Random item slope regression: an alternative measurement model that accounts for both similarities and differences in the association with individual items

Article

Published Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1037/met0000587

the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Psychological Methods

Random Item Slope Regression: An Alternative Measurement Model That Accounts for Both Similarities and Differences in Association With Individual Items

Ed Donnellan, Satoshi Usami, and Kou Murayama

Random Item Slope Regression: An Alternative Measurement Model That Accounts for Both Similarities and Differences in Association With Individual Items

Ed Donnellan[1, 2], Satoshi Usami[3], and Kou Murayama[2, 4]

[1] Department of Experimental Psychology, University College London
[2] School of Psychology and Clinical Language Sciences, University of Reading
[3] Graduate School of Education, University of Tokyo
[4] Hector Research Institute of Education Sciences and Psychology, University of Tübingen

Abstract

In psychology, researchers often predict a dependent variable (DV) consisting of multiple measurements (e.g., scale items measuring a concept). To analyze the data, researchers typically aggregate (sum/average) scores across items and use this as a DV. Alternatively, they may define the DV as a common factor using structural equation modeling. However, both approaches neglect the possibility that an independent variable (IV) may have different relationships to individual items. This variance in individual item slopes arises because items are randomly sampled from an infinite pool of items reflecting the construct that the scale purports to measure. Here, we offer a mixed-effects model called *random item slope regression*, which accounts for both similarities and differences of individual item associations. Critically, we argue that random item slope regression poses an alternative measurement model to common factor models prevalent in psychology. Unlike these models, the proposed model supposes no latent constructs and instead assumes that individual items have direct causal relationships with the IV. Such operationalization is especially useful when researchers want to assess a broad construct with heterogeneous items. Using mathematical proof and simulation, we demonstrate that random item slopes cause inflation of Type I error when not accounted for, particularly when the sample size (number of participants) is large. In real-world data ($n = 564$ participants) using commonly used surveys and two reaction time tasks, we demonstrate that random item slopes are present at problematic levels. We further demonstrate that common statistical indices are not sufficient to diagnose the presence of random item slopes.

Translational Abstract

In psychology, researchers often predict a dependent variable (DV) consisting of multiple measurements (e.g., nine scale items measuring conscientiousness). To analyze these data, researchers typically try and create a single value (e.g., collapse eight-item responses to one value indicating conscientiousness). Typically, researchers sum/average item scores, or use structural equation modeling, which posits a single hypothetical value representing a common element captured by the items. However, both approaches neglect the possibility that an independent variable (IV, e.g., birth order) may have different relationships to individual items. This variance results from the fact that items are randomly sampled from an infinite number of items that reflect the construct (i.e., all items that could measure conscientiousness). Using mathematical proof and simulation, we demonstrate that the chance of finding a falsely significant relationship between and IV and DV increases when using the standard approaches described above, particularly when the number of participants is large. In contrast, we offer a statistical model that accounts for similarities and differences

Ed Donnellan https://orcid.org/0000-0002-2739-7322
Kou Murayama https://orcid.org/0000-0003-2902-9600

Correspondence concerning this article should be addressed to Ed Donnellan, Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom, or Kou Murayama, Hector Research Institute of Education Sciences and Psychology, Europastrasse 6, University of Tübingen, 72072 Tübingen, Germany. Email: ed.donnellan@ucl.ac.uk or k.murayama@uni-tuebingen.de

of relationships between an IV and individual items, which does not have this problem. In real-world data ($n = 564$ participants) using commonly used surveys and two reaction time tasks, we demonstrate that though large variation in relationships between an IV and individual items measuring a concept are not very common, they are both present and problematic. Our proposed model should therefore be considered when analyzing data with multiitem DVs, especially when assessing a broad construct with heterogeneous items.

Think about a common situation in psychology in which one predicts scale scores from an independent variable (IV). For example, a researcher may be interested in predicting a personality trait from birth order. Normally the researcher measures the personality trait (e.g., extroversion) using multiple self-reported items on a survey, aggregating across them (i.e., computing sum or average scores), and conducts a regression analysis with the personality trait as the dependent variable (DV) and birth order as the IV. Of course, this aggregation approach is not limited to researchers using surveys. For example, a researcher may be interested in predicting participants' response times to particular types of stimuli from some IV (e.g., an experimental condition) and would use average response times across trials. Or a researcher may get participants to fill out a behavioral checklist (e.g., "in the last week have you X'd, have you Y'd?") to generate a total score of how prone to some behavioral trait a participant is, and predict this using some IV that is thought to be related. Throughout this paper, we use common nomenclature regarding survey data, talking about items (and item slopes), but equally this generalizes to other experimental stimuli (e.g., reaction time [RT] data). Our focus is not the specific type of items or stimuli used in research in particular fields (e.g., social psychology), but the research design typically used when investigating individual differences: where researchers aggregate across multiple measurements to create a DV.

The aggregation approach mentioned above (using birth order to predict a personality trait measured by multiple scale items) is the standard approach taken toward this kind of data. However, an implicit assumption of this approach is that there is a single fixed relationship between the IV (e.g., birth order) and the DV (e.g., aggregated trait scores). Therefore, this approach neglects the possibility that an IV may have different unique relationships to people's responses to each individual item in the personality scale (or RT to each individual stimuli in a RT task). This variation results from the fact that the items are sampled randomly from an infinite number of items (item pool, item population) that assess the same thing, e.g., some personality trait (Cronbach et al., 1963; Mõttus, 2016). We shall call these variations of the relationships *random item slopes*, as they represent random variation in the slopes between individual items and the IV.

The idea that items in a scale are randomly sampled from a population is not uncommon in test theories (Cronbach et al., 1963; De Boeck, 2008; McDonald, 1999; Shavelson & Webb, 1981) and is also an implicit assumption of measurement models based on common factors (latent variables), which are typically estimated by structural equation modeling (SEM; see Bollen & Lennox, 1991). Furthermore, the idea of random intercepts and slopes is

at the heart of mixed-effects modeling (Baayen et al., 2008; Barr et al., 2013; Judd et al., 2012; Murayama et al., 2014). To the best of our knowledge, however, while there are some isolated examples where these effects have been modeled (see Bayesian regression models predicting extrapair desire in Arslan et al., 2021; "category-specific effects" in Bürkner & Vuorre, 2019; equations in Appendix C in James et al., 2018; and multilevel SEM in Kessels et al., 2021), random item slopes in this type of research design have not been explicitly and systematically discussed in the literature. The aim of this article is to introduce the idea of incorporating random item slopes in this type of research design, which we shall call *random item slope regression*, and to discuss potential statistical and practical implications.

Random Item Slope Regression: A Mixed-Effects Model With Random Item Slopes

Let's first consider a common aggregation approach (i.e., an aggregated DV is predicted by an IV). Under the aggregation approach (hereafter called *aggregation regression*), where there is only one measurement of *y* per participant, a model predicting individual differences in *y* from an IV (*x*) can be represented by the equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1}$$

where *i* denotes participants ($i = 1, 2, …, I$). Parameter estimates and their standard errors (SEs) can be obtained using the ordinary least squares (OLS) method. Although we consider a case where there is a single IV for the purpose of simplicity, we can apply the following logic to cases where there are multiple IVs.

Random item slope regression treats items as crossed with participants in the data because all *I* participants respond to the same set of *J* items ($I \times J$ data points in total).[1] Mixed-effects modeling can effectively analyze nested or crossed data (for a review of advantages, e.g., statistical power, see Baayen et al., 2008). To conduct mixed-effects modeling, data are normally organized in so-called long format, in which every row represents a single item, resulting in data with $I \times J$ rows (see Figure 1 for an illustration of the transformation between wide format and long format data). First, consider a model with no IV, simply looking at the items that would typically be aggregated across:

$$y_{ij} = \beta_0 + u_{0i} + u_{0j} + \varepsilon_{ij}, \tag{2}$$

---

[1] This is analogous to situations where *I* participants' performance is rated by the same set of *K* raters (e.g., in multirater designs).

where u_{0i} represents a random intercept for each participant and u_{0j} represents a random intercept for each item. These random effects are independent and assumed to follow a normal distribution, $u_{0i} \sim N(0, \omega_{00})$, $u_{0j} \sim N(0, \tau_{00})$, $\varepsilon_{ij} \sim N(0, \sigma^2)$. The model is a two-factor random-effects analysis of variance and has been extensively discussed in the context of generalizability theory (Brennan, 2001). Generalizability theory quantifies the variance components of observed scores obtained in various factorial designs, which allows us to flexibly estimate reliability of test scores in a new study (Cronbach et al., 1963; Shavelson & Webb, 1981, 2006). In generalizability theory, test items or raters are some of the major sources of variance (in addition to the occasion of testing). An important assumption from this theory is that test items in the data are, like participants, randomly drawn from an infinite number of items in the population (called an "item universe"). In other words, the model assumes that all the items are *exchangeable*, meaning that the statistical conclusion is not dependent on the specific items used in the study. In standard statistical models (like the model in Equation 1), exchangeability is assumed for participants, and the results are guaranteed to generalize to the participant population. In this model, the same logic is applied to items as well: τ_{00} represents variance in this item population, as ω_{00} represents variance in the participant population. Note that Cronbach's α under this model is $\omega_{00}^2 / \left(\omega_{00}^2 + \dfrac{\sigma^2}{J} \right)$ (Shrout & Fleiss, 1979).

Although not normally considered in the literature of generalizability theory, we can add an IV assessed at the participant level (i.e., each participant has one value), x_i, to the model as follows:

$$y_{ij} = \beta_0 + u_{0i} + u_{0j} + \beta_1 x_i + \varepsilon_{ij}, \qquad (3)$$

where β_1 is a regression coefficient of x_i. Additionally, u_i and ε_{ij} are assumed to be independent from x_i. This model (hereafter *random intercepts regression* as it considers random intercepts for items and participants) is different from aggregation regression in Equation 1 and cannot be estimated using OLS. However, random intercepts regression in Equation 3 produces the identical parameter estimate and SE for β_1 with those from aggregation regression (and the same β_0 parameter estimate) using restricted maximum likelihood method (see Appendix A for mathematical proof). In other words, random intercepts regression in Equation 3 is mathematically equivalent to aggregation regression in Equation 1.

Critically, this model can be extended further by assuming that there is random variation of the slopes between items, for example, random item slopes. Specifically, the model now includes a population slope β_1 (i.e., a slope that we can obtain if we conducted the analysis for the entire population of participants and items) as well as individual slopes between items and the IV. The final model with random item slopes is as follows:

$$y_{ij} = \beta_0 + u_{0i} + u_{0j} + (\beta_1 + u_{1j}) x_i + \varepsilon_{ij}. \qquad (4)$$

The current manuscript calls this model *random item slope regression*. The deviation between the population and individual slopes is assumed to follow a normal distribution, $u_{1j} \sim N(0, \tau_{11})$, and there is a covariance between random item intercepts and random item slopes, $\mathrm{cov}(u_{0j}, u_{1j}) = \tau_{10}$. When there is a DV assessed by common multiple items and a single participant-level predictor, this is the full mixed-effects model with complete specification of all possible random effects. The model is easily estimated by any software of mixed-effects modeling that can specify crossed random effects, for example, lme4 in R (Bates et al., 2015; R Core Team, 2019), Hierarchical Linear Modeling (Raudenbush & Bryk, 2002), Mplus (Muthén & Muthén, 2017), and so on. By comparing Equations 4 and 3, and the fact that Equations 3 and 1 produce identical parameter estimates and SE (of β_1), aggregation regression in Equation 1 can be seen as a special case of random item slope regression in Equation 4, in which random item slopes are nonexistent.

Conceptual Ground: Random Item Slope Regression as an Alternative Measurement Model

The idea of random item slope regression is a natural extension of the standard regression model from a perspective of mixed-effects modeling. Although previous literature has repeatedly underscored the importance of incorporating possible random effects in various research designs (Baayen et al., 2008; Clark, 1973; Judd et al., 2012; Kajimura et al., 2023; Murayama et al., 2014; Usami & Murayama, 2018), the proposed model specification has rarely been discussed. One possible reason is that most of the previous work using crossed random effects focuses on factorial experiments and has not paid much attention to studies on individual differences.

Another possible reason is that, when researchers predict psychological constructs, they automatically suppose (either explicitly or implicitly) that there is a common factor underlying it; a predominant measurement framework to model people's responses to items on a scale. Using SEM, a common factor model with regression (hereafter called *common factor regression*) can be explicitly specified as depicted in Figure 2 (left). We also depict random item slope regression in Figure 2 (right) to clarify the differences between the models. In common factor regression, a common factor (i.e., a latent variable) representing the psychological construct of interest (e.g., "conscientiousness") is supposed to cause the items (observed variables), and the factor is predicted by the IV. From this perspective, item scores are the manifestation of the underlying construct with measurement errors. Interestingly, aggregation regression in Equation 1 can be considered as a special case of common factor regression. Specifically, aggregation regression is essentially equivalent to a constrained common factor regression model, with all factor loadings being 1 and error variances being equal (regression with a parallel factor model; McNeish & Wolf, 2020; see also Rose et al., 2019). Thus, aggregation regression can also be considered as supposing a common factor to explain an IV–DV relationship.[2]

A strength of common factor regression is that it can effectively separate measurement errors ($e_1, e_2, \ldots, e_5$ in Figure 2) from the construct of interest (separately from the residuals of the regression itself, which is denoted as d in Figure 2). This is clearly an

---

[2] As noted earlier, aggregation regression is also essentially equivalent to random intercepts regression in Equation 3. By implication, random item slope regression and common factor regression can be seen as extending aggregation regression in a different manner. Random item slope regression extends random intercepts regression, which is equivalent to aggregation regression, by incorporating random item slopes. Common factor regression extends a constrained common factor regression model, which is also equivalent to aggregation regression, by freeing the equality constraints on the factor loadings and error variances.

Figure 1

Wide Format Data With I Participants and J Items (With I Rows, as Used for Aggregation Regression, Left) Transformed to Long Format Data for Mixed-Effects Modeling (With I × J *Rows, Right)*

Note. IV = independent variable; DV = dependent variable.

advantage, but in light of the model we put forth in Equation 4, this strength also entails a cost. Specifically, common factor regression defines measurement errors as the components that are not common to all items. That is, if there are important item-specific components for a construct (i.e., the components that are specific to one or some, *but not all* items), these components are regarded as errors (e.g., e_1 in Figure 2) in common factor regression (McClure et al., 2021; McDonald, 1999, pp. xi, 485). Importantly, common factor regression normally considers that these item-specific components are uncorrelated with the IV (because they are considered measurement errors), and instead assumes that there is a single true effect between the IV and latent variable (β in Figure 2). Potential differential relationships between the IV and individual items that define the latent variable are supposed to be proportional to the factor loadings of the common factor ($\gamma_1, \gamma_2, \ldots, \gamma_5$ in Figure 2). In other words, individual item slopes are all attributed to the difference in measurement properties, rather than the substantive contents of the individual items. For example, if an item loads highly onto the common factor in comparison to other items, the item has more of the common component than the other items (i.e., it includes less measurement errors) and the model supposes that the item should have a stronger relationship with the IV. In short, common factor regression does not allow for the differential association between the IV and individual items beyond what is expected by factor loadings (i.e., higher factor loadings = stronger association).

On the other hand, random item slope regression understands the IV–DV relationship from a substantially different perspective. Specifically, while random item slope regression aims to evaluate a common relationship with the IV across all items, it also allows for heterogeneous items, evaluating different item-specific true associations with the IV (Figure 2). Heterogeneous items are

still seen as assessing the same construct, as the model assumes that individual items are randomly sampled from a common item population. As such, random item slope regression quantifies both similarities (represented by the regression coefficient β_1; this can be seen as the population mean of item slopes) and differences (represented by random item slope u_{1j}) of individual items, both of which explain the relationship with external variables. In other words, random item slope regression posits that a set of overlapping but heterogeneous items represent a construct as a whole.

Given this fundamental difference between common factor regression (and aggregation regression) and random item slope regression in how they conceptualize the relationship between individual items and the construct of interest, one can view random item slope regression as an alternative measurement model. At the very least, by adopting random item slope regression, researchers should be aware that this implicitly endorses a different way of conceptualizing measurement than if they adopt common factor regression. Common factor regression has dominated the analysis of multiitem constructs for decades in the psychological literature, and the model has indeed been useful in statistical analysis of psychological constructs. However, there is a good reason to believe that common factor regression is not the only correct measurement model to describe psychological constructs. In fact, common factor regression has been criticized in the literature, especially in terms of the critical assumption that there is a single latent construct that causally affects observed variable (Borsboom et al., 2003), with some researchers arguing that this assumption is restrictive and even unrealistic (Edwards & Bagozzi, 2000; van Bork et al., 2017). Recent literature suggests that we should take measurement models more seriously, underscoring the importance of understanding both strengths and

Figure 2

SEM Common Factor Model for a Scale With Five Items Predicted by an IV (Common Factor Regression; Left) Compared to Random Item Slope Regression (Right)
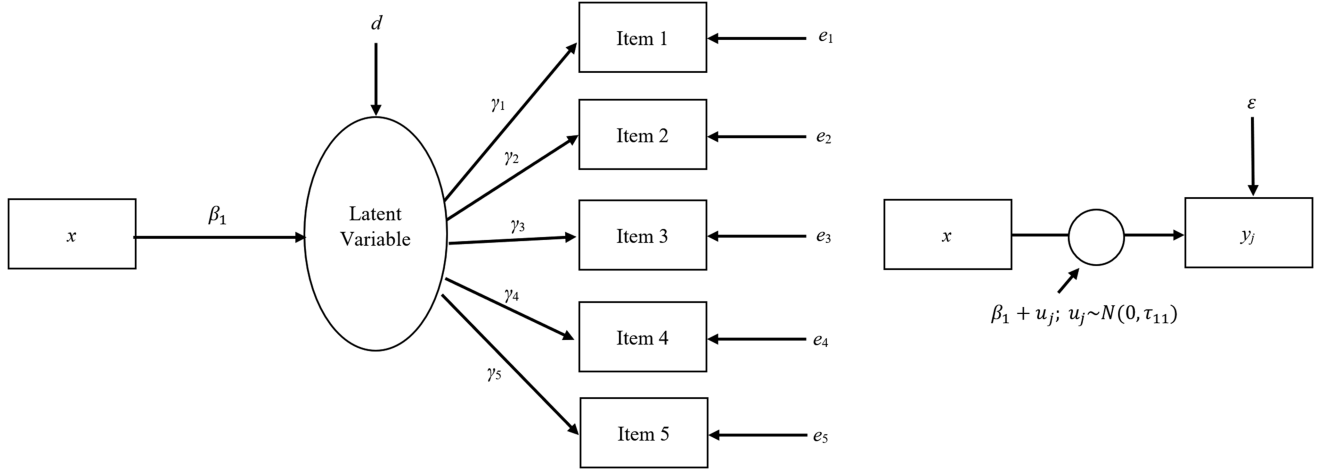
Note. j represents items ($j = 1, 2, …, 5$). For both diagrams, the participant dimension i is not explicitly expressed, which is common in SEM diagrams. Intercepts are all omitted for the purpose of simplicity. SEM = structural equation modeling; IV = independent variable.

weakness of common factor regression, and calling for consideration of potential alternative measurement models (Bollen & Lennox, 1991; Borsboom & Cramer, 2013; Fried, 2020; Rhemtulla et al., 2020). We believe that random item slope regression is a valuable option for consideration when there are substantial item-specific associations in the IV–DV relationship.

Therefore, the decision to use random item slope regression should primarily be based on substantial theory, that is, how researchers conceptualize the construct that they are investigating. If one is assessing a relatively narrow construct that could plausibly underscore each individual item, common factor regression based on SEM is more appropriate. On the other hand, if one is assessing a relatively large construct with a heterogeneous set of items, and thinks that item-specific elements are a constituent part of the construct, we believe random item slope regression is more appropriate. At the same time, common factor regression and random item slope regression are two distinct statistical models with different expected variance–covariance structures. As such, it is theoretically possible to distinguish between these two models purely with empirical data (e.g., by looking at model fit). We will evaluate the potential empirical distinguishability of these models in this paper.

Differences From Other Seemingly Related Models

Random item slope regression is not the only model that addresses item-specific effects, and it is worth considering differences from other seemingly related approaches, for example, generalizability theory and item response theory (IRT). Generalizability theory explicitly models random item effects to address the contribution of different sources of measurement variance (Brennan, 2001). However, to the best of our knowledge, generalizability theory only concerns random item *intercepts* without explicitly incorporating random item *slopes*. As shown above,

our proposed model (Equation 4) can be considered as an extension of a common model in generalizability theory (Equation 2) specifically incorporating item-specific slopes while also incorporating an IV. IRT also focuses on item-specific effects. However, like generalizability theory, the IRT literature primarily concerns item-specific properties within a measurement, not in relation to IVs.[3] For example, Rijmen et al. (2003) provide an analytic framework to understand IRT based on generalized mixed-effects modeling. However, item-specific parameters are defined only in relation to measurement (and they are fixed rather than random parameters), not slopes between items and IVs. One remarkable exception is De Boeck and Wilson (2004), who discussed a collection of IRT models based on generalized mixed-effects modeling including person-level predictors (see also De Boeck, 2008). However, a model with random item slopes (participant effects randomly varying among items) was not directly discussed.

In the context of SEM, our model could be analyzed with multilevel SEM with cross-classified design given that this model subsumes mixed-effects models (Muthén & Muthén, 2017). In fact, it is challenging to specify the proposed random item slope regression within the SEM framework without using multilevel cross-classified SEM. For example, Mehta and Neale (2005) describe a way to specify random-effects models based on standard SEM. However, the models they discuss only handle single random effects with *nested* data, while the critical feature of the proposed model is that we analyze data with a multiple-item scale with *crossed* random effects (random effects of participants crossed with random effects of items). Another SEM approach to deal with item-specific slopes is a bifactor regression

---

[3] As an exception, literature in differential item functioning (DIF) discusses cases in which item-specific parameters differ between externally defined groups. However, DIF effects are normally treated as fixed, not random.

model, in which external variables relate to both a single general factor as well as subgroup factors defined by a bifactor model (Martel et al., 2017; Wiernik et al., 2015). While the model allows for differential associations between the items and IVs, bifactor models are normally used to test a "hierarchical" structure of a broad construct, aiming to account for homogeneous "subgroups" across items. This perspective is critically different from random item slope regression where each individual item is supposed to cover different aspects of the target construct. In addition, the model has been criticized in terms of its susceptibility to overfitting and unstable parameter estimates, making the interpretation of these factors difficult (Watts et al., 2019).

Although not commonly observed in practice, perhaps the most relevant SEM specification is a model in which individual items are directly predicted by an IV (a version of seemingly unrelated regression [SUR]; Srivastava & Giles, 1987; Zellner, 1962). When the structure of residual variances/covariance matrix has compound symmetry (i.e., all variances are equal and all covariances are equal), the model is very similar to random item slope regression, and one can compute the average of Individual Item $\sim$ IV regression coefficients to statistically test the overall DV $\sim$ IV relationship. Alternatively, one can add a latent factor predicting individual items with a fixed path coefficient of 1 while constraining the covariance of residuals being zero (see Figure B1 in Appendix B). This model is mathematically equivalent to the former model, and the latent factor exactly represents the random participant intercept. In either case, however, this model misses the critical element of random item slope regression: It treats variations of regression coefficients as fixed rather than random (i.e., the model does not estimate the population variance of item-specific slopes). As a consequence, the model suffers from another critical issue which we will discuss below: the inflation of Type I error rates.[4]

Consequences of Ignoring Random Items Slopes in Statistical Inference: Type I Error Inflation

As discussed so far, if we do not incorporate random item slopes in the model, we implicitly ignore the possibility that an IV has different direct relationships with individual items. Another (somewhat less obvious) implication is that, by not accounting for random item slopes in the model, statistical results for β_1 cannot generalize to the item population, limiting our interpretation of the obtained findings. In other words, if we do not consider random item slopes in the model, the statistical results cannot guarantee that the same results (e.g., statistical significance) will hold if the same construct was assessed by a different set of items measuring the same construct/concept. This is especially problematic for many assessments in social and personality psychology, as these assessments normally purport to measure an abstract-level construct (e.g., "conscientiousness") and it is difficult to justify that a particular set of items in a scale uniquely and sufficiently assess the construct (Mõttus, 2016).

These conceptual issues also come with a serious statistical problem when ignoring random item slopes: the inflation of Type I error rates. Suppose that the true relationship between an IV (e.g., birth order) and a personality trait (e.g., extroversion) is exactly zero ($\beta_1 = 0$). A researcher may use eight items to measure the personality trait (e.g., using the Big Five Inventory; John & Srivastava, 1999). As these are sampled from the population of items that measure the trait, the average relationship between the IV and the eight items is unlikely to equal exactly zero due to random variation

affecting each individual slope (e.g., random item slopes). Therefore, for the given eight individual items, the estimated aggregated slope ($\hat{\beta}_1$: averaged slope between the IV and these individual items) will be nonzero even though the true mean is zero (β_1: average relationship between the IV and all items in the item population). This phenomenon is directly related to the number of items chosen to measure the construct: Increasing the number of items used means $\hat{\beta}_1$ will more closely approach β_1, that is, the influence of random item slopes decreases. In the example above, when more items are used the aggregated slope will be closer to zero.

However, even if random variation in item slopes is very small (meaning $\hat{\beta}_1$ more closely approaches β_1), as power increases (e.g., when the number of participants is larger), even small nonzero relationships become significant, resulting in a Type I error rate above the nominal level (e.g., 5%). In statistical terms, this increased Type I error is explained by underestimation of SE about $\hat{\beta}_1$ when random item slopes are present in the data but not accounted for in calculating $se(\hat{\beta}_1)$.[5] In fact, in the mixed-effects modeling literature, it is well known that standard errors are generally underestimated when one fails to specify random effects which are present in data (Barr et al., 2013; Clark, 1973; Usami & Murayama, 2018). We provide a mathematical proof of the underestimation of SE in Appendix A. Importantly, the mathematical proof (see Equation A19 in Appendix A) demonstrates that when random item slopes are present the degree of SE underestimation by aggregation regression in Equation 1 compared with random item slope regression in Equation 4 is a function of:

1. the number of participants (I): larger underestimation when there are more participants,
2. the number of items (J): larger underestimation when there are fewer items,
3. variation in participant intercepts (i.e., average scores of participants, ω_{00}): larger underestimation when there is smaller variation, and
4. variation in random residuals (σ^2): larger underestimation when there is larger residual error variance.[6]

The issue of a larger number of participants is particularly noteworthy because in research focusing on individual differences (e.g., investigating whether birth order predicts personality traits), researchers often collect data from a large number of participants with the good faith intention to increase statistical power. A large sample size is important to ensure high statistical power, but when the true effect does not exist, ironically it makes it more likely that researchers find false-positive effects. In theory, as the number of participants increases, Type I error rate asymptotically reaches

---

[4] In Appendix B, we run a simulation of this model following the simulation work described in this paper. We demonstrate that this model suffers from the same critical issue as aggregation regression and common factor regression.

[5] $\hat{\beta}_1$ is actually unbiased (see Appendix A for a proof). This is because the average of $\hat{\beta}_1$ converges to the true value if we repeatedly and randomly sample items to compute $\hat{\beta}_1$. Note also that even if $\beta_1 \neq 0$ (i.e., there is a nonzero true relationship), this underestimation of SE means that confidence intervals around $\hat{\beta}_1$ are narrower than they should be. This situation is discussed in the General Discussion section.

[6] The proof also shows that underestimation is larger when variance in the IV is larger (s_x^2). However, this is simply a scaling factor; if the variance of an IV is larger, random item slopes decrease.

100%. The issue of a smaller number of items is also worth mentioning as researchers often seek to reduce the number of items in their scales with the intention of reducing the burden on their participants, developing shorter scales to measure the same constructs (and often to facilitate data collection from a larger sample). Both decisions theoretically have big implications if the possibility of random item slopes is ignored.

A false-positive effect due to random item slopes may have a particular hallmark, namely a "small but statistically significant effect in a large sample." This is because, when the true effect is absent and random item slopes are present, the estimate of β_1 is likely to be small (due to the absence of the true effect) with underestimated SE. Such a small effect becomes falsely significant particularly when the sample size (number of participants) is large. Accordingly, false-positive effects due to random item slopes sometimes manifest as small statistically significant effects with a large sample.

It is worth noting that the inflation of Type I error rates is considered true only when the items are sampled from a population of possible items that could assess a construct, and where the effect is not present across this item population. However, if we only cared about the specific set of items used to collect data (e.g., testing the effect on scores of extroversion on a particular scale developed by a particular researcher), and did not care about generalizing the results to a wider set of items, then Type I error rate is accurately controlled for (i.e., these effects are not false positives). The problem is that, when random item slopes are not considered, a statistically significant effect does not guarantee that the results from one study would replicate when using different items assessing the same construct. Again, as we are normally interested in a broader construct assessed by scale items, and as such we expect the effect to generalize regardless of the items used (assuming item exchangeability), it may be difficult to substantially justify the omission of random item slopes when they are indeed present.

Illustrating the Effects of Random Item Slopes

To illustrate the properties and implications of the proposed random item slope regression, in the following, we try to address three issues through statistical simulations and analysis of real-world data. First, we demonstrate how aggregation regression and common factor regression inflate Type I error rates when random item slopes exist (at varying degrees) while random item slope regression shows no Type I error inflation. Second, we empirically evaluate the potential existence of such random item effects using real-world data. Existence of item-specific associations has been examined and discussed in the literature of personality ("trait nuances" in Mõttus et al., 2017; Mõttus & Rozgonjuk, 2021), psychometrics (Method of Correlated Vectors regarding g factor of intelligence, Jensen, 1998), and psychiatry (regarding p factor of psychopathology, Caspi et al., 2014; see also McClure et al., 2021), while recent literature on causal inference (VanderWeele, 2022) as well as network science has also underscored their importance (Borsboom & Cramer, 2013; Fried, 2015). The current paper empirically evaluates potential item-specific associations from the perspective of random item slope regression.

Third and finally, we also explore the potential effectiveness of commonly used statistical indices that could alert researchers to the presence of random item slopes: Cronbach's α (a measure of scale reliability) and the fit indices from SEM. As discussed above, common factor regression and random item slope regression are related but different statistical models. As such, it is theoretically possible to distinguish between them based on empirical data in order to select a more appropriate model. This could be achieved by using Cronbach's α or SEM fit indices generated when using common factor regression. In fact, the presence of random item slopes could reduce Cronbach's α because random item variance reduces the relative contribution of random participant intercepts, which is the major source in calculating this reliability index (see equation to compute Cronbach's α above). Also, the presence of random item slopes means that the association between an IV and individual items varies above and beyond what is expected from the differences in factor loadings. This deviation should manifest itself by decreasing the fit of the data to common factor regression. Poor fit of the SEM could therefore potentially alert a researcher to the presence of problematic random item slopes.

Simulation

Using simulation, we demonstrate that, when items are differentially predicted by the IV (i.e., random item slopes are present), aggregation and common factor regression inflate Type I error rates, while random item slope regression does not. Furthermore, we demonstrate that the inflation is more pronounced as a function of larger sample sizes, fewer items, and smaller participant intercepts.

Method

Models

Simulations were conducted using R 3.6.2 (R Core Team, 2019). We tested aggregation regression (a simple linear regression model where the DV is aggregated across multiple items for each participant, see Equation 1; implemented using *stats::lm*, R Core Team, 2019), which does not control for random item slopes. Crucially, we compared this model with random item slope regression (a linear mixed-effects model that controls for the random item slope in addition to participant and item intercepts, see Equation 4; implemented using *lme4::lmer*, Bates et al., 2015) and common factor regression (a structural equation model defining the DV as a latent variable, predicted by the IV and with each item as an indicator, Figure 2; implemented using *lavaan::sem*, Rosseel, 2012).

Simulation Parameters

The models were tested on simulated data sets. Simulated data were generated from the model in Equation 4. We systematically varied the number of participants ($I = 100, 200, 400, 1,000$) and the number of items ($J = 5, 10, 20$) per data set. Additionally, we systematically manipulated the variance of two random effects in the data generation model: random item slopes ($\tau_{11}^2 = 0$ [no slope], 0.01 [moderate], 0.09 [high]) and random participant intercepts ($\omega_{00}^2 = 0.36$ [low], 0.81 [high]). We simulated 1,000 data sets from each of the 72 unique parameter sets (with the same random seed used for each parameter set). For a given set of parameters, we generated a data set by randomly sampling from these parameters, where I participants had J items measuring the DV. We generated an IV value for each participant (x_i), randomly sampling a continuous value from a normal distribution ($M = 0$, $\sigma^2 = 1$). Random item intercepts (u_{0j}) are essentially means of each item

and would not affect any of the statistical inferences that we are interested in; as such, we fixed τ_{00}^2 to 0.25 (note these were sampled from a separate distribution than for τ_{11}^2, i.e., they were uncorrelated). Variance of the random errors in Equation 4 (σ^2) was fixed to 1.0. Crucially we set the fixed effect slope to 0 (also the intercept was also set to 0 for model interpretability, e.g., $\beta_0 = \beta_1 = 0$), meaning that the true model has no overall relationship between the IV and DV. Therefore any significant relationships found by the models in the simulated data sets result from the random item slopes and are Type I errors.

As the mathematical proof demonstrates that SE underestimation is a function of variance in random item slopes, participant intercepts, and residual error, for ease of interpretation we express variance of random item slopes and participant intercepts as a percentage of this relevant error variance (e.g., ω_{00}^2 as a percentage of relevant error variance $= \dfrac{\omega_{00}^2}{\omega_{00}^2 + \tau_{11}^2 + \sigma^2}$; see Table 1). The values for our simulation were initially arbitrarily chosen with two realistic constraints: (a) the prior assumption that random intercepts are likely to be larger than random item slopes, and (b) that these values would produce a reasonable range of Cronbach's α typically observed in empirical research (see Table 1). Though somewhat arbitrary, we demonstrate later that random item slopes of a magnitude tested in the simulation were present in real-world data (as a percentage of relevant error variance; see Real Data below).

Code/Data Availability

The code for the simulation and results from the simulation are available at https://osf.io/g7nbw/.

Results and Discussion

The proportion of significant results (i.e., Type I errors) over 1,000 simulations are shown in Figure 3 for each of the 72 parameter sets (4 [I] × 3 [J] × 2 [ω_{00}^2] × 3 [τ_{11}^2] = 72). This demonstrates that random item slopes cause Type I error inflation when they are not controlled for, with larger variation in random item slopes causing higher Type I error rates for aggregation regression, and no inflation for random item slope regression. The risk of Type I error increases for larger numbers of participants, but decreases to some extent for larger number of items. A larger variation in participant intercepts also slightly decreases the risk of Type I error. Thus, the Type I error rate is highest when the variance due to random item slopes represents more of the total variance across the model (see Table 1

for the size of random item slopes as a percentage of relevant error variance). However, we note that inflation is noticeable even when the random item slope is small (0.549% of relevant error variance) when I is large.

We calculated Cronbach's α for the different parameter sets (averaged across the four I, Table 1) and calculated the average fit statistics across parameter sets for common factor regression (Figure 4). Cronbach's α did not substantially decrease as random item slopes increased and were clearly much more of a function of random participant intercepts. In fact, even when random item slopes were present, Cronbach's α was at a level commonly thought to be satisfactory (>.70, see Bland & Altman, 1997). This means that Cronbach's α is not sensitive to the presence of random item slopes.

Fit indices for common factor regression showed poorer fit as random item slopes increased. However, when random item slopes are small, fit is often within the range of "good/modest fit" (e.g., $p > .05$, comparative fit index [CFI] > .90, and root-mean-square error of approximation [RMSEA] < .08; see Browne & Cudeck, 1993), suggesting that fit may not be sufficient to warn researchers of the potential presence of random item slopes.

Real Data

We have shown in the mathematical proof and simulations that random item slopes, and in particular the relative size of the variation in random item slopes in proportion to other error variance, causes inflation of Type I error. However, are there really differential relationships between items and IVs in real-world data from surveys and tasks?

Method

We collected some popular survey responses and data from two RT tasks to provide a range of real-world measures (see Table 2) with different numbers of items, reported scale reliability, and types of IV (all measures collected are reported in Table 3). Data were collected via Prolific Academic from August to September 2020 using an online survey implemented using jsPsych (de Leeuw, 2015). The study received ethical approval from the School of Psychology and Clinical Language Sciences Research Ethics Committee, University of Reading, UK.

Participants

$N = 579$ participants consented to take part; however, $n = 15$ were excluded for failing to complete the full set of surveys and RT tasks.

Table 1

Variance in Participant Intercepts and Item Slopes Expressed as a Proportion of Relevant Error Variance, and Mean Cronbach's α (Across I) for Different Participant Intercept and Item Slopes

Participant intercept		Item slope		Cronbach's α		
Variance	% Relevant error variance	Variance	% Relevant error variance	$J = 5$	$J = 10$	$J = 20$
0.36	26.471	0	0.000	.639	.780	.877
0.36	26.277	0.01	0.730	.636	.779	.876
0.36	24.828	0.09	6.207	.618	.766	.867
0.81	44.751	0	0.000	.800	.888	.941
0.81	44.505	0.01	0.549	.798	.888	.941
0.81	42.632	0.09	4.737	.785	.881	.936

Figure 3

Simulation Results Showing Type I Error and Convergence Failure of Aggregation Regression, Random Item Slope Regression and Common Factor Regression on Data With Varying Participant Intercepts and Item Slopes
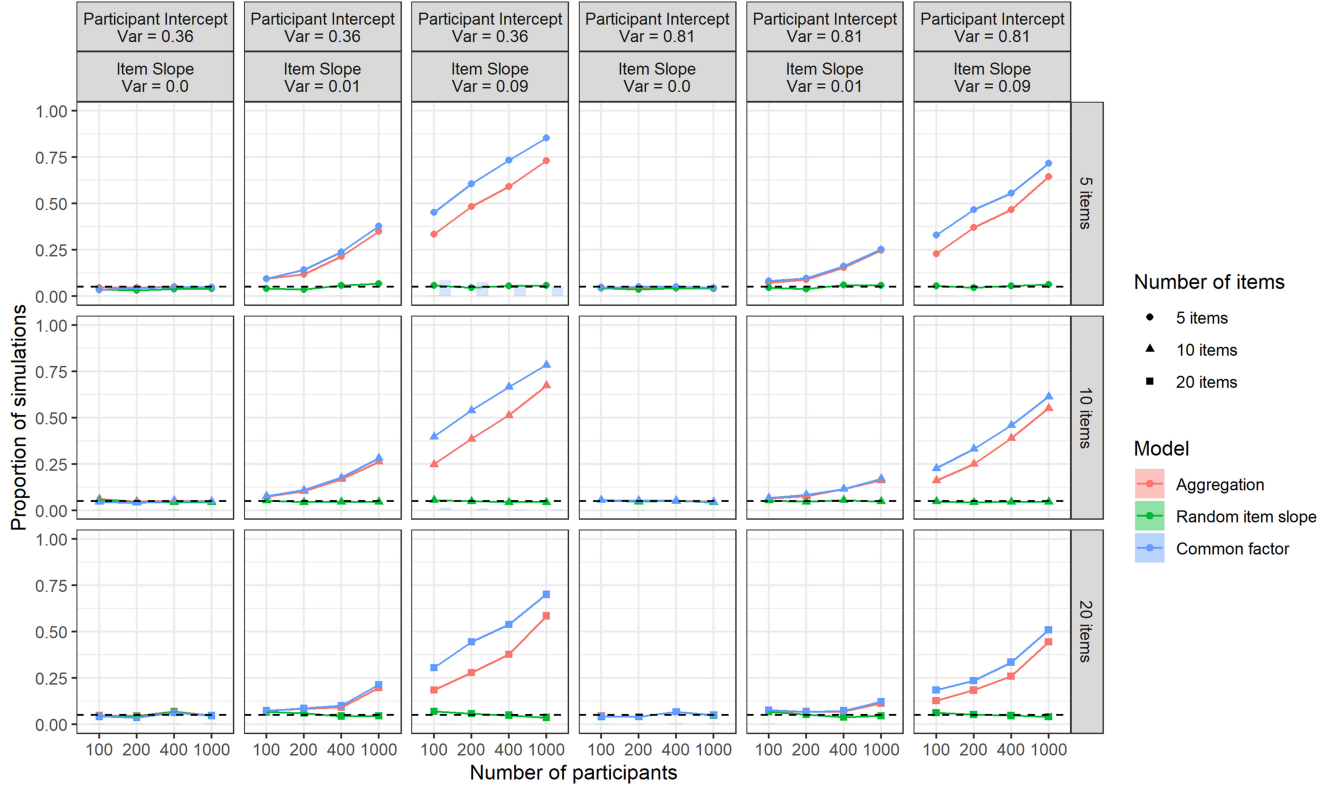
Note. Bars indicate the proportion of simulations in which models failed to converge. Data points indicate the proportion of simulations (in which the model converged) where the model found a significant relationship between IV and DV (e.g., Type I error rate), colored by model. IV = independent variable; DV = dependent variable. See the online article for the color version of the figure.

The final sample consisted of $n = 564$ participants (Female = 382, Male = 180, Prefer not to say = 2; Age: $M = 32.08$, $SD = 12.01$; Ethnicity: Asian = 45, Black = 22, Describe differently = 6, Mixed ethnicity = 27, Prefer not to say = 4, White = 460; Highest level of education: No formal education = 3, Secondary school/General Certificate of Secondary Education or equivalent = 38, College/Advanced levels or equivalent = 192, Undergraduate degree = 228, Postgraduate degree = 83, Doctorate = 18, Prefer not to say = 2).

Survey Data

The main purpose of our study was to analyze data using psychological measures that are commonly aggregated when used as DVs. Participants completed a number of commonly used psychological surveys (Table 2). All surveys in Table 2 were presented in the same order for all participants, that is, the Balanced Inventory of Desirable Reporting (Paulhus, 1991), the Big Five Inventory (John & Srivastava, 1999), the Short Index of Self-actualization (Jones & Crandall, 1986), and then the Epistemic Curiosity Scale (Litman, 2008). These were chosen as they are commonly used and have a range of reported α values. Additionally, most had facets or subscales, and one had a short version (e.g., the Balanced Inventory of Desirable Reporting; see Hart et al., 2015), again

with a range of reported α values. In addition to the measures reported in Table 3, we collected the Biographical Inventory of Creative Behaviours (Furnham & Bachtiar, 2008) to be analyzed as part of a separate project where item scores are binary variables that are treated as continuous when aggregated. Table 2 also shows the scale reliability calculated for each measure (Cronbach's α).

Reaction Time Experiments

Alongside survey data, we were keen to demonstrate that random item slope regression can be generalized to other kinds of psychological measures, for example, RT tasks where participant RTs to a number of stimuli (items) are aggregated. After completing the surveys, participants completed two Go/No-Go RT tasks. On Go trials, participants had to press the spacebar, and on No-Go trials, participants had to refrain from pressing the spacebar. One RT used simple two-dimensional shapes (Go: blue shapes; No-Go: orange shapes), while the other used pictures from the Open Affective Standardized Image Set (OASIS) database (Go: animals, No-Go: nonanimal objects—note all pictures were selected with a valence ≥5; see Kurdi et al., 2017). For both RTs, stimuli were separated with a blank screen with a fixation point for 1,000 ms, and

Figure 4
Averaged SEM Fit Statistics for Common Factor Regression Across 1,000 Simulations for Each Parameter
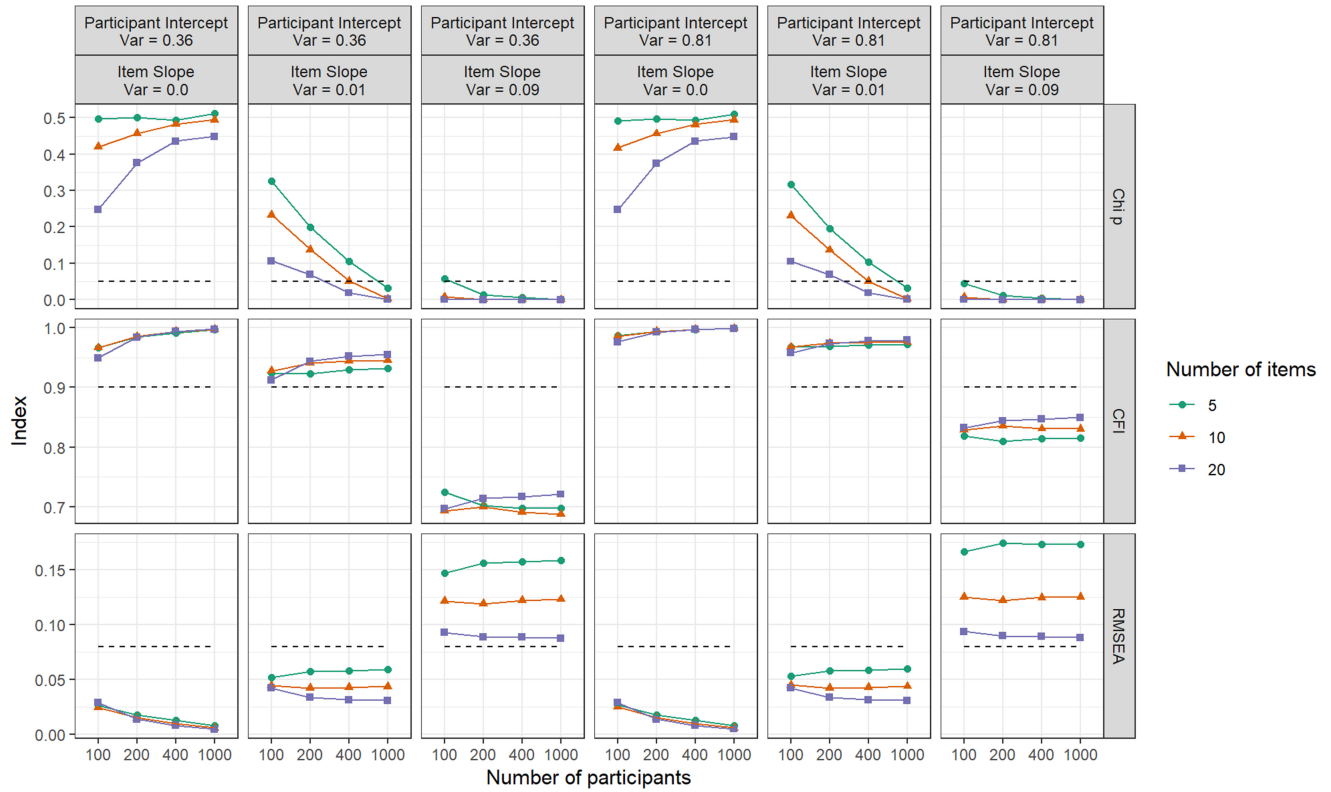
Note.   Dotted line indicates threshold for good fit (chi-square $p > .05$, CFI $> .90$, RMSEA $< .08$). SEM = structural equation modeling; CFI = comparative fit index; RMSEA = root-mean-square error of approximation. See the online article for the color version of the figure.

stimuli appeared for 2,000 ms. Participants had 10 practice trials in which they received feedback (i.e., if they pressed the spacebar in a No-Go trial they were informed that they should not have done). In the actual tasks, there were 50 stimuli shown in a preselected random order (all participants saw the same order, two blocks of 25 trials with a break in the middle, for both shapes and pictures), with 25 Go and 25 No-Go stimuli. We calculated the mean response time for each participant for both RTs (see Table 2).

Independent Variables

Demographics.   We tested the relationship between a range of IVs (Table 3). We focused on commonly used demographic information (e.g., age, gender, birth order, and number of siblings) as well as self-reported data on height.

Mood Induction.   In addition, we wanted to demonstrate that random item slope regression could generalize from observational studies of individual differences to experiments, so we conducted a mood induction, assigning participants to one of two experimental conditions: a neutral or positive mood induction. This consisted of viewing 17 images of scenery from the OASIS (Kurdi et al., 2017) for 4 s each. Images from the OASIS database were selected because they were open access and have previously been assigned scores for valence and arousal. We selected 17 positively valanced and high-arousal

images (e.g., beautiful lakes, sunsets, fireworks with a reported valence ≥ 6 out of 7, and arousal ≥ 4 out of 7 as reported in Kurdi et al., 2017) and 17 neutrally valanced low-arousal images (e.g., bare earth, concrete with a reported valence from 3.25 to 4.75, and arousal ≤ 4 out of 7).

Measures of participants valence and arousal were taken at baseline and post-mood induction—Valence: "How pleasant are you currently feeling?" Responses on a 9-point scale ranging from 1 (*extremely unpleasant*) to 9 (*extremely pleasant*); Arousal: "How aroused (i.e., feeling sleepy or feeling activated) are you currently feeling?" Responses on a 9-point scale ranging from 1 (*low arousal/sleepy*) to 9 (*high arousal/activated*). The mood induction was successful as those in the positive condition experienced feeling more pleasant after the induction relative to baseline and those in the neutral condition. A mixed-effects model (with condition [effect coded: *neutral* = -1, *positive* = 1], measurement time [*baseline* = -1, *post* = 1], and Condition × Measurement Time, with a random participant intercept) showed that the interaction was significant for valence, $b = 0.146$, $SE = 0.017$, $t(562) = 8.394$, $p < .001$. Participants in the positive condition reported feeling more pleasant after the mood manipulation—baseline: $M = 4.896$, $SD = 1.443$; post: $M = 5.408$, $SD = 1.401$; post hoc pairwise comparison corrected using Satterthwaite method, $t(562) = 11.088$, $p < .001$— while participants in the neutral condition did not—baseline: $M = 5.020$, $SD = 1.427$; post: $M = 4.948$, $SD = 1.329$; $t(562) = 1.391$,

Table 2

Descriptive Information for DVs Used (Prestandardization)

Subscale/facet	Items	n	Scale	Aggregated scores M (SD)	α
Balanced Inventory of Desirable Reporting					
Self-deception enhancement	20	564	7-point	3.98 (0.60)	.71
Impression management	20	564	7-point	4.09 (0.79)	.77
Balanced Inventory of Desirable Reporting (Short Version)					
Self-deception enhancement	8	564	7-point	3.75 (0.82)	.67
Impression management	8	564	7-point	4.04 (0.91)	.70
Big Five Inventory					
Agreeableness	9	564	5-point	3.68 (0.60)	.75
Conscientiousness	9	564	5-point	3.58 (0.65)	.81
Extraversion	8	564	5-point	2.92 (0.84)	.87
Neuroticism	8	564	5-point	3.3 (0.83)	.86
Openness	10	564	5-point	3.48 (0.63)	.80
Epistemic Curiosity Scale					
Deprivation-type	5	564	4-point	2.39 (0.67)	.84
Interest-type	5	564	4-point	2.88 (0.61)	.80
Short Index of Self-actualization					
Short index	15	564	7-point	3.68 (0.63)	.61
Go/No-Go Tasks					
Shapes RT	25	564	0–2,000 ms	457.89 (85.99) [a]	.94
Pictures RT	25	564	0–2,000 ms	578.84 (103.37) [a]	.94

Note. DV = dependent variable. RT = reaction time.
[a] Aggregated RT across trials where participants responded only.

$p = .506$. However, for a mixed-effects model with the same specification but with arousal as the DV, the interaction was not significant—$b = 0.035$, $SE = 0.192$, $t(562) = 1.794$, $p = .073$. Participants in the positive condition reported feeling marginally more aroused/activated—baseline: $M = 4.228$, $SD = 1.725$; post: $M = 4.430$, $SD = 1.669$; $t(562) = 3.971$, $p < .001$—whereas participants in the negative condition did not—baseline: $M = 4.097$, $SD = 1.654$; post: $M = 4.161$, $SD = 1.558$; $t(562) = 1.121$, $p = .677$.

Coding and Standardization

All IVs were effect coded (if dichotomous) or standardized ($M = 0$, $SD = 1$). All DVs were also standardized to allow us to compare the size of random item slopes (and fixed effects) across measures and models; specifically, for each DV, we mean-centered and scaled a vector containing all item responses (e.g., if the DV consisted of five items for 100 participants, this created a vector of 500 values, which were then scaled). To create standardized aggregated scores, these scaled values were then aggregated. This process of standardization of DVs does not affect any test statistics for random item slope regression.

Analysis

Aggregation, common factor, and random item slope regression were implemented as in the simulation. In addition, random intercepts regression was also implemented using lme4:lmer (Bates et al., 2015).

Data Availability

The data analyzed here and scripts are available at https://osf.io/g7nbw/.

Table 3

Descriptive Information for IVs (Prestandardization)

Variable	Type	Scaled/coded	n	M (SD)	Categories
Birth order	Continuous	Scaled	564	1.82 (1.11) [a]	
Total siblings	Continuous	Scaled	564	1.73 (1.43)	
Height	Continuous	Scaled	563 [b]	169.26 (9.67)	
Age	Continuous	Scaled	564	32.08 (12.01)	
Condition	Dichotomous	Effect coded	564		Neutral, $n = 248$
					Positive, $n = 316$
Gender	Dichotomous	Effect coded	562 [c]		Female, $n = 382$
					Male, $n = 180$

Note. IV = independent variable.
[a] 1 = *first born*, 2 = *second born*, and so on. [b] One outlier removed for reporting a height of 60 cm. [c] Variable dichotomized (male/female), removed $n = 2$ that gave different responses.

Results and Discussion

We investigated models with every combination of IV and DV (single predictor models), using random item slope regression in Equation 4 on the full sample. Table 4 shows the size of the random item slope estimated by random item slope regression, expressed as a percentage of relevant error variance for each DV $\sim$ IV relationship.

Table 4 shows that we did not observe random item slopes at the large sizes in our simulation (e.g., 4.737%–6.207%). However, random item slopes were estimated to be present in real-world data using commonly used scales. Furthermore, for some DV $\sim$ IV relationships, variance in random item slopes was estimated to be higher than the percentage of the relevant error variance that is known from our simulations to cause Type I error inflation (0.549%). Note that the percentage of the relevant error variance smaller than 0.549% is not a "cutoff"; this is an arbitrary value that we used in the simulation and smaller random item slope variance still produces inflated Type I error rates (as mathematically demonstrated in Appendix A). We chose 0.549% simply because the consequences were well quantified in our simulation.

The two largest item slopes estimated from the data were for Self-actualization $\sim$ Age, and Interest-Type Curiosity $\sim$ Age. To understand how different model specification changes the results, Table 5 shows the results from the three models represented by Equations 1, 3, and 4 for these two relationships, along with the relationship between Agreeableness and Age (with a random item slope estimated to be near zero) for comparison.

Table 5 shows that for the larger item slopes (Interest-Type Curiosity $\sim$ Age and Self-actualization $\sim$ Age), error variance that is considered residual by models that ignore random item slopes (Equations 1 and 3) could be due to random item slopes, which, if modeled (Equation 4), affects the estimate of SE around the fixed effect. This, in turn, affects whether a relationship is found to be significant or not. When there is a negligible random item slope (Agreeableness $\sim$ Age), SE is effectively identical (after rounding) for Equations 1, 3, and 4, that is, the underestimation is small, with only a marginal affect on t and p. Note that $\hat{\beta}_1$ does not differ across the models, with the exception of some differences for common factor regression. As discussed earlier, aggregation regression (Equation 1) is mathematically equivalent to random intercepts regression (Equation 3), and fixed effects of these models are not biased (see Footnote 5). Therefore, we expected equivalent fixed effects among aggregation, random intercepts, and random item slope regression. The difference between these models and common factor regression is also expected because, as discussed earlier, aggregation regression is equivalent to a *constrained* common factor regression model, not the commonly used *unconstrained* common factor regression model tested here. As illustrated by McNeish and Wolf (2020), these different models could produce nonnegligible differences in parameter estimates. For interested readers, we also included the figure comparing the estimated slopes of each item based on random item slope and common factor regression in Figure S1 online supplemental materials.

Figure 5 shows the relationship between the aggregated scores from the Short Index of Self-actualization as predicted by age (the relationship with the second largest random item slope

estimate). Figure 6 (bottom) then shows the relationship of each item from the Index with age (the red dotted line indicates the overall relationship from Figure 5). This shows the large variance in the slopes across different items on the scale; while around seven out of 15 of the items show a smaller (or indeed opposite) slope to the aggregated slope with two items contributing large negative slopes. Aggregation regression found a significant relationship between age and self-actualization ($b = -0.062$, $SE = 0.015$, $t = -4.182$, $p < .001$) while random item slope regression did not ($b = -0.062$, $SE = 0.032$, $t = -1.957$, $p = .066$) due to a smaller $se(\hat{\beta}_1)$ for aggregation regression. In addition, common factor regression also found a significant relationship ($b = -0.129$, $SE = 0.023$, $z = -5.547$, $p < .001$). The self-actualization scale had poor scale reliability (Cronbach's $\alpha = .607$). Furthermore, common factor regression showed poor model fit ($\chi^2 = 762.168$, $df = 104$, $p < .001$, CFI $= 0.447$, RMSEA $= 0.106$).

A strength of random item slope regression is that it can allow us to identify when items differently relate to a predictor. Figure 5 shows that items measuring self-actualization have different direct relationships with age. For example, two items showing strong negative relationships with age both relate to fears (Item 8: I fear failure" and Item 14: "I am bothered by fears of being inadequate"). Other items do not mention fear at all, so we may speculate on why age is related to the fear element of self-actualization, while having less of an (or an opposite) effect on other elements. If we can quantify these item-specific properties, we can even incorporate them in the model to see whether item-specific characteristics can indeed predict the differential regression slopes (see also General Discussion). Additionally, there was negative covariance between random item intercepts and slopes (see Table 5). Note that mean age in the sample was 32.08 ($SD = 12.01$; see Table 3) and age was standardized before the analysis. The results mean that items with larger negative slopes (e.g., Items 8 and 14) also had higher intercepts, suggesting that items that participants aged 32 tended to more strongly agree with showed stronger negative relationships with age (Figure 5). One potential interpretation is that items relating to fear were more highly endorsed than other items, but this difference disappears as people age. An alternative interpretation is that this negative covariance reflects a measurement artifact, for example, items with higher intercepts have more room to decrease with age. We feel the alternative interpretation is unlikely because this pattern is only observed for this particular IV–DV pair among many different combinations, and results for interest-type curiosity and age showed the opposite (positive) covariance (Table 5). In any case, the relationship between intercept and slopes should also provide useful information to interpret findings.

The next largest item slope was for the Interest-Type Curiosity subscale from the Epistemic Curiosity scale. Figure 6 shows the relationship between aggregated scores and age and the by-item slopes. Item 9 has a stronger negative relationship with age than all other items, causing the aggregated slope to be more negative than 4/5 of the items. While aggregation and common factor regression found a significant relationship between age and interest-type curiosity (aggregation: $b = -0.120$, $SE = 0.030$, $t = -4.034$, $p < .001$; common factor: $b = -0.089$, $SE = 0.027$, $z = -3.283$, $p = .001$), random item slope regression did not ($b = -0.120$, $SE = 0.054$, $t = -2.216$, $p = .063$). The Interest-Type

Table 4

Random Item Slopes (Expressed as a Percentage of Relevant Error Variance, Estimated by Random Item Slope Regression)

Measures/facets	Random item slope					
	Birth order	Age	Condition	Gender	Height	Total siblings
Balanced Inventory of Desirable Reporting						
Self-deception enhancement	0.017%	0.545%[a]	0.071%	**0.836%** [a]	0.228%[a]	0.192%[a]
Impression management	0.095%[a]	**1.060%** [a]	0.015%	**0.608%** [a]	0.187%[a]	0.109%[a]
Balanced Inventory of Desirable Reporting (short version)						
Self-deception enhancement	0.030%	0.070%	0.008%	**0.942%** [a]	0.257%	0.089%
Impression management	0.073%	0.276%[a]	0.120%	**0.952%** [a]	0.467%[a]	0.046%
Big Five Inventory						
Agreeableness	0.021%	0.002%	0.069%	0.321%[a]	0.125%	0.013%
Conscientiousness	0.018%	0.306%[a]	0.000%	0.167%	0.025%	0.088%
Extraversion	0.043%	0.296%[a]	0.013%	0.388%[a]	0.225%[a]	NA
Neuroticism	0.002%	0.268%[a]	0.046%	**1.00%** [a]	0.488%[a]	0.007%
Openness	0.054%	0.492%[a]	0.038%	**0.702%** [a]	0.203%[a]	0.017%
Epistemic Curiosity Scale						
Deprivation-type	0.050%	0.309%[a]	NA	0.301%[a]	0.230%	0.006%
Interest-type	0.058%	**1.120%** [a]	0.000%	0.220%[a]	0.090%	0.043%
Short Index of Self-actualization						
Short index	0.007%	**1.430%** [a]	0.003%	**0.781%** [a]	0.181%[a]	0.000%
Go/No-Go Task						
Shapes RT	0.070%	0.065%	0.000%	0.090%[a]	0.035%[a]	0.049%
Pictures RT	0.014%	0.135%[a]	0.001%	0.022%	0.116%[a]	0.001%

Note. Bold text denotes when a random item slope was above 0.549% of relevant error variance (i.e., the smallest size slope tested in simulations that showed Type I inflation). NA = not applicable; RT = reaction time.

[a] Indicates where the random item slope was significant, as tested by a likelihood ratio test comparing the random intercepts and random slope model.

Curiosity scale showed acceptable scale reliability (Cronbach's $\alpha = .797$). Furthermore, common factor regression showed good model fit according to CFI and only marginally poor fit on RMSEA ($\chi^2 = 47.442$, $df = 9$, $p < .001$, CFI = 0.956, RMSEA = 0.087).

Again, by inspecting Figure 6, we can see that age has a stronger effect on one item (Item 5: "I enjoy discussing abstract concepts") relating to interest-type curiosity. This is the only item to mention abstract concepts, and the others relate to acquisition of new or unfamiliar information. Age could therefore be particularly negatively related to the element of interest-type curiosity concerning interest in abstract concepts, in comparison to the elements captured by the remaining items (e.g., enjoyment of learning about new information). Positive covariance between random item intercepts and slopes (see Table 5) suggests that while the abstract concepts item had a stronger negative relationship with age, participants aged 32 also tended to agree less with this item when compared to other items.

Table 5

Regression Model Results for Three Example Relationships (Agreeableness ~ Age, Interest-Type Curiosity ~ Age, Self-actualization ~ Age)

| Regression model	Fixed effects				Random effects				
	$\hat{\beta}_1$	SE	t	p	ω^2_{00}	τ^2_{00}	τ^2_{11}	σ^2	τ_{10}
Agreeableness ~ Age									
Aggregation in Equation 1	0.034	0.023	1.507	.132				0.536	
Random intercepts in Equation 3	0.034	0.023	1.507	.132	0.214	0.142		0.660	
Random item slope in Equation 4	0.034	0.023	1.504	.136	0.214	0.142	<0.001 [a]	0.660	>−0.001 [b] (−0.264)
Common factor	0.023	0.020	1.125	.261					
Interest-Type Curiosity ~ Age									
Aggregation in Equation 1	−0.120	0.030	−4.034	<.001				0.705	
Random intercepts in Equation 3	−0.120	0.030	−4.034	<.001	0.394	0.093		0.520	
Random item slope in Equation 4	−0.120	0.054	−2.216	.063	0.396	0.093	0.010	0.509	0.026 (0.842)
Common factor	−0.089	0.027	−3.283	.001					
Self-actualization ~ Age									
Aggregation in Equation 1	−0.062	0.015	−4.182	<.001				0.351	
Random intercepts in Equation 3	−0.062	0.015	−4.182	<.001	0.073	0.188		0.747	
Random item slope in Equation 4	−0.062	0.032	1.957	.066	0.074	0.188	0.012	0.736	−0.029 (−0.617)
Common factor	−0.129	0.023	−5.547	<.001					

Note. τ_{10} is reported as covariance and as the standardized correlation coefficient in parentheses. For common factor regression, we provide z values instead of t values.

[a] 0.00002. [b] 0.0004.

Figure 5

Self-Actualization Scores Predicted by Age
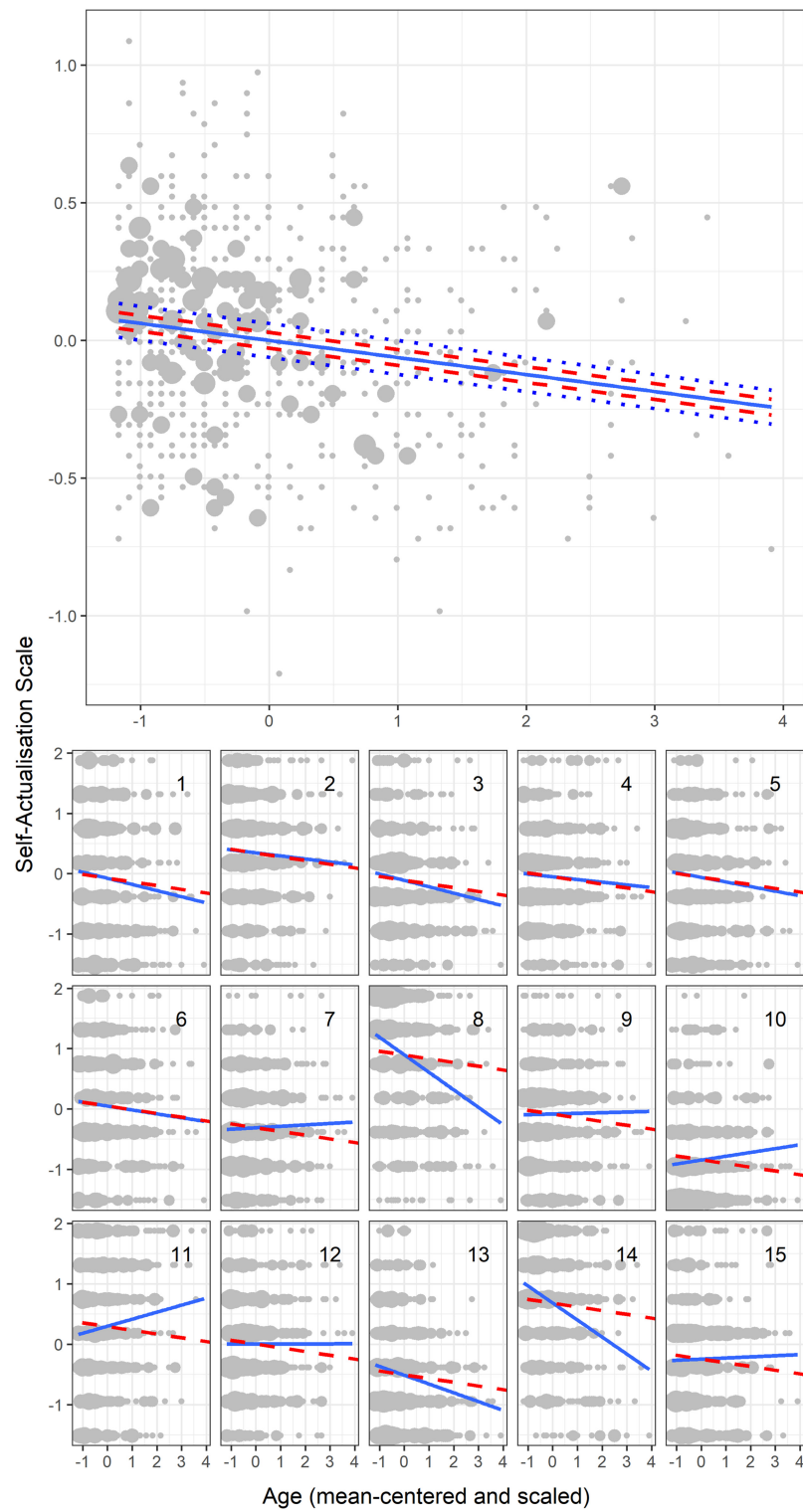
Note. Top (aggregated score): Blue line represents the slope, red dashed line represents the 95% confidence interval from aggregation regression in Equation 1, and blue dotted line represents the 95% confidence interval from random item slope regression in Equation 4. Bottom (by-item score): Red dashed line indicates the slope of the Aggregated Scores ~ Age (as in Figure 5), with intercept adjusted for comparability. See the online article for the color version of the figure.

Figure 6

Interest-Type Curiosity Scores Predicted by Age
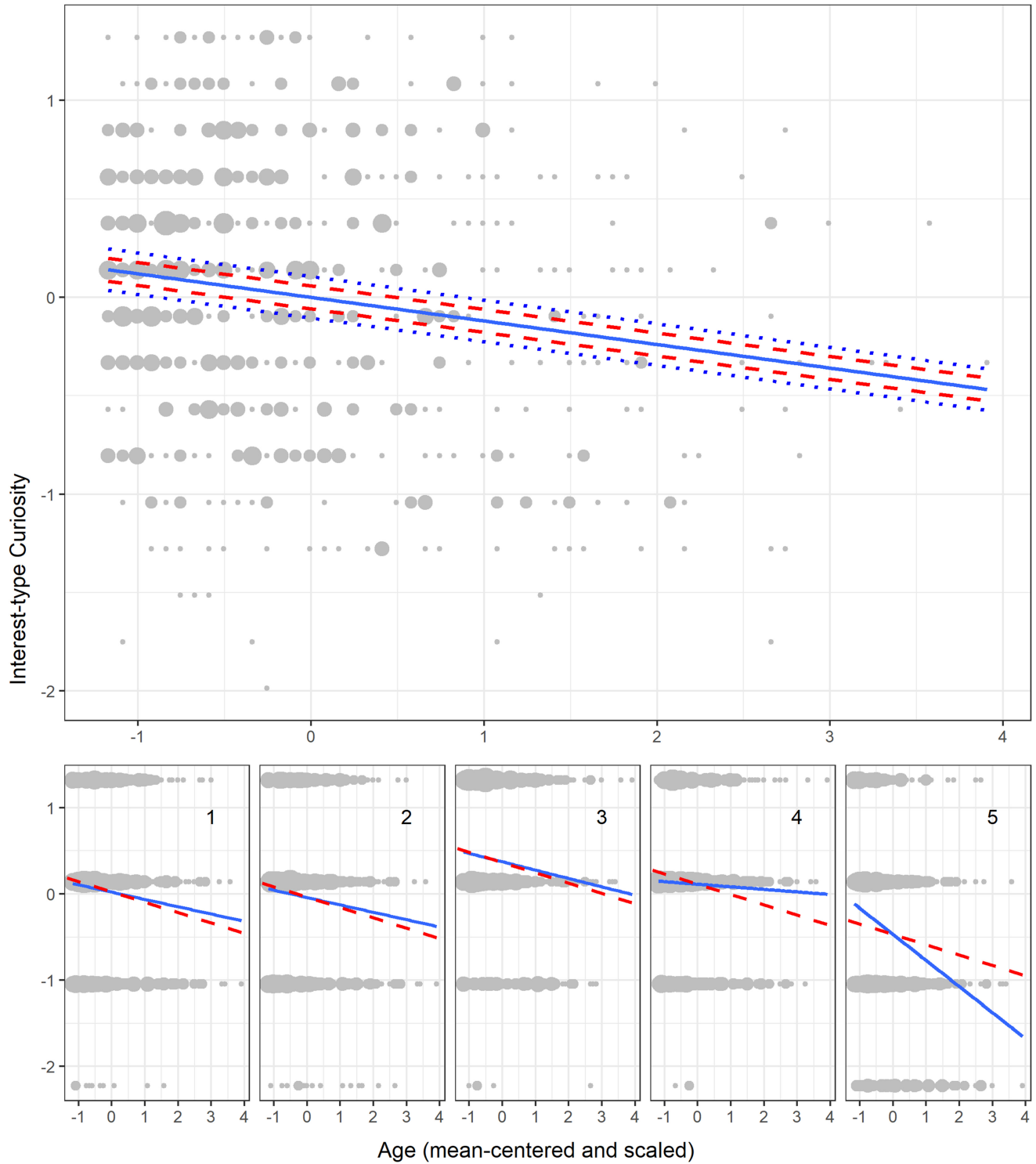
Note. Top (aggregated score): Blue line represents the slope, red dashed line represents the 95% confidence interval from aggregation regression in Equation 1, and blue dotted line represents the 95% confidence interval from random item slope regression in Equation 4. Bottom (by-item scores): Red dashed line indicates the slope of the Aggregated Scores ~ Age (as in Figure 5), with intercept adjusted for comparability. See the online article for the color version of the figure.

Without knowing the true model, we cannot say that the significant result found by aggregation regression for Self-actualization ∼ Age is a Type I error. However, we can still demonstrate that as the sample size increases (larger number of participants), the chance of finding a significant result with aggregation regression increases relative to the chance of finding a significant result with random item slope regression for the two cases discussed above (where random item slope regression estimated large variation in random item slopes). Figure 7 shows the absolute t value for the estimated slope of Agreeableness ∼ Age, Interest-Type Curiosity ∼ Age, and Self-actualization ∼ Age, for 100, 200, 400, and 564 participants (for 100, 200, and 400 participants, the data were averaged over 1,000 resamples). This shows that for the two relationships with high random item slopes (Figures 5 and 6), as the sample increases, these show the characteristic pattern observed in the simulations for random item slopes (i.e., an increasingly higher likelihood of finding a significant relationship when the sample size increases when random item slopes are not controlled for; see Figure 2). Comparatively, when there are no random item slopes (for Agreeableness ∼ Age) the t values remain the same for aggregation regression, random intercepts regression, and random item slope regression, that is, the interaction between sample size and model misspecification does not increase the chance of finding a significant result (simply sample size, suggesting simply a small effect of age on Agreeableness). These results indicate that ignoring random item slopes when they are present becomes particularly problematic when sample size is larger.

General Discussion

The current paper proposes random item slope regression as an alternative to commonly used approaches predicting the DV, which consists of a multiitem scale (aggregation regression, common factor regression). We discussed the conceptual and theoretical implications of the model, especially from the perspective of measurement models. Then to evaluate the statistical and practical implications of the proposed model, we conducted the statistical simulations and empirical study. The results demonstrate three points. First, when random item slopes are not controlled for, they can cause Type I error inflation (as predicted by the mathematical derivation in Appendix A). This is the case even when the magnitude of random item slope variance is small (0.549% of relevant error variance), if sample size (number of participants) is large. Second, in real-world data, using commonly used survey measures, we found evidence that random item slopes may exist at levels known from simulations to cause Type I error inflation. Finally, we have shown that measures of scale reliability and SEM fit statistics are not sufficient to warn of the presence of random item slopes. It should be noted that, while our empirical example utilized data from social and personality psychology, the practical risk of Type I error inflation that can be averted by using random item slope regression is not limited to just these subdisciplines of psychology; any DV that consists of multiple measurements sampled from a pool of possible items could be susceptible, regardless of discipline (see Judd et al., 2012; Kajimura et al., 2023; Westfall et al., 2017; Wolsiefer et al., 2017).

Figure 7

Absolute t *Values Estimated by Aggregation Regression in Equation 1 and Random Item Slope Regression in Equation 4 for Example Relationships*
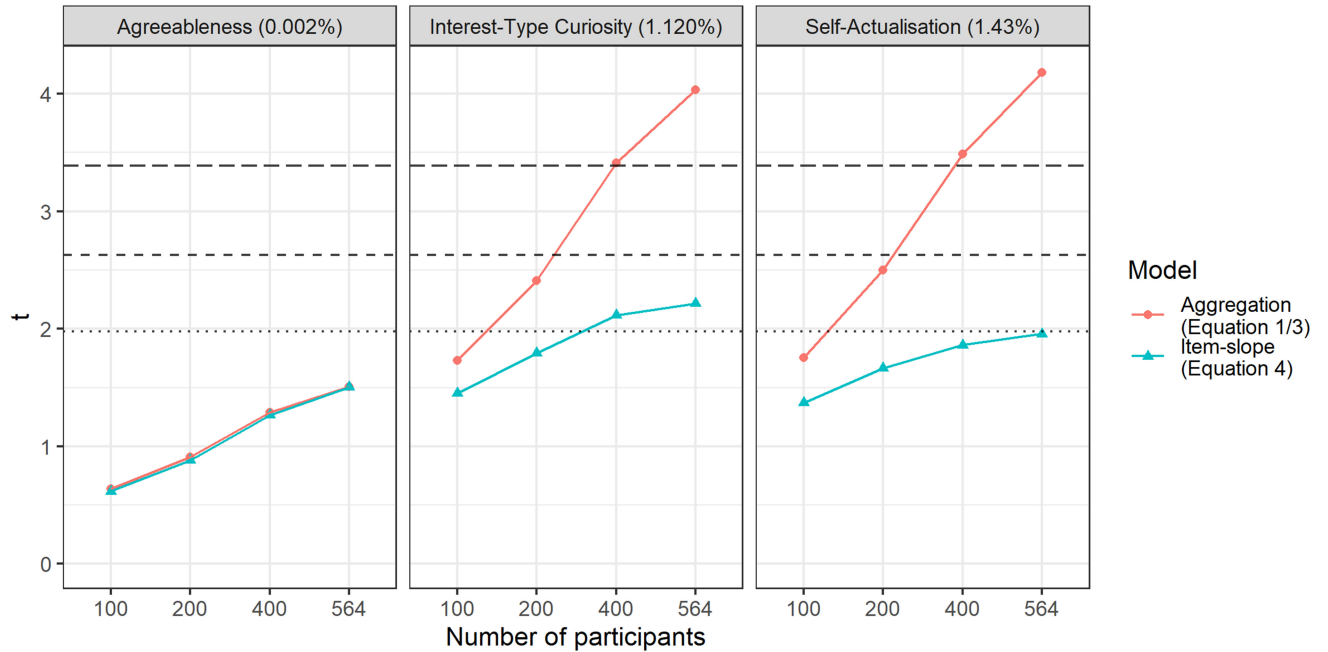
Note. Facets are labeled with the size of the item slope as a percentage of relevant error variance in parentheses. Horizontal lines indicate significance where $\alpha = .05$ ($z = 1.96$: dotted line), $\alpha = 0.01$ ($z = 2.326$: short-dashed line), and $\alpha = 0.001$ ($z = 3.291$: longer-dashed line). See the online article for the color version of the figure.

One may ask why our proposed model is necessary given that, in many cases, random item slope regression and the standard analytical approach based on latent variables (e.g., common factor regression) show similar regression coefficients (see Table 5), and that one could also use seemingly related existing models within the latent variable framework (e.g., SUR). In response, we cannot stress more strongly that switching to the proposed model has much broader theoretical and practical implications than those that arise from simply switching to existing models. Crucially, random item slope regression represents a new alternative measurement model that does not posit latent variables. This provides a qualitatively different way of theoretically understanding psychological constructs from the standard analytical approach. Even if models show similar regression coefficients, random item slope regression is different in that we do not interpret the coefficients as the magnitude of the effects on a single latent construct. The model rather supposes that there is no such latent causal entity and instead posits that a psychological construct is an emergent property from individual items (as in network models, discussed below). Thus, this focuses on the direct causal effects from IVs on individual items, as opposed to indirect effects entirely through a single hypothetical latent variable. In practice, item-specific effects are not uncommon in psychological measurements (McClure et al., 2021; VanderWeele, 2022) and should not be ignored. Furthermore, no existing latent variable model considers generalizability to the item population and, as we have demonstrated in the current paper, therefore sometimes commits to considerable inflated Type I error rates when making this natural inference (even using SUR, see Appendix B). The proposed model, on the other hand, is immune to such errors as it naturally incorporates random variation of items.

The Ubiquity of Random Item Slopes

In the real-world data that we examined, random item slopes exist, and in some cases exist to an extent that would increase Type I error rates to a considerable degree. We should be aware of this, especially when dealing with small but significant relationships found across a large sample of participants. However, the good news is that large random item slopes were not ubiquitous in the current data. Although we only tested a small set of scales, these results suggest that many findings based on aggregation regression or common factor regression may not suffer from this issue. If there are no random item slopes, we could safely return to aggregation or common factor models for reasons of parsimony. However, we stress that from a practical perspective, nothing prevents us at this point in time (or far less prevents us) from testing for the presence of random item slopes. When item-specific effects are important or expected (see discussion below), it is worth examining data to see if random item slopes exist in the first place before proceeding to common factor models.

Four things should be noted. First, even when random item slope variance is statistically significant (i.e., model comparison shows that random item slope regression is a significantly better fit than random intercept regression for some data), this does not necessarily mean that random item slope regression is the true model for the data. As noted earlier, different associations may be accounted for by the differences in factor loadings in common factor regression. As discussed in the introduction, the final decision should be based on a substantive theory as well as empirical data. Second, while

large random item slopes may not be ubiquitous, we should remain vigilant to the reality that even small random item slopes can have a substantial impact on Type I error rates. In our real-world data, few measures showed truly negligible estimates of random item slopes. While the impact of these slopes is greater when combined with a larger number of participants, our simulations demonstrated that small random item slopes showed some Type I error inflation with as few as 100 participants when not controlled for. Therefore while SE underestimation undoubtedly gets worse for large samples of participants, they should not be ignored even by researchers who are collecting from smaller samples. Third, while random item slope regression may be particularly advantageous to draw a correct inferential conclusion when N is large and effects are small, in these situations, regardless of the modeling approach, researchers should of course rely on effect sizes and other information to identify practically significant effects over a reliance on statistical significance. The proposed method should be seen as having a complementary advantage with effect size approach to fight against spurious statistically significant effects. Finally, in the literature of mixed-effects modeling, it is well known that SE is underestimated when cluster size is small (McNeish & Stapleton, 2016). In such cases, it is recommended that researchers use correction methods such as Kenward–Roger correction (Kenward & Roger, 1997), which is implementable using R package pbkrtest (Halekoh & Højsgaard, 2014). This recommendation also applies to the proposed model.

The current manuscript focuses on Type I error rate but it is also worth noting the effects of random item slopes on statistical power. In Appendix C, we showed additional simulation results to examine the effect of random item slopes on statistical power in the presence of different sized fixed effects slopes for aggregation and random item slope regression (Appendix C). Generally, when random item slopes are present and aggregation regression is used (red line), statistical power is higher than when applying random item slope regression (green line) due to SE underestimation. However, this does not mean that aggregation regression is better as aggregation regression gains high statistical power by paying an important (and unacceptable) price of inflated Type I error rates, the extent of which we cannot calibrate from the data. In terms of the effects of various factors on statistical power of random item slope regression (i.e., comparison within the green lines), generally, power increases as the number of participants and items increases, which is consistent with our common intuition. In addition, larger magnitudes of random item slope variance generally decrease statistical power because increased random item slope variance adds "noise" to the data, increasing sampling error.

When Should We Use Random Item Slope Regression?

There are some potential empirical clues of the presence of random item slopes, which may indicate when random item slope regression is a more appropriate choice than common factor regression or aggregation regression, e.g., poor fit of common factor regression. However, our results showed that simply checking traditional reliability or fit indices is not sufficient grounds on which to disregard random item slope regression; we have demonstrated that they are not sufficiently diagnostic for random item slopes. Thus, when random item slopes are present but small, there is a possibility that the data may be well approximated by a common factor regression model (i.e., the model shows acceptable fit). These observations suggest that random item slope regression and common factor regression are sometimes

empirically indistinguishable. Another complication is that, in practice, poor fit could be caused by a number of different factors (e.g., multidimensionality of the scale), and thus the observation of poor fit alone does not immediately indicate the presence of random item slopes. For example, without knowledge of the true model, it is unclear whether poor fit of the SEM observed in our real-world data example (e.g., predicting Self-actualization from Age) resulted from (a) random item slopes (which according to random item slope regression were significant), (b) the scale not being unidimensional (indicated by poor Cronbach's α and poor fit statistics for the SEM), (c) something else, or (d) a combination of a, b, and/or c.

As such, while researchers should be wary of empirical clues that random item slopes are present, their judgment should also be based on a substantive theoretical perspective about the construct that they are assessing. That is to say that their decision should rest on how they construe the measurement model for their data. If researchers believe that the construct is best described by a common factor, that is, the commonality of items exactly represents the construct of interest such that item-specific effects are irrelevant, and common factor regression is a reasonable choice (unless there is strong evidence of model misfit). On the other hand, if researchers believe that the construct is not strictly defined by such a model, and item-specific slopes reflect the important part of the construct, random item slope regression is a viable choice.

We believe that there are a number of situations in which random item slope regression is a more attractive practical alternative to common factor regression. Specifically, random item slope regression aligns with the typical strategy for developing psychological scales, whereas common factor regression does not. A good scale with substantial predictive validity tends to consist of a comprehensive set of items covering the broad spectrum of the psychological construct as a whole (e.g., Big Five personality scales). As such, researchers are often encouraged to generate heterogeneous, nonredundant sets of items to develop a scale (e.g., starting from a large pool of heterogeneous items; Loevinger, 1957). Random item slope regression is well suited to scales developed in this way. This is because random item slope regression captures common as well as unique (heterogeneous) construct-relevant elements of items in relation to the IV, acknowledging that idiosyncratic components of individual items are also an important part of the construct. In fact, while fixed effects slopes give the relationship between an IV and the commonality between items, inspection of the individual item slopes can provide more insight into what element(s) of the construct is related to external variables. For example, as discussed above, in our data, age could be particularly related to certain elements of self-actualization (e.g., relating to fear of failure or inadequacy) or interest-type curiosity (e.g., relating to enjoying discussion of abstract concepts) over other elements.

In contrast, common factor regression is not well suited to scales developed in way described above. As discussed, the latent factor in common factor regression only represents the "conjunction" between items (Figure 2, left) with item-specific elements considered irrelevant and modeled as measurement error (see Introduction). Furthermore, even elements that are shared by multiple, *but not all* items can also be regarded as a source of model misfit. In good scale design, increasing the number of items aims to increase the conceptual coverage of items, incorporating items that capture unique elements of the concept that is being assessed. However, common factor regression assesses the common component only, diluting the effects of item-specific effects, that is, marginalizing unique elements of the concept captured by individual items. In short, as noted in the introduction, when researchers are interested in assessing a relatively broad construct with heterogeneous set of items, we believe random item slope regression would be a valuable choice, well reflecting the nature of the psychological construct in focus.

Note that, when item-specific effects are present which cannot be explained by common factor regression, random item slope regression is not the only option. For example, studies using network models as a way to describe the relationships of multiple items from a scale are increasing. According to the network model, a psychological construct (e.g., depression) is an emergent property of the interaction of constituent elements such as behavioral symptoms assessed with individual items (Borsboom & Cramer, 2013; Fried et al., 2017). Elements assessed by individual items are supposed to have dynamic causal relationships with each other. Importantly, a network model assumes that each element has its own functions, underscoring the importance of item-specific effects. Network models and random item slope regression have similarities in that individual items are supposed to have differential relationships with an external variable (e.g., the IV). In fact, if a network model is the correct measurement model, we can imagine that substantive random item slope variance would be observed in random item slope regression.[7] In a way, random item slope regression could be seen as a convenient, theory-free model which allows researchers to examine the relationship between IV and constituent elements of DV without directly specifying the causal network structure of the elements. However, the fundamental difference is that, in network models, each element has its own functionality that is not exchangeable with other elements. As a result, it is essential to have a comprehensive set of items in order to correctly understand the dynamics between the items. On the other hand, random item slope regression assumes that items are exchangeable. That is, items are a small sample from the large item population, and the model accounts for this by correcting for statistical precision (i.e., increasing SE; see Appendix A). As such, there is no need for the scale to be comprehensive: One can still make an inference about the item population even from a limited set of items (although a larger number of items would improve statistical precision). Ultimately, the choice between using a network model and random item slope regression depends substantively on a researcher's theoretical perspective about the psychological constructs under examination as well as researcher's confidence in the comprehensiveness of the items in the assessment.

Extensions and Limitations

Although the current study covered a simple case in which a continuous multiitem DV is predicted by IVs, we propose several extensions of random item slope regression. First, the model is easily extended to cases in which the DV consists of noncontinuous multiitem DV. Examples include binary behavioral checklists such as the Biographical Inventory of Creative Behaviours (Batey, 2007; used as a DV in Furnham & Bachtiar, 2008). In this situation, researchers

---

[7] In this case, associations between the IV and individual items represent the total effect. For example, the item-specific slope between an IV and Item 1 represents the direct causal relationship between the IV and Item 1 as well as the sum of indirect effect through other items.

can use generalized linear mixed-effects models (Stroup, 2012) to explicitly model the nonlinear relationship between IVs and DV. Second, we can extend the model to deal with the case in which an IV consists of multiple items. Although mixed-effects models are applicable only when a DV (but not IV) has a nested/crossed structure, cross-classified SEM in the framework of multilevel SEM (Asparouhov & Muthén, 2016; González et al., 2008; Rabe-Hesketh et al., 2004) should allow researchers to model random item slopes with regard to the IV. As such, multilevel SEM can further extend the model to include both multiitem IVs and DVs. In this case, the model takes into account the slopes of every combination of the items between the IV and DV. Third, the model can also be extended to include item-specific covariates (i.e., characteristics of the items) to explain random item slopes (for a discussion on item level covariates see Rijmen et al. 2003). Such a model provides us with a great opportunity to understand why there are some variations in the slopes between items.

Fourth, while the proposed model in this article assumes homogenous error variances across items, that is, we assumed $\varepsilon_{ij} \sim N(0, \sigma^2)$, this assumption can be relaxed in a way that individual items have different error variance, that is, $\varepsilon_{ij} \sim N(0, \sigma_j^2)$. There are already several statistical models that relax this assumption, such as the heterogeneous variance model and mixed-effects location-scale model (see Lester et al., 2021 for an overview). Such a model can be easily implemented using a Bayesian framework (e.g., brms package in R; see Bürkner, 2017; see also McNeish, 2021 for implementation in Mplus). When we see large differences in item variance, it may be better to use such a less-constrained model, although more investigation on the effects of heterogeneous error variance is needed. We have reanalyzed the example relationships considered in Table 5 using Bayesian linear mixed-effects models that allow for heterogeneous error variances for each item, and included this analysis in Appendix D. We observed only slight changes in SE about the fixed effect compared to random item slope regression proposed in this manuscript (i.e., with no heterogeneous error variance), suggesting that the relaxed model offers marginal advantages in this particular context.

Finally, one limitation of random item slope regression is that it cannot control for measurement errors. As the model assumes that item-specific components constitute the important part of the construct, it cannot dissociate measurement errors in the same way as common factor regression. However, if researchers can collect the same data more than once (e.g., using test–retest design), we can explicitly model time as an additional random effect, which would represent measurement errors (defined as time-varying elements). This essentially makes it possible for us to draw an inference after correcting for measurement errors. Future studies should examine the model properties of these potential extensions.

References

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

Arslan, R. C., Schilling, K. M., Gerlach, T. M., & Penke, L. (2021). Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *Journal of Personality and Social Psychology*, *121*(2), 410–431. https://doi.org/10.1037/pspp0000208

Asparouhov, T., & Muthén, B. (2016). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In R. Harring, L. Stapleton, & S. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 163–192). Information Age.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batey, M. (2007). *A psychometric investigation of everyday creativity*. [Doctortal dissertation]. University College London.

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, *314*(7080), 572–572. https://doi.org/10.1136/bmj.314.7080.572

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Brennan, R. L. (2001). *Generalizability theory* (1st ed.). Springer. https://doi.org/10.1007/978-1-4757-3456-0

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor. *Clinical Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

De Boeck, P. (2008). Random item IRT models. *Pyschometrika*, *73*, 553–559. https://doi.org/10.1007/s11336-008-9092-x

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer-Verlag.

de Leeuw, J. R. (2015). Jspsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

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

Fried, E. I. (2015). Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. *Frontiers in Psychology*, *6*, Article 309. https://doi.org/10.3389/fpsyg.2015.00309

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: A review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 1–10. https://doi.org/10.1007/s00127-016-1319-z

Furnham, A., & Bachtiar, V. (2008). Personality and intelligence as predictors of creativity. *Personality and Individual Differences*, *45*(7), 613–617. https://doi.org/10.1016/j.paid.2008.06.023

González, J., De Boeck, P., & Tuerlinckx, F. (2008). A double-structure structural equation model for three-mode data. *Psychological Methods*, *13*(4), 337–353. https://doi.org/10.1037/a0013269

Halekoh, U., & Højsgaard, S. (2014). A Kenward–Roger approximation and parametric bootstrap methods for tests in linear mixed models—The R package pbkrtest. *Journal of Statistical Software*, *59*(9), 1–32. https://doi.org/10.18637/jss.v059.i09

Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The Balanced Inventory of Desirable Responding short form (BIDR-16). *SAGE Open*, *5*(4), https://doi.org/10.1177/2158244015621113

James, A. N., Fraundorf, S. H., Lee, E.-K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, *102*(1), 155–181. https://doi.org/10.1016/j.jml.2018.05.006

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). Guildford Press.

Jones, A., & Crandall, R. (1986). Validation of a Short Index of self-actualization. *Personality and Social Psychology Bulletin*, *12*(1), 63–73. https://doi.org/10.1177/0146167286121007

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kajimura, S., Hoshino, T., & Murayama, K. (2023). Stimulus-specific random effects inflate false-positive classification accuracy in multivariate-voxel-pattern-analysis: A solution with generalized mixed-effects modeling. *NeuroImage*, *269*, Article 119901. https://doi.org/10.1016/j.neuroimage.2023.119901

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558

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*(8), 1652–1672. https://doi.org/10.1002/bimj.202100046

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods*, *49*(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Lester, H. F., Cullen-Lester, K. L., & Walters, R. W. (2021). From nuisance to novel research questions: Using multilevel models to predict heterogeneous variances. *Organizational Research Methods*, *24*(2), 342–388. https://doi.org/10.1177/1094428119887434

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, *44*(7), 1585–1595. https://doi.org/10.1016/j.paid.2008.01.014

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Martel, M. M., Pan, P. M., Hoffmann, M. S., Gadelha, A., do Rosário, M. C., Mari, J. J., Manfro, G. G., Miguel, E. C., Paus, T., Bressan, R. A., Rohde, L. A., & Salum, G. A. (2017). A general psychopathology factor (P factor) in children: Structural model analysis and external validation through familial risk and child global executive function. *Journal of Abnormal Psychology*, *126*(1), 137–148. https://doi.org/10.1037/abn0000205

McClure, K., Jacobucci, R., & Ammerman, B. (2021). *Are items more than indicators? An examination of psychometric homogeneity, item-specific effects, and consequences for structural equation models*. https://doi.org/10.31234/osf.io/n4mxv

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

McNeish, D. (2021). Specifying location-scale models for heterogeneous variances as multilevel SEMs. *Organizational Research Methods*, *24*(3), 630–653. https://doi.org/10.1177/1094428120913083

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Mõttus, R. (2016). Towards more rigorous personality trait–outcome research. *European Journal of Personality*, *30*(4), 292–303. https://doi.org/10.1002/per.2041

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*(3), 474–490. https://doi.org/10.1037/pspp0000100

Mõttus, R., & Rozgonjuk, D. (2021). Development is in the details: Age differences in the Big Five domains, facets, and nuances. *Journal of Personality and Social Psychology*, *120*(4), 1035–1048. https://doi.org/10.1037/pspp0000276

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1287–1306. https://doi.org/10.1037/a0036914

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

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. https://doi.org/10.1007/BF02295939

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

R Core Team. (2019). *R: A language and environment for statistical computing* (Version 3.6.2) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*(2), 185–205. https://doi.org/10.1037/1082-989X.8.2.185

Rose, N., Wagner, W., Mayer, A., & Nagengast, B. (2019). Model-based manifest and latent composite scores in structural equation models. *Collabra: Psychology*, *5*(1), Article 9. https://doi.org/10.1525/collabra.143

Rosseel, Y. (2012). {Lavaan}: An {R} package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*(2), 133–166. https://doi.org/10.1111/j.2044-8317.1981.tb00625.x

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In J. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 309–322). Lawrence Erlbaum Associates.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Srivastava, V. K., & Giles, D. E. A. (1987). *Seemingly unrelated regression equation models: Estimation and inference*. Marcel Dekker.

Stroup, W. W. (2012). *Generalized linear mixed models: Modern concepts, methods and applications*. CRC Press.

Usami, S., & Murayama, K. (2018). Time-specific errors in growth curve modeling: Type-1 error inflation and a possible solution with mixed-effects models. *Multivariate Behavioral Research*, *53*(6), 876–897. https://doi.org/10.1080/00273171.2018.1504273

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology*, *27*(6), 759–773. https://doi.org/10.1177/0959354317737185

VanderWeele, T. J. (2022). Constructed measures and causal inference. *Epidemiology*, *33*(1), 141–151. https://doi.org/10.1097/EDE.0000000000001434

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier tests of the validity of the bifactor model of psychopathology. *Clinical Psychological Science*, *7*(6), 1285–1303. https://doi.org/10.1177/2167702619855035

Westfall, J., Nichols, T. E., & Yarkoni, T. (2017). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, *1*, 1–23. https://doi.org/10.12688/wellcomeopenres.10298.1

Wiernik, B. M., Wilmot, M. P., & Kostal, J. W. (2015). How data analysis can dominate interpretations of dominant general factors. *Industrial and Organizational Psychology*, *8*(3), 438–445. https://doi.org/10.1017/iop.2015.60

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209. https://doi.org/10.3758/s13428-016-0779-0

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, *57*(298), 348–368. https://doi.org/10.1080/01621459.1962.10480664

(Appendices follow)

Appendix A

Let I be the number of participants and J be the number of items. In Equation 4, we posit $u_{0i} \sim N(0, \omega_{00})$, $\boldsymbol{u}_j = (u_{0j}, u_{1j})' \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}\right)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and assume that these random effects and residuals are uncorrelated with each other. We also assume that these random effects and residuals are uncorrelated with the IV x_i. Then the variance of the DV y_{ij} [= $\text{Var}(y_{ij}|x_i)$], its covariance between participants for the same item j [= $\text{Cov}(y_{ij}, y_{i'j}|x_i, x_{i'})$], its covariance between items for the same participant i [= $\text{Cov}(y_{ij}, y_{ij'}|x_i)$], and its covariance between different items and participants [= $\text{Cov}(y_{ij}, y_{i'j'}|x_i, x_{i'})$] can be expressed as follows, respectively:

$$\text{Var}(y_{ij}|x_i) = \omega_{00} + \tau_{00} + 2x_i\tau_{01} + x_i^2\tau_{11} + \sigma^2, \quad \text{(A1)}$$

$$\text{Cov}(y_{ij}, y_{i'j}|x_i, x_{i'}) = \tau_{00} + (x_i + x_{i'})\tau_{01} + x_ix_{i'}\tau_{11}, \quad \text{(A2)}$$

$$\text{Cov}(y_{ij}, y_{ij'}|x_i) = \omega_{00}, \quad \text{(A3)}$$

$$\text{Cov}(y_{ij}, y_{i'j'}|x_i, x_{i'}) = 0. \quad \text{(A4)}$$

We then define the vector of the DV with size IJ as $Y = (Y_1^t, Y_2^t, \ldots, Y_i^t, \ldots, Y_I^t)^t$ $Y_i = (y_{i1}, y_{i2}, \ldots, y_{ij}\ldots, y_{iJ})^t$, the vector of regression coefficients with size 2 as $\boldsymbol{\beta} = (\beta_0, \beta_1)^t$, and the matrix of the IV with size $IJ \times 2$ as $X = (1_{IJ}, \ X^+)$, $X^+ = (X_1^{+t}, X_2^{+t}, \ldots, X_i^{+t}, \ldots, X_I^{+t})^t$, $X_i^+ = x_i 1_J$ (1_a is a vector of ones with size a), and vector of random effects with size IJ as $\boldsymbol{\varepsilon}$. Then Equation 4 can be expressed in a matrix form as follows:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad \text{(A5)}$$

Here $\text{Cov}(X, \boldsymbol{\varepsilon}) = 0$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \boldsymbol{\Sigma})$. Elements of $\boldsymbol{\Sigma}$ (which is a nondiagonal matrix) can be derived from Equations A1 to A4. Specifically, $\sigma_{ij,i'j'}$, which is the covariance of ith participant in jth item and i'th participant in j'th item can be generally written as follows.

$$\sigma_{ij,ij} = \omega_{00} + \tau_{00} + 2x_i\tau_{01} + x_i^2\tau_{11} + \sigma^2 \ (i = i' \& j = j'), \quad \text{(A6)}$$

$$\sigma_{ij,i'j} = \tau_{00} + (x_i + x_{i'})\tau_{01} + x_ix_{i'}\tau_{11} \ (j = j'), \quad \text{(A7)}$$

$$\sigma_{ij,ij'} = \omega_{00} \ (i = i'), \quad \text{(A8)}$$

$$\sigma_{ij,i'j'} = 0. \quad \text{(A9)}$$

Considering the current case in which the IV does not depend on j ($X_i^+ = x_i 1_J$), the point estimate of the generalized least squares (GLS) of the regression coefficient is the same as the point estimate of OLS (see Amemiya, 1985), which is the proportion between the covariance and the variance of the IV. Specifically, $\hat{\beta}_1 = \dfrac{\text{Cov}(y_{ij}, x_i)}{s_x^2}$. The SE of the GLS estimator of the regression

coefficient can be analytically computed as:

$$se(\hat{\beta}_1) = \sqrt{(X^t\boldsymbol{\Sigma}^{-1}X)_{[2,2]}^{-1}} = \sqrt{\frac{I\tau_{11}s_x^2 + \sigma^2 + J\omega_{00}}{IJs_x^2}}. \quad \text{(A10)}$$

Here $A_{[m,n]}$ refers to (m, n) element in the matrix A. We can see that $se(\hat{\beta}_1)$ is not dependent of τ_{00}, τ_{01}.

Note that the derived formula makes it clear that random item slope variance τ_{11} adds to the SE, as $I\tau_{11}s_x^2 \geq 0$. As such, if the model in Equation 4 is true and a researcher mistakenly applied a model in Equation 3, we can expect underestimation of the SE. To show the consequence of this model misspecification more specifically, let us consider the elements of $\boldsymbol{\Sigma}$ when we apply a model in Equation 3:

$$\sigma_{ij,ij} = \omega_{00}^* + \tau_{00}^* + \sigma^{*2} \ (i = i' \& j = j'), \quad \text{(A11)}$$

$$\sigma_{ij,i'j} = \tau_{00}^* \ (j = j'), \quad \text{(A12)}$$

$$\sigma_{ij,ij'} = \omega_{00}^* \ (i = i'), \quad \text{(A13)}$$

$$\sigma_{ij,i'j'} = 0. \quad \text{(A14)}$$

Here, $\omega_{00}^*, \tau_{00}^*$, and σ^{*2} are the parameters in Equation 3, which are similarly defined as ω_{00}, τ_{00}, and σ^2. By comparing these elements with those obtained under the assumption that the model in Equation 4 is true (i.e., the elements described above), we can derive the asymptotic parameter estimates $\hat{\omega}_{00}^*, \hat{\tau}_{00}^*, \hat{\sigma}^{*2}$ if one mistakenly applied the model in Equation 3 when Equation 4 is the true model. Specifically, by comparing the elements, we can derive the following relations:

$$\hat{\omega}_{00}^* + \hat{\tau}_{00}^* + \hat{\sigma}^{*2} = \frac{1}{IJ}\sum_j\sum_i\left[\omega_{00} + \tau_{00} + 2x_i\tau_{01} + x_i^2\tau_{11} + \sigma^2\right]$$
$$= \omega_{00} + \tau_{00} + 2\bar{x}\tau_{01} + (\bar{x}^2 + s_x^2)\tau_{11} + \sigma^2, \quad \text{(A15)}$$

$$\hat{\tau}_{00}^* = \frac{1}{I(I-1)J}\sum_j\sum_i\sum_{i'=i}[\tau_{00} + (x_i + x_{i'})\tau_{01} + x_ix_{i'}\tau_{11}]$$
$$= \tau_{00} + 2\bar{x}\tau_{01} + \left(\bar{x}^2 - \frac{1}{I-1}s_x^2\right)\tau_{11}, \quad \text{(A16)}$$

$$\hat{\omega}_{00}^* = \omega_{00}. \quad \text{(A17)}$$

Here $\bar{x}$ is the sample mean of x. By solving the equations with regard to $\hat{\sigma}^{*2}$, we have

$$\hat{\sigma}^{*2} = \frac{I}{I-1}\tau_{11}s_x^2 + \sigma^2. \quad \text{(A18)}$$

Accordingly, if one mistakenly applied the model in Equation 3 when Equation 4 is the true model, the SE of the

GLS estimator of the regression coefficient $se(\hat{\beta}_1^*)$ can be expressed as:

$$se(\hat{\beta}_1^*) = \sqrt{\frac{\hat{\sigma}^{*2} + J\hat{\omega}_{00}^*}{IJs_x^2}} = \sqrt{\frac{\frac{I}{I-1}\tau_{11}s_x^2 + \sigma^2 + J\omega_{00}}{IJs_x^2}}$$

$$\leq \sqrt{\frac{I\tau_{11}s_x^2 + \sigma^2 + J\omega_{00}}{IJs_x^2}} = se(\hat{\beta}_1). \quad \text{(A19)}$$

As I is a positive constant, Equation A19 indicates that SE would be underestimated unless $\tau_{11} = 0$. This also shows that the underestimation is larger when (a) I is larger, (b) $\tau_{11}s_x^2$ is larger, and (c) $\sigma^2 + J\omega_{00}$ is smaller. It should be noted that the point estimate of the GLS of the regression coefficient is $\hat{\beta}_1^* = \dfrac{\text{Cov}(y_{ij}, x_i)}{s_x^2} = \hat{\beta}_1$, meaning that the model misspecification would not bias parameter estimate itself.

Next, we consider the case where we aggregated item scores for each participant to apply a model in Equation 1. First, assume that Equation 4 is true. Given that $\sum_j^J u_{0j} = 0$ and $\sum_j^J u_{1j} = 0$, we can derive y_i, the aggregated score for each participant as follows:

$$y_i = \sum_j^J y_{ij} = \sum_j^J (\beta_0 + u_{0i} + u_{0j} + (\beta_1 + u_{1j})x_i + e_{ij})$$

$$= \beta_0^{**} + \beta_1^{**}x_i^{**} + e_i^{**} \quad . \quad \text{(A20)}$$

The equation takes a similar form with Equation 1. Here, $\beta_0^{**} = J\beta_0$, $\beta_1^{**} = \beta_1$, $x_i^{**} = Jx_i$, $e_i^{**} = Ju_{0i} + \sum_j^J e_{ij}$. Also, we can derive that $\text{Var}(e_i^{**}) = J^2\omega_{00} + J\sigma^{*2}$. Accordingly, considering that $\text{Var}(x_i^{**}) = s_x^{**2} = J^2s_x^2$, SE of the OLS estimator for Equation 1, $se(\hat{\beta}_1^{**})$, can be derived as follows:

$$se(\hat{\beta}_1^{**}) = \sqrt{\frac{\text{Var}(e_i^{**})}{Is_x^{**2}}} = \sqrt{\frac{J^2\hat{\omega}_{00}^* + J\hat{\sigma}^{*2}}{IJ^2s_x^2}} = \sqrt{\frac{J\hat{\omega}_{00}^* + \hat{\sigma}^{*2}}{IJs_x^2}} \quad \text{(A21)}$$

$$= se(\hat{\beta}_1^*).$$

This is equivalent to Equation A19. Therefore, when the model in Equation 4 is true, SE of the GLS estimator of regression coefficient in Equation 3 [$se(\hat{\beta}_1^*)$] is mathematically identical to that of the OLS estimator of regression coefficient in Equation 1 [$se(\hat{\beta}_1^{**})$]. Importantly, regardless of the true regression model that explains outcome from IV, the aforementioned relations $\text{Var}(e_i^{**}) = J^2\omega_{00}^* + J\sigma^{*2}$ and $\text{Var}(x_i^{**}) = s_x^{**2} = J^2s_x^2$ always hold because these equations do not include any parameters of the true model. As such, the equivalence of the SE holds for any empirical data.

As noted earlier, the variance of the IV in Equation 1 is $\text{Var}(x_i^*) = J^2s_x^2$. In other words, by aggregating the variable across items, the variance increased by J^2 times. Similarly, the covariance between the IVs and DVs in Equation 1 also increases by J^2 times, that is, $\text{Cov}(y_i, x_i^*) = J^2\text{Cov}(y_{ij}, x_i)$). The estimator of the regression coefficient is equal to the proportion between the covariance and the variance of the IV, regardless of whether one uses GLS (in Equation 3, $\hat{\beta}_1^*$) or OLS (in Equation 1, $\hat{\beta}_1^{**}$). Accordingly,

$$\hat{\beta}_1^{**} = \frac{\text{Cov}(y_i, x_i^{**})}{\text{Var}(x_i^{**})} = \frac{J^2\text{Cov}(y_{ij}, x_i)}{J^2s_x^2} = \frac{\text{Cov}(y_{ij}, x_i)}{s_x^2} = \hat{\beta}_1^*, \quad \text{meaning}$$

that the point estimates of the regression coefficient are also equivalent between Equations 1 and 3. Therefore, for any empirical data, the regression coefficient estimate and its SE from Equation 1 are always equivalent with those from Equation 3.

(Appendices continue)

Appendix B

We tested an SEM version of an extended SUR model (described in the introduction, depicted in Figure B1) on the same simulated data sets as reported in the Simulation Study. To obtain an estimate of the DV ∼ IV relationship, we averaged across Individual Item ∼ IV relationships (e.g., β_{11}, β_{12}, …, β_{15} in Figure B1) and statistically tested it. This model also demonstrates Type I error inflation (Figure B2).

Figure B1

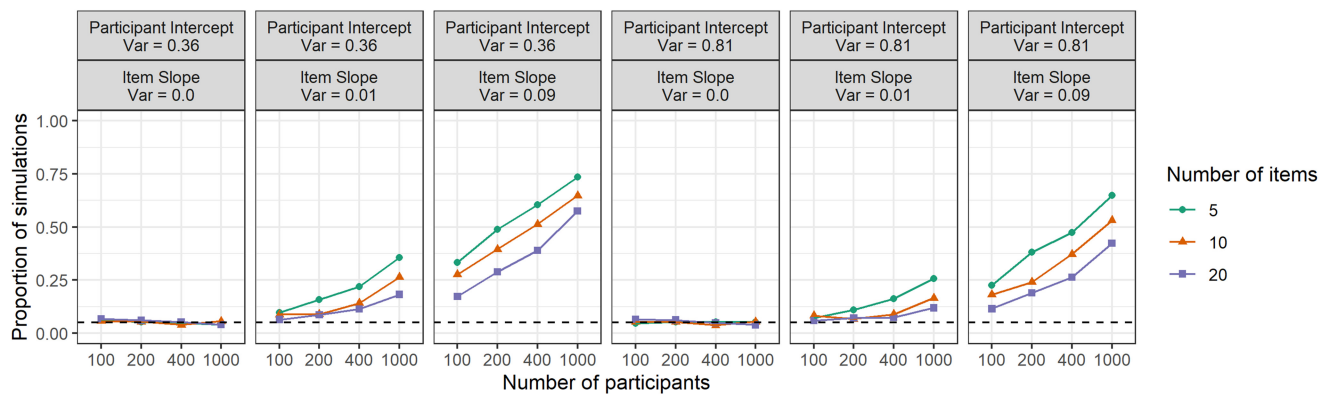*SEM Version of a SUR Model*
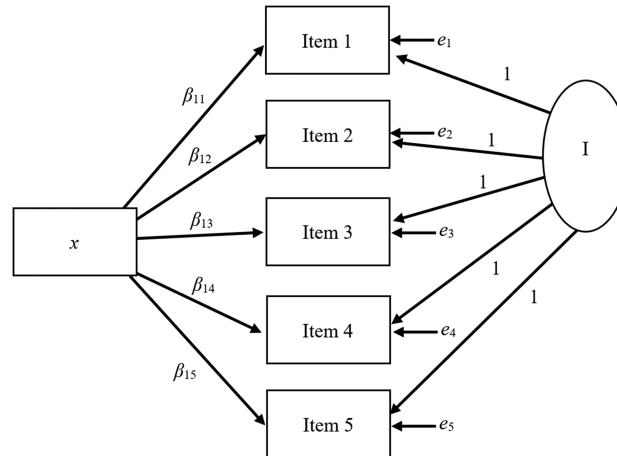
Note. var (e_1) = var (e_2) = … = var (e_5). SEM = structural equation modeling; SUR = seemingly unrelated regression.

Figure B2

Simulation Results Showing Type I Error of the SEM Version of a SUR (Figure B1) on Data With Varying Participant Intercepts and Item Slopes

Note. SEM = structural equation modeling; SUR = seemingly unrelated regression. See the online article for the color version of the figure.
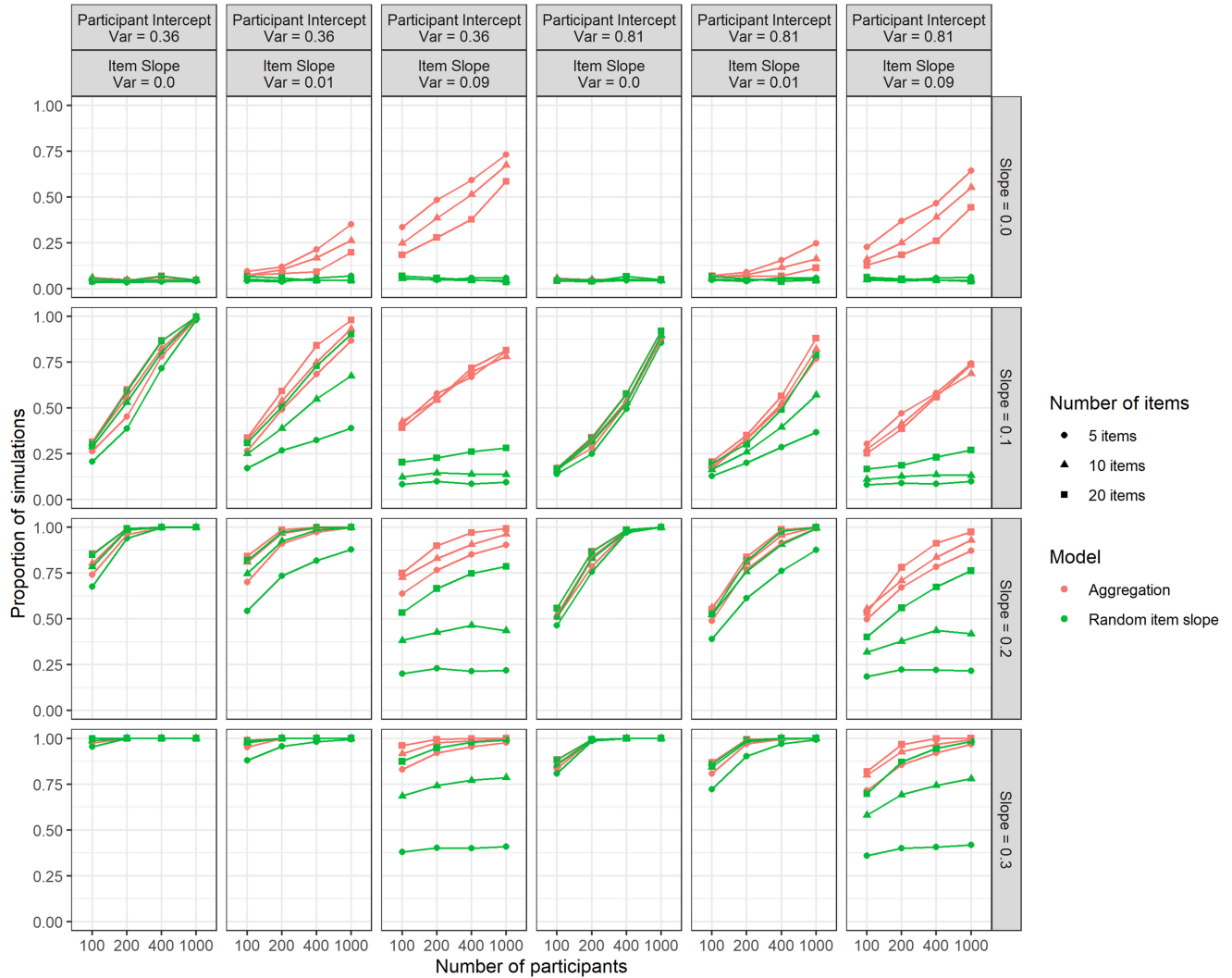
(Appendices continue)

Appendix C

Figure C1 shows the result of a simulation using the same parameters as in the main simulation, but introducing different sized fixed effects slopes (0, 0.1, 0.2, 0.3) for both aggregation and random item slope regression.

Figure C1

Simulation Results Showing Proportion of Simulations With Significant Results for Aggregation and Random Item Slope Regression on Data With Varying Participant Intercepts, Item Slopes, and Fixed Effects Slopes

Note. See the online article for the color version of the figure.

Appendix D

Using *brms* (Bürkner, 2017) in *R*, we ran a Bayesian multilevel model equivalent to Equation 4, but where $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, that is, allowing the variance across items to be heterogeneous for the three example relationships in Table 5. The models used default priors, four chains with 2,000 iterations (including 1,000 warm-up iterations). Estimated population-level effects (equivalent to fixed effects) and group-level effects (equivalent to random effects) are shown in Table D1. Note that σ_j is assumed to follow a log-normal distribution: $\log(\sigma_j) \sim N(\mu, v^2)$, where the mean $= \exp(\mu + v^2/2)$ and variance $= \exp(2\mu + v^2)(\exp(v^2) - 1)$. This means that μ can take a negative value (as seen in Table D1).

Table D1

Bayesian Multilevel Model Results for Three Example Relationships (Agreeableness ∼ Age, Interest-Type Curiosity ∼ Age, Self-actualization ∼ Age)

Model	Population-level effects				Group-level effects			
	$\hat{\beta}_1$	*SE*	μ	*SE*	ω_{00}^2	τ_{00}^2	τ_{11}^2	v^2
Agreeableness ∼ Age	0.031	0.025	−0.226	0.071	0.206	0.220	0.001	0.040
Interest-Type Curiosity ∼ Age	−0.119	0.087	−0.358	0.121	0.410	0.227	0.026	0.059
Self-actualization ∼ Age	−0.063	0.035	−0.160	0.027	0.071	0.234	0.014	0.009