual variances**
Pleasure	1.00	0.25 (0.03)	1.00	0.07 (0.03)
Inhibition	0.91 (0.04)	0.28 (0.04)	0.91 (0.09)	0.33 (0.09)
Desire	0.94 (0.07)	0.37 (0.05)	1.00 (0.11)	0.19 (0.08)
Bod. Arousal	1.03 (0.05)	0.21 (0.02)	1.14 (0.07)	0.07 (0.04)
Subj. Arousal	1.11 (0.05)	0.18 (0.02)	1.20 (0.08)	0.02 (0.02)

Note. Numbers in parentheses represent the standard error.
aFactor variance at the within-level was 0.68 (0.12).
bFactor variance at the between-level was 0.71 (0.16).

TABLE 4 Uniform and nonuniform factorial invariance testing using the ML-MIMIC procedure

Item tested for factorial invariance						**Constrained**
Parameter	**Pleasure**	**Inhibition**	**Desire**	**Bod. Arousal**	**Subj. Arousal**	**Model**
Within patients						
Study period[a]	0.10	−0.23	−0.05	0.05	0.09	0
Interaction[b]	−0.20	0.11	0.10	−0.13	−0.02	0
Between patients						
Treatment group[a]	−0.03	−0.05	0.23	−0.07	0.01	0
Interaction[b]	−0.06	−0.15	0.18	0.51	−0.54	0
Model fit						
Free parameters	26	26	26	26	26	22
Log likelihood	−3564.043	−3565.587	−3562.118	−3554.134	−3554.948	−3570.83
Correction factor	4.733	4.455	4.608	4.499	4.699[c]	4.870
SB LR	3.411	4.827	5.502	13.582*	8.451	

Abbreviations: Bod, Bodily; Subj, Subjective; SB LR, Satorra–Bentler scaled likelihood ratio test.
aUnstandardized regression coefficients as uniform invariance ($\mathbf{\Gamma}_y$ in Equation (8)).
bUnstandardized regression coefficients as nonuniform invariance ($\mathbf{\Gamma}_{\eta_{ei(w),y}}$ and $\mathbf{\Gamma}_{\eta_{i(b),y}}$ in Equation (8)).
cSB LR test yielded a negative χ^2 value, so an adjusted correction factor was used (Satorra & Bentler, 2010).
*$p = .009$.

for the item bodily arousal at each level separately using the SB LR test to detect the source of noninvariance. This procedure revealed that only the intercepts at the within-patient level between study periods were invariant, suggesting only uniform invariance between study periods (results not shown). When considering the regression coefficients of the covariates (upper part of Table 4), it shows that for bodily arousal (as well as for all other items), all four direct covariate effects were not significant implying uniform and nonuniform factorial invariance. These findings contradict the result of the SB LR test for bodily arousal, whereas for all other items, the SB LR test and the results of the covariate effects lead to the same conclusion.

The results of the ML-MIMIC modeling procedure to test factorial invariance reveal uniform and nonuniform factorial invariance across the covariates study period and treatment condition for the SED items such as Pleasure, Inhibition, Desire, and Subjective Arousal, whereas for Bodily Arousal, this can be questioned. However, as the likelihood ratio test may have lead to a false positive result for the item Bodily Arousal and because all direct covariate effects were not significant, we concluded that all item intercepts and factor loadings were invariant across study periods and treatment conditions.

Under the assumption that the intercepts and factor loadings are invariant across study periods and treatment conditions, we continued with the configural invariant model presented in Figure 1 that was tested at the beginning of this section.

The fit indices of the ML-CFA of Figure 1 with the factor loadings constrained to be equal across levels are presented in the second column of Table 2. Also for this model, the (scaled) chi-square value suggests a bad model fit, whereas the other fit indices show the model has an excellent fit. The scaled chi-square difference between the model with and without

TABLE 5 Unstandardized parameter estimates ML-CFA with invariant factor loadings

	Within patients[a]		Between patients[b]	
Item	**Factor loadings**	**Residual variances**	**Factor loadings**	**Residual variances**
Peasure	1.00	0.25 (0.03)	1.00	0.07 (0.03)
Inhibition	0.92 (0.04)	0.28 (0.04)	0.92 (0.04)	0.33 (0.09)
Desire	0.95 (0.06)	0.37 (0.05)	0.95 (0.06)	0.19 (0.08)
Bod. Arousal	1.05 (0.04)	0.21 (0.02)	1.05 (0.04)	0.07 (0.04)
Subj. Arousal	1.13 (0.04)	0.18 (0.02)	1.13 (0.04)	0.03 (0.02)

Note. Numbers in parentheses represent the standard error.
aFactor variance at the within-level was 0.67 (0.11).
bFactor variance at the between-level was 0.79 (0.15).

equal factor loadings across levels was $\Delta\chi^2 = 3.002$ (df = 4, $p = .557$), indicating that factor loadings can be considered equal. The unstandardized parameter estimates of the model with invariant factor loadings are presented in Table 5 and are generally very similar to the estimates of the unconstrained ML-CFA in Table 3. Assuming a configural invariant model with across-level factor loading invariance indicates that the multiple indicators reflect within- and between-components of the same latent variable. This is an important finding, because the common factor at the patient level can now be interpreted as the aggregate of the event-level factor. In other words, there now is a within-patient latent variable with a random intercept at the between-patient level. Cross-level invariant factor loadings also allow us to calculate the ICC of the common latent variable. The variance for the random intercept, or latent trait variance, is 0.79 and the corresponding within-patient variance, or latent state variance, is 0.67 resulting in an ICC of 0.54. This ICC means that there is an approximately equal amount of variability at the state and trait level. This implies that a patient's sexual functioning fluctuates over time that is comparable to how sexual functioning fluctuates between patients in the population. These results show that the variability within a patient takes place on the same dimension that describes the difference between patients.

3.3 | ML-MIMIC model: Including covariates predicting the latent variable

The findings of uniform and nonuniform factorial invariance across study periods and treatment conditions indicates that differences in SED items between study periods and treatment conditions are only attributable to differences in sexual functioning. This allows us to compare the latent variable sexual functioning between study periods and treatment conditions. Furthermore, the finding of across-level invariant factor loadings indicates that the latent variable at the within-patient level has a random intercept. Together, these findings make it possible to use the latent variable sexual functioning as outcome in the structural equation part of the model with covariates predicting the latent variable at both levels: the ML-MIMIC model. In this ML-MIMIC model, the main interest was focused on the cross-level interaction effect between study period and treatment condition on the latent variable sexual functioning to evaluate whether the active drug T+S had a benefit over placebo. Study period was included as a predictor of the latent variable sexual functioning at the event level together with a latent random slope at the patient level. Treatment condition was included as a predictor of the latent variable sexual functioning at the patient level as well as a predictor of the latent random slope of study period at the patient level, which covers the cross-level interaction of interest (as described by Model 14 in Section 2.4). The results of the structural equation part of the ML-MIMIC model are presented in the first column of Table 6. The results of the measurement part of the ML-MIMIC model (the factor model) are not included in this table, but these results are fairly close to the results presented in Table 5. The effect of study period, γ_{10}, on the within-patient latent variable sexual functioning is .59 ($p < .001$). This illustrates that the sexual functioning latent score increases significantly during the ATP for the placebo condition. The effect of treatment condition on the between-patient latent variable sexual functioning is −.35 and was not significant ($p = .158$), indicating that there exists no difference between T+S and placebo at baseline. The cross-level interaction effect between study period and treatment condition is .56 ($p = .040$) and can be interpreted as an additional increase in change from baseline for the T+S patients, revealing that there is a significant benefit of the active drug T+S compared to placebo. The variance of the latent random slope of study period is 0.71, which means that there is a substantial variation in the effect of study period at the patient level (varying regression slopes), after differences due to treatment condition have already been accounted for. These varying regression slopes for study period are assumed to follow a normal distribution. The standard deviation of the slope is equal to $\sqrt{0.71} = 0.84$. This standard deviation has a

TABLE 6 Parameter estimates structural equation part of the ML-MIMIC model

	Parameter	ML-MIMIC model Estimate (SE)	ML-MIMIC model† Estimate (SE)
Intercepts	Pleasure	2.84 (0.17)	2.84 (0.17)
	Inhibition	2.75 (0.17)	2.85 (0.17)
	Desire	2.55 (0.16)	2.53 (0.17)
	Bod. Arousal	2.51 (0.17)	2.56 (0.18)
	Subj. Arousal	2.54 (0.18)	2.49 (0.19)
Covariates	Study period	0.59 (0.16)**	0.61 (0.15)**
	Treatment group	-0.35 (0.25)	-0.32 (0.25)
	Study period × Treatment group	0.56 (0.27)*	0.54 (0.28)
Variances	Factor within-level	0.32 (0.05)	0.32 (0.07)
	Factor between-level	0.66 (0.12)	0.55 (0.20)
	Random slope study period	0.71 (0.19)	0.65 (0.19)

Note. Numbers in parentheses represent the standard error.

†ML-MIMIC model while controlling for noninvariance across study periods and treatment conditions for Bodily Arousal.

*$p < .05$, **$p < .001$.

useful characteristic, because with normally distributed individual patient, slopes approximately 68.3% of the individual patient slopes are predicted to lie between one standard deviation below and one standard deviation above the average slope. This standard deviation can therefore be used to derive a predictive interval of individual slopes (Hox et al., 2018, p. 16). For placebo patients, the average increase in sexual functioning from the BLE period to the ATP was 0.59, which indicates that approximately 68.3% of the placebo patients have a predicted change from the BLE period to the ATP lying within the predictive interval −0.25 to 1.43. For the T+S patients, the average increase in sexual functioning from the BLE period to the ATP was $0.59 + 0.56 = 1.15$, indicating that approximately 68.3% of the T+S patients have a predicted change from the BLE period to the ATP lying within the predictive interval 0.31–1.99. These predictive intervals reveal that some placebo patients have a predicted decrease during the ATP compared to the BLE period.

To investigate whether the possible uniform and nonuniform factorial invariance violation of the SED item Bodily Arousal affected the results of the ML-MIMIC model analyzed in this section, we reestimated the model parameters presented in the first column of Table 6 while controlling for possible noninvariance across study periods and treatment conditions for Bodily Arousal. This was done by using the unconstrained model presented in Table 4 for Bodily Arousal while also including study period and treatment condition as predictors of the latent variable. The results of this model are presented in the second column of Table 6 and do not differ much compared to the estimated parameters of the ML-MIMIC model where uniform and nonuniform factorial invariance is assumed across all items. This result strengthens our conclusion that all items are invariant across study periods and treatment conditions.

When comparing the results of the ML-MIMIC model applied in this paper to the results of the sum score multilevel regression (Equation (19)) applied in Kessels et al. (2019), we find similar results leading to the same conclusion regarding the drug effect of T+S. The reason for these similar results is because the (unstandardized) factor loadings and residual variances of the different items of the measurement part of the ML-MIMIC model do not differ much from each other. McNeish and Wolf (2020) showed that the sum score model can be specified as a factor model by constraining all standardized factor loadings to the same value across items. To obtain equal standardized factor loadings across items, the unstandardized factor loadings and residual variances of the items need to be constrained to the same value when fitting the model. Because our results reveal that the unstandardized factor loadings and residual variances do not differ much from each other, the fitted ML-CFA model approximates the sum score model, which is why the results of the ML-MIMIC do not differ much from an analysis on sum scores. For comparison, we fitted the ML-CFA sum score model by constraining the unstandardized loadings and residual variances to be equal across items of which the fit indices are presented in the third column of Table 2. The (scaled) chi-square value shows that this model has a bad model fit and compared to the unconstrained ML-CFA model (first column of Table 2), the sum score model fits the data worse (scaled chi-square difference was $\Delta\chi^2 = 84.218$ ($df = 18$, $p < .001$)). This indicates that the sum score model should not be preferred over the unconstrained ML-CFA model. However, the RMSEA and $SRMR_{within}$ of the sum score model show that the model fit is still satisfactory and given these fit indices, the sum score model has not a great misfit to the data.

The results between the ML-MIMIC model and the univariate multilevel regression also tend to be more similar the more reliable the SED items are. We calculated the reliability coefficients (Cronbach's) Alpha and Omega at the within-patient and between-patient level separately as described by Geldhof et al. (2014), who explain how to estimate reliability coefficients in an ML-CFA framework. The Alpha coefficient is specified as a function of the average interitem covariance matrix within a scale, the variance of the scale score, and the number of items. Alpha is considered a consistent estimate of the reliability when all items have equal loadings on the same scale. The Omega coefficient is an estimate of composite reliability and takes into account possible heterogeneity in item factor loadings. At the within-patient level, Alpha was estimated to be .93 and Omega was also .93. At the between-patient level, Alpha was .88 and Omega .97. These results show that the items are very reliable and that the Alpha and Omega values are very close to each other, which confirm our findings that the analysis on the sum score is similar to an analysis on factor scores using the ML-MIMIC model.

Compared to the results in Table 5, the factor variances at both levels decreased in Table 6, indicating that part of the factor variance is now explained by the inclusion of the covariates.

Finally, we investigated the effect of two patient-level covariates: age (mean = 43.1, standard deviation (SD) = 11.24) and body mass index (BMI) score (mean = 26.7, SD = 7.27). These were added to the model as predictors of the patient-level latent variable sexual functioning. Both covariates have a very small nonsignificant effect (age: −0.002, $p = .870$, BMI: 0.013, $p = .450$) and the remaining parameters only changed slightly.

4 | DISCUSSION

The main purpose of this study was to demonstrate how a drug effect on a patient reported outcome with measurements taken at multiple events could be assessed more accurately using a data set that consists of two key characteristics. First, the patient-reported outcome measures a latent variable. Second, the data follow a hierarchical structure with events nested within patients and where the number of available events is highly unbalanced across patients. We utilized an ML-SEM approach (ML-CFA model and ML-MIMIC model) to overcome several limitations that arise when aggregating data possessing these key characteristics. Specifically, the ML-SEM model was employed to test whether a one-factor model at both levels with equal across-level factor loadings fitted the data, to examine if uniform and nonuniform factorial invariance across groups at the within-patient level and at the between-patient level could be assumed, and to assess the effect of an active drug compared to placebo on the latent variable measured over multiple events divided over two study periods.

The current research is an extension of Kessels et al. (2019) and was intended to solve possible issues regarding the aggregation of the five SED items that was ignored in Kessels et al. (2019): creating a sum score of the five SED items prior to analysis. The ML-SEM model presented in this paper prevented the need for sum scoring *a-priori* because the measurement model of the multiple items was directly embedded into a model with a structural equation part to estimate all parameters simultaneously: the ML-MIMIC model. This approach overcomes possible limitations that may arise when first computing a sum score, such as giving all items equal weight. By first fitting a factor model at the event level and at the patient level simultaneously, we showed that a one-factor model at both levels fitted the data reasonably well. This result is consistent with previous findings on the same sexual event data (Kessels et al., 2019; Van Nes et al., 2018, 2017) and confirmed our expectation that the patient reported outcome measured the latent variable sexual functioning. Subsequently, in the same model where this latent variable was measured, this latent variable was used as outcome predicted by the covariates study period and treatment condition. The main results of the ML-MIMIC model revealed a benefit of T+S compared to placebo on the change from baseline on sexual functioning.

The treatment effect of T+S in the ML-MIMIC model was very similar compared to the treatment effect of a regular multilevel analysis using the sum score of the five SED items as outcome (Kessels et al., 2019). This finding can easily be misinterpreted, because it suggests that the ML-SEM procedure described in this paper (multilevel factor analysis, measurement invariance testing, and assessing group differences on a latent variable) would be redundant. Especially when considering the relative complexity of the ML-SEM procedure compared to the regular multilevel model, the question could be asked why all these complicated models should be performed. We acknowledge that this would be a valid argument for this specific example, but we still recommend applying the ML-SEM model. It is essential that researchers realise that sum scoring is, in fact, a latent variable model (McNeish & Wolf, 2020). As explained in Section 3.3, the sum score model can be specified as a latent variable model with equal unstandardized factor loadings and residual variances across items to obtain equal standardized loadings. Therefore, a justified use of sum scoring requires statistical evidence and

should be reported identically as a latent variable model. Our results revealed that the sum score model had acceptable fit for some fit indices. Consequently, it is imperative to always start studying the factor model to verify if the items measure the intended latent variable before doing any analysis and drawing conclusions based on potentially untrustworthy sum score. Once it has been verified, the items measure the intended latent variable to some extent (i.e., factor loadings that exceed a certain threshold), caution is still advised when subsequently calculating sum scores, because sum scores only match factor scores when the standardized factor loadings of the items are identical. Congruence between true factor scores of a latent variable model and sum scores can also be investigated by inspecting their Pearson's correlation coefficient. Many researchers would interpret a correlation coefficient with a magnitude of .80 as statistical evidence that scores are nearly equal and that sum scores can then be used rather than factor scores. However, this is very misleading. It has namely been explained that with a correlation of .80, there still exists $1 - .80^2 = 36\%$ variability in scores between sum scores and factor scores (McNeish & Wolf, 2020). This means that for two patients having the same sum score, there still can be a large discrepancy between true factor scores. In fact, only in the situation where the correlation exceeds .99, the variability in scores becomes negligible. Consequently, only in such extreme cases, preferring sum scores over factor scores is actually justified and the exact same results are then expected. Admittedly, a small discrepancy between factor scores and sum scores will not have an adverse impact when sum scores are used to obtain an approximation of a quantitative scale. However, when the goal is to evaluate treatment efficacy in clinical trials, scores need to be as precise as possible. Therefore, we encourage researchers always to apply the ML-SEM approach if applicable, because equal factor loadings and error variances across the items are rarely expected in practice.

In addition to sum scores, there exist multiple other ways of computing scale scores prior to analysis, which generally can be divided into nonrefined methods and refined methods. Nonrefined methods include the sum score and the weighted sum score where the weights can be determined by clinical experts. Weighted sum scores have been proposed in assessing quality of life in clinical research (Schumacher et al., 1991). In contrast to nonrefined methods, refined methods, such as factor score regressions, are based on linear combinations of the observed items (Devlieger et al., 2016). From a methodological viewpoint, refined methods are superior to nonrefined methods, but refined methods can be quite complex to comprehend for clinicians. Moreover, it has been illustrated that different refined scoring methods can lead to different conclusions, even in situations where the correlation between the scores is close to 1 (McNeish & Wolf, 2020). In clinical applications, nonrefined methods are most often used due to their simplicity. As pointed out in the previous paragraph, nonrefined methods require clear statistical justification and scores based on nonrefined methods are only valid when their correlation with factor scores is extremely high. The issue of which scoring method to use can be avoided if the measurement model for the items is directly embedded into the structural equation model (Devlieger et al., 2016). The ML-MIMIC presented in this paper meets this ability and is therefore recommended over an approach where scores needs to be computed prior to analysis.

In this paper, we also demonstrated how to test for measurement invariance across grouping covariates at the within-patient level and between-patient level. Measurement invariance of intercepts and factor loadings is implicitly assumed in an ML-MIMIC model. However, violation of this assumption complicates the interpretation when comparing latent means between groups. Because multigroup factor analysis cannot be used to test equality of intercepts and factor loadings of grouping covariates at the within-patient level, we employed an ML-MIMIC procedure described by Kim et al. (2015), where intercept and factor loading invariance testing was implemented by including the covariate study group as a predictor of the observed items. ML-MIMIC modeling has great potential in testing factorial invariance for both within- and between-level grouping covariates simultaneously, whereas this is challenging in multigroup factor analysis (Kim et al., 2015). In this paper, we concluded that there existed uniform and nonuniform factorial invariance for the event-level and patient-level covariates indicating equality of intercepts and factor loadings across study periods and treatment conditions. This resulted in a clear interpretation regarding differences on sexual functioning. In this paper, we argue that full measurement invariance (i.e., all factor loadings and intercepts are invariant) is a prerequisite to properly interpret latent mean differences. However, studies have shown that the presence of partial measurement invariance (some, but not all factor loadings and/or intercepts are noninvariant; Vandenberg & Lance, 2000) across groups can still result in unbiased covariate effects as long as the model is corrected for these partially invariant factor loadings and/or intercepts (Guenole & Brown, 2014; Hsiao & Lai, 2018). We showed that our modeling procedure is able to correct for partially invariant factor loadings and intercepts by adding the covariates study period and treatment condition predicting the latent variables while allowing the intercepts and factor loadings of the item Bodily Arousal to differ across study periods and treatment conditions. Researchers should be aware that the accuracy of covariate effects is affected by the measurement invariance status (Hsiao & Lai, 2018). This emphasizes the importance of testing the measurement invariance assumption before assessing covariate effects in multilevel data.

Another possibility to test for uniform and nonuniform measurement invariance across groups is to use a selection of anchor items. Anchor items are items that are presumed to be invariant across groups (Woods, 2009). The MIMIC model has been shown to be good method to select these anchor items (Chun et al., 2016). These anchor items are then used to identify a minimally constrained MIMIC model and this minimally constrained MIMIC model can then be used to test invariance of other items. This procedure has been shown to obtain less inflated type 1 error rates than the item invariance detecting procedure described in this paper (Chun et al., 2016). This could offer a promising method for testing measurement invariance in ML-MIMIC models.

We further explained the importance of across-level factor loading invariance as invariance of factor loadings implies that the latent factor has a random intercept itself that can then be used in a multilevel regression. Recently, Cao et al. (2019) showed that when the across-level factor loading invariance assumption was violated, the performance of the ML-MIMIC model in detecting an accurate (unbiased) cross-level interaction effect was comparable to the condition where factor loadings were assumed to be invariant across levels. The authors conclude that the measurement part in an ML-MIMIC model does not affect the performance in estimating the structural part of the model, or the cross-level interaction effect, as long as the measurement part is correctly specified. This finding suggests that the across-level invariance assumption may be relaxed without harming the model performance and applicability. However, it remains unclear how this will alter interpretations of treatment effects. Overall, future research is required to better understand the influence of noninvariance in ML-MIMIC models and how to solve this.

Another reason to prefer the ML-SEM model compared to a regular multilevel model is the ability to study the latent state and trait paradigm. When across-level invariance in factor loadings holds, the factor variances at both levels can be compared using the ICC to study the trade-off between the amount of latent state and trait variability, which offers additional insights when interpreting the data at the within-patient level and at the between-patient level. In this paper, we found an equal amount of variability at the state and trait level, indicating that the differences within patients are approximately equal to differences between patients. Note that states and traits can also be analyzed using regular multilevel analyses (Nezlek, 2007). However, in an ML-CFA model, it is possible to explore if different factor structures exist at the state and trait level. This enables one to explain why factor structures found between patients are not necessarily equal to factor structures found within patients. This additional possibility of the ML-CFA model increases the flexibility of the ML-SEM approach compared to the univariate multilevel regression, but this also introduces a severe limitation when extending the ML-CFA model as ML-MIMIC model. Different factor structures found at the within- and between-patient level violate the first necessary configural invariance assumption discussed in this paper. This means that the different latent variables cannot be used in a multilevel regression as the second-level latent variable, then no longer is an aggregate of a first-level latent variable. Although such a finding would result in the inapplicability of the ML-SEM approach, it also strengthens our argument that researchers should always start evaluating the factor structure at both levels rather than ignorantly using sum scores.

Another point of discussion is under what circumstances the ML-MIMIC model has sufficient power to detect the cross-level interaction effect between study period and treatment condition. In our study, there was one latent variable at both levels with an ICC of .54, a total sample size of 53 patients and 627 events, and a cross-level interaction effect of .56. Under similar simulation conditions, Cao et al. (2019) showed that the ML-MIMIC model has sufficient power to detect an unbiased cross-level interaction effect between two dichotomous covariates, indicating that the power for testing the cross-level interaction effect in our study is acceptable. A related question would be whether and when the ML-MIMIC model has more power in testing the (fixed) interaction effect compared to multilevel regression analyses using the univariate sum score as outcome, either in the situation where the individual events are considered or in the situation where the sum score is aggregated over the events. To answer this question, substantive simulation studies are required, because multiple design factors are expected to have an influence, such as the sample size, heterogeneity of factor loadings, degree to which measurement invariance assumptions are violated, ICC, and the size of the interaction effect. However, we consider this to be beyond the scope of this paper and see this as future work.

As a final note, we would like to point out that the ML-SEM procedure described in this manuscript does not resolve the issue of possible heterogeneous first-level residual variances across groups. In standard two-level multilevel models, it is assumed that the residual error variances at the first level are constant across the units specified at the first level (i.e., events): homogeneity of residual error variances. In multilevel models, it is possible to test hypotheses regarding sources of possible variance heterogeneity. This can be done by extending the standard univariate multilevel model presented in this manuscript and our previous work (Kessels et al., 2019) by modeling the residual variance as a log-linear function of first-level and second-level predictors and by allowing the first-level predictors of this function to vary across second-level units with random regression coefficients (Raudenbush & Bryk, 2002). Research has shown that ignoring possible

heterogeneity in first-level residual variances can result in spurious precision in estimating the residual variance (Leckie et al., 2014). As the ML-MIMIC model implicitly assumes invariance across groups, it also assumes that the factor error variances are invariant. Previous research has shown that factor (error) variance heterogeneity does not have an impact on the performance of ML-MIMIC models in estimating interaction effects with balanced groups and data that are normally distributed (Kim & Cao, 2015). However, whether the ML-MIMIC model can model heterogeneous first-level residual (factor) variance the same way it can be performed using standard multilevel models is not clear yet. Furthermore, studying the precision on estimating residual factor variances under variance heterogeneity with ML-MIMIC models would be an interesting topic for future work.

In conclusion, the ML-SEM procedure presented in this paper combines both the advantages of ML-CFA and multilevel analysis in one model thereby solving many issues (apart from the possible heterogeneous first-level residual variance across groups problem) concerned with aggregating the data prior to analysis. The results of this paper confirm the advantages of the ML-SEM model, described by Rabe-Hesketh et al. (2004): the model was able to cope with unbalanced on-demand medication data, the model included a factor structure at both levels (event level and patient level) of the hierarchy precluding the need for calculating sum scores *a-priori*, and the model enabled the estimation of a fixed and random drug effect. We showed by calculating predictive intervals around the random slope using the estimated slope variance how to give these fixed and random drug effects meaningful interpretations. The present paper reflects an illustration of an ML-SEM analysis method (ML-CFA model and ML-MIMIC model) to assess a drug effect on a patient reported outcome measured at multiple events. We therefore believe that the presented ML-SEM model in this paper offers a promising analysis method for testing treatment effects in clinical trials that use patient reported outcome measures. Furthermore, we demonstrated how to test for measurement invariance across groups and levels using the same ML-MIMIC model. In addition to our recommendation of adopting the ML-SEM approach, we also believe our demonstration of how to test for measurement invariance will be a valuable contribution to the literature in its own right.

ACKNOWLEDGMENT

This work was supported by Emotional Brain BV. The clinical trial on which data the current research was based was funded by Emotional Brain BV as part of a drug development program. For the current research, Emotional Brain BV provided support in the form of salaries for authors R. Kessels and J. Bloemers, and a consultation fee for author P.G.M. van der Heijden. The funder did not have any additional role in the study design and analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST

R. Kessels and J. Bloemers are (former) employees of Emotional Brain BV and P.G.M. van der Heijden was a consultant to Emotional Brain BV. J. Bloemers and P.G.M. van der Heijden own share options in Emotional Brain BV.

DATA AVAILABILITY STATEMENT

Data and Mplus code to generate the results are available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/bimj.202100046/suppinfo)

ORCID

Rob Kessels https://orcid.org/0000-0002-2479-3872
Mirjam Moerbeek https://orcid.org/0000-0001-5537-1237

REFERENCES

Barendse, M., Oort, F. J., & Garst, G. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. *Advances in Statistical Analysis*, *94*, 117–127.

Barendse, M., Oort, F. J. Werner, C.Ligtvoet, R., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 561–579.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*, 135–167.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. John Wiley & Sons.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Brown, H., & Prescott, R. (2015). *Applied mixed models in medicine* (3rd ed.). John Wiley & Sons.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230–258.

Cao, C., Kim, E. S., Chen, Y., Ferron, J., & Stark, S. (2019). Exploring the test of covariate moderation effects in multilevel MIMIC models. *Educational and Psychological Measurement*, *79*, 512–544.

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. *Applied Psychological Measurement*, *40*, 486–499.

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistical Medicine*, *16*, 2349–2380.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, *114*, 174–184.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*, 529–569.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*, 741–770.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*, 1–11.

Donaldson, G. W. (2003). General linear contrasts on latent variable means: Structural equation hypothesis tests for multivariate clinical trials. *Statistics in Medicine*, *22*, 2893–2917.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2003). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980.

Hox, J., Moerbeek, M., & Van De Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Taylor & Francis Group.

Hsiao, Y. Y., & Lai, M. H. (2018). The impact of partial measurement invariance on testing moderation for single and multi-level data. *Frontiers in Psychology*, *9*, 740.

Jak, S., Oort, F. J., & Dolan, C. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 265–282.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639.

Kaplan, D., Kim, J. S., & Kim, S. Y. (2012). Multilevel latent variable modeling: Current research and recent developments. In R. E. Millsap, & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 592–612). Publications Ltd.

Kelly, K. (2007b). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*, 1–24.

Kessels, R., Bloemers, J., Tuiten, A., & Van Der Heijden, P. G. M. (2019). Multilevel analyses of on-demand medication data, with an application to the treatment of female sexual interest/arousal disorder. *PLoS One*, *14*, e0221063.

Kifley, A., Heller, G. Z., Beath, K. J., Bulger, D., Ma, J., & Gebski, V. (2012). Multilevel latent variable models for global health-related quality of life assessment. *Statistics in Medicine*, *31*, 1249–1264.

Kim, E. S., & Cao, C. (2015). Testing group mean differences of latent variables in multilevel data using multiple-group multilevel CFA and multilevel MIMIC modeling. *Multivariate Behavior*, *50*, 436–456.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*, 469–492.

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (2015). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 603–616.

Leckie, G., French, R., Charlton, C., & Brown, W. (2014). Modeling heterogeneous variance–covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, *39*, 307–332.

McNeish, D., & Wolf, M. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*, 2287–2305.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Merz, E. L., & Roesch, S. C. (2011). Modeling trait and state variation using multilevel factor analysis with PANAS daily diary data. *Journal of Research in Personality*, *45*, 2–9.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nezlek, J. (2007). A multilevel framework for understanding relationships among traits, states, situations and behaviours. *European Journal of Personality: Published for the European Association of Personality Psychology*, *21*, 789–810.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*, 107–124.

Rabe-Hesketh, S., & Skrondal, A. (2008). Classical latent variable models for medical research. *Statistical Methods in Medical Research*, *17*, 5–32.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications.

Roesch, S., Aldridge, A., Stocking, S., Villodas, F., Leung, Q., Bartley, C. E., & Black, L. J. (2010). Multilevel factor analysis and structural equation modeling of daily diary coping data: Modeling trait and state variation. *Multivariate Behavioral Research*, *45*, 767–789.

Rovine, M. J., & Molenaar, P. C. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, *35*, 51–88.

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye, & C. Clogg (Eds.), *Latent variable analysis: Applications for development research* (pp. 399–464). Sage.

Satorra, A., & Bentler, P. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*, 243–248.

Schumacher, M., Olschewski, M., & Schulgen, G. (1991). Assessment of quality of life in clinical trials. *Statistics in Medicine*, *10*, 1915–1930.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press Inc.

Song, X. Y., Lee, S. Y., & Hser, Y. I. (2008). A two-level structural equation model approach for analyzing multivariate longitudinal responses. *Statistics in Medicine*, *27*, 3017–3041.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.

Stapleton, L., Yang, J., & Hancock, G. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*, 481–520.

Tuiten, A., van Rooij, K., Bloemers, J., Eisenegger, C., van Honk, J., Kessels, R., Kingsberg, S., Derogatis, L. R., de Leede, L., Gerritsen, J., Koppeschaar, H. P. F., Olivier, B., Everaerd, W., Frijlink, H. W., Höhle, D., de Lange, R. P. J., Böcker, K. B. E., & Pfaus, J. G. (2018). Efficacy and safety of on-demand use of 2 treatments designed for different etiologies of female sexual interest/arousal disorder: 3 randomized clinical trials. *Journal of Sexual Medicine*, *15*, 201–216.

Van Nes, Y., Bloemers, J., van der Heijden, P. G. M., Van Rooij, K., Gerritsen, J., Kessels, R., DeRogatis, L., & Tuiten, A. (2017). The Sexual Event Diary (SED): Development and validation of a standardized questionnaire for assessing female sexual functioning during discrete sexual events. *Journal of Sexual Medicine*, *14*, 1438–1450.

Van Nes, Y., Bloemers, J., Kessels, R., Van Der Heijden, P. G. M., van Rooij, K., Gerritsen, J., DeRogatis, L., & Tuiten, A. (2018). Psychometric properties of the Sexual Event Diary in a sample of Dutch women with female sexual interest/arousal disorder. *Journal of Sexual Medicine*, *15*, 722–731.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42–57.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kessels, R., Moerbeek, M., Bloemers, J., & van der Heijden, P. G. M. (2021). A multilevel structural equation model for assessing a drug effect on a patient-reported outcome measure in on-demand medication data. *Biometrical Journal*. *63*:1652–1672. https://doi.org/10.1002/bimj.202100046