Methods in Psychology 15 (2026) 100275

Contents lists available at ScienceDirect

Methods in Psychology

journal homepage: www.sciencedirect.com/journal/methods-in-psychology

To aggregate or not to aggregate: Effect variant measurement models and their advantages for questionnaire data in experimental psychology

Giorgia Tosi *, Marcello Gallucci, Daniele Romano

Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, Milan, 20126, Italy

ARTICLE INFO

Keywords:
Random effects
Mixed models
Item analysis
Reliability

ABSTRACT

Experimental psychological research frequently relies on aggregating items, assuming that the effects of other variables remain constant across items. This assumption can lead to biased statistical inferences, most notably an inflated Type I error rate, as it disregards the random variability of item-specific effects. This paper highlights the potential biases that can arise from using aggregated scores. It shows the conditions under which effect-variant models, which treat both participants and items as random factors, are more appropriate than traditional aggregation approaches. We introduce the Aggregation Standard Error Inflation (ASEI), a novel metric designed to quantify the variance inflation associated with conventional aggregation methods and the related risks. Through a series of simulations and real data analyses, we show that neglecting item-specific variability in aggregated models increases Type I error rates, particularly in larger samples. The extent of this inflation can be effectively assessed using the ASEI. These findings highlight the importance of incorporating random variability across both items and participants to improve the validity of statistical inferences in psychological research that relies on questionnaires and scales.

In Classical Test Theory, stimuli measuring the same construct are correlated and parallel, meaning that each item should be distinct from the others, yet similar and comparable (Algina and Penfield, 2009). The use of having more than one item or stimulus to assess a construct is to have a more reliable measure of the construct itself. However, when many items are utilized to measure a single construct, the standard practice is to aggregate them by averaging (or summing) participants' responses and analyzing the resulting averages. This is a typical situation in experimental psychology: for example, studies evaluating intervention efficacy on psychological symptoms often adopt this procedure (Doorn et al., 2023; Harley et al., 2007; Lush et al., 2009). Sometimes, the averaging procedure is preceded by a factor analysis or an internal consistency check to confirm the questionnaire/task structure and the internal consistency of the aggregated score. However, in cognitive and experimental psychology, experiments typically have small sample sizes ($N < 100$; Bogdan, 2025), thus preventing an accurate and reliable preliminary check of the items' structure. As a multi-stimuli measurement case study, we can examine the case of body illusions (Botvinick and Cohen, 1998; Frisco et al., 2024; Serino et al., 2023; Tosi et al., 2022). The illusion effects are typically assessed through questionnaires (Botvinick and Cohen, 1998; Longo et al., 2008; Romano et al., 2021) comprising varying numbers of items, designed to capture different psychological constructs (e.g., embodiment, disembodiment, ownership, physical sensations). When multiple items target a single construct – such as the sensation of ownership – it is common practice to average participants' responses across those items and to analyze these aggregated scores.

A crucial assumption of this approach is that all properties of a set of items or stimuli are independent of the specific context in which the measurement is used: the reliability of a scale remains the same irrespective of the independent variables used to predict it, or its factor structure does not change when its scores are compared among groups or experimental conditions. In other words, the only psychometric properties considered in the classical measurement model are those that do not depend on the relationships, or effects, the variable shares with other variables in a study. We shall call this assumption *effect invariance*. Effect invariance can be conceived as a form of fixed effects linear model. In this context, an effect (or more generally a coefficient) is named *fixed* when it remains constant for all cases in the population (Green and Tukey, 1960; Winer, 1962). The Pearson correlation between two variables, for instance, is assumed to have one single value in the population, which applies to all cases in the sample. On the other

* Corresponding author.
E-mail address: giorgia.tosi@unimib.it (G. Tosi).

https://doi.org/10.1016/j.metip.2026.100275
Received 14 January 2026; Received in revised form 22 July 2026; Accepted 24 July 2026
Available online 25 July 2026

hand, effects can also be *random*, meaning that they vary across levels of a random factor in the design and are sampled from a larger population of levels (Green and Tukey, 1960; Winer, 1962). Participants, which are frequently treated as a random factor, are sampled from a population to which we want to make inferences. Thus, the estimated effects are allowed to vary across participants, such as in repeated measures analyses. Following the same logic, when we select a specific set of stimuli for an experiment, a set of items for a questionnaire, or a sequence of tasks in a neuropsychological test, we are sampling from a population of potential stimuli to which we want to make inferences (Cronbach et al., 1963; McDonald, 2013; Shavelson and Webb, 1981). Thus, the effects of other variables may randomly vary across stimuli. However, items are classically aggregated by summing them, or factor scores are computed as weighted combinations of the items, yielding a combined score for each participant (cf. Donnellan et al., 2025). This means that when the multi-stimuli measurement is used to estimate the effects it receives from other variables, it is assumed that these effects are fixed across stimuli. This practice precludes the possibility of treating stimuli as random factors, and thus, their potential variability is ignored. The consequences of this assumption may be profound and unexpected (Donnellan et al., 2025; Judd et al., 2012; Locker et al., 2007), potentially biasing the results obtained in the study.

Locker et al. (2007) examined how selecting words for experimental designs affects statistical analysis, noting that traditional ANOVA approaches assume materials are fixed factors (Clark, 1973). The authors showed that randomizing the selection of items introduces random variance that could bias the F-test, thus increasing the likelihood of Type I error (2007). To face this problem, Clark (1973) proposed using a quasi-F ratio, a random effects model that considers both item and participant variability. This method is limited by needing complete and balanced samples, and cannot accommodate continuous covariates. Locker et al. (2007) advocate the crossed random factors multilevel model, where both participants and items can be considered as random effects simultaneously. Judd et al. (2012) extended this reasoning to social psychology, showing that ignoring stimulus sampling biases F-tests even under the null hypothesis, and advocated mixed models to control over an appropriate rate of Type I errors. This issue has been recently framed by Yarkoni (2022) as a generalizability crisis, where the failure to model stimuli as random effects (i.e., the stimulus-as-fixed-effect fallacy) can inflate false-positive rates by up to 300%.

More recently, Donnellan et al. (2025) took the matter to a more general level, proposing to adopt the idea of *effect variance* as a general measurement model. They proposed *random item slopes models* for multi-item measures, arguing that aggregated scores underestimate standard errors and inflate Type I error. They showed that this model is inherently different from several other measurement models, such as the aggregated model (scores computed by summing up the items of the measurement), the multilevel SEM (Mehta and Neale, 2005), or random coefficients IRT (Rijmen et al., 2003). Random item slopes, however, are closely related to Differential Item Functioning (*DIF*; Angoff, 1981; Holland and Wainer, 2012; Holland and Thayer, 1986; Meredith, 1993), and we return to this issue in the discussion. Crucially, Donnellan and collaborators (2025) demonstrated analytically that the standard errors of the estimated effects in a linear model predicting the aggregated score of a multi-item measure are downwardly biased by the potential variability of the effects across items, thus increasing the Type I error of the associated tests.

Gilbert et al. (2025) reinforced this argument by applying an item-level heterogeneous treatment effect (IL-HTE) model to 75 RCT datasets, providing a formula for standard error inflation. They demonstrated that standard error inflation ratios can exceed a factor of 4, making significant results statistically insignificant when item sampling uncertainty is considered. In line with this, Ahmed et al. (2025) found that, among significant RCTs, 41% allowed for the possibility of a null effect when item-level sensitivity was accounted for. They also noted that, in extreme cases, a single item can account for nearly 40% of the observed treatment effect. Crucially, Halpin and Gilbert (2024) showed that effects on the latent variable and on the item aggregate scores are not equivalent in the presence of item-level heterogeneity. Although some item sets might be truly fixed, making a random intercepts or fixed effects model more conceptually suitable (Gilbert et al., 2023; Miratrix et al., 2021), if there is genuine heterogeneity among the items and this is ignored by treating the items as fixed when a generalization is desired, there is a risk of underestimating the standard errors and increasing the false positive rate (Donnellan et al., 2025; Gilbert et al., 2023). An unresolved issue, however, is whether these considerations extend to within-subject designs and how to estimate the risk associated with the aggregation procedure. In their discussion, Donnellan and collaborators (2025) acknowledged that the choice between a standard common factor measurement model and a *random item slopes model* should be guided by both substantive theory and empirical data.

We further developed the latter argument, providing a psychometric index to quantify the error inflation associated with aggregating items. We further explore the estimated effects arising from disregarding the variability across items and how the error inflation is related to the proposed psychometric indices.

1. The model

Consider a multi-item measure, such as a scale or a psychological test, with I items. Assume the data are organized in the long-format, where each row of the data represents the score of the item i for the participant p. A common factor measurement model may be represented in terms of linear models as a random intercept model, with intercepts varying across the random factor participants.

$$y_{ip} = a + a_p + a_i + e_{pi} \tag{1}$$

where a is the fixed intercept, a_p is the participant intercept expressed as deviation from a, a_i is the item (deviation) intercept, and e_{pi} is the residual error, uncorrelated with any other coefficient in the model. For simplicity, we assume that all parameters are normally distributed, with mean equal to zero and variance σ_P^2 for the participant random intercept a_p, σ_I^2 for a_i, and σ^2 for the residual term e_{pi}. If scores of the measure are aggregated across items to compute the scale score, one obtains the classical aggregation score $\overline{y_P} = a + a_p$. The score reliability, expressed as Cronbach's alpha, is (Shrout and Fleiss, 1979):

$$\alpha = \frac{\sigma_P^2}{\sigma_P^2 + \frac{\sigma^2}{I}} \tag{2}$$

We now consider a model with an independent variable x predicting the y score. The independent variable x is measured across participants, so it has the same value for all items. We start with the simple case in which the independent variable x has unitary variance and varies between participants, such as an experimental manipulation with two between-participant conditions x_p. The model is specified for the item-level responses y_{ip} , with e_{pi} retaining the same distributional assumptions as in Equation (1) ($e_{pi} \sim N(0,\sigma^2)$, uncorrelated with all other terms.

$$y_{ip} = a + a_p + a_i + b' \cdot x_p + e_{pi} \tag{3}$$

The effect coefficient b' is equivalent to the coefficient one would obtain by aggregating the y scores, and its standard errors are the same as the ones obtained in a linear model on the aggregated data[1] (see

[1] The exact Correspondence between the mixed model in (3) and the aggregated model standard errors and inferential tests requires the mixed model coefficients to be tested with *Kenward-Roger* degrees of freedom. Nonetheless, the logical equivalence between the models remains independent of the degrees of freedom being used.

Donnellan et al. (2025); Appendix 1), and we refer to the model as the *aggregation model*. As a consequence, considering items as a random factor in which only the intercepts are random does not change the underlying measurement model as compared with the classical aggregation model. The crucial step is to allow the effect of the independent variable x to vary across items, so to set the coefficient as a random effect. The *effect variant model* can be defined as[2]:

$$y_{ip} = a + a_p + a_i + (b + b_i)\cdot x_p + e_{pi} \tag{4}$$

where b_i is the random slope of x varying across items, expressed as deviations from b, with variance τ_I^2. The crucial feature of the model is that it expands generalizability theory (Brennan, 2001) by taking into account not only the source of variations of the measure scores, but also possible variability of the effects of the independent variable across items.

1.1. Effect variance sensitivity index

Judd et al. (2012) and Donnellan et al. (2025) have both demonstrated that disregarding the random effects of an independent variable across items or stimuli may largely increase the Type I error associated with the inferential test. Judd et al. (2012) demonstrated that the F-test of an aggregation model, such as model (3), is upward-biased proportionally to the variance of the random effects across stimuli. Donnellan et al. (2025) demonstrated that the standard error computed for the aggregation model (3) is smaller than expected under the presence of random effects across items. Both contributions lead to the conclusion that, in the presence of random effects, the results obtained by aggregating items or not accounting for this variability would bias the conclusions of the tests. Interestingly, both contributions showed that the fixed effects estimates (i.e., the b's) do not vary between models, so the fact that the models yield the same fixed effects coefficients is not informative about the presence or the size of the error inflation.[3]

The size of the bias associated with not considering the random effects of the independent variable across items depends on the size of the variance of the random effects, but relatively to the other parameters of the model. More specifically, in the between-participants case, Donnellan et al. (2025) showed that when random effects are present in the data, but an aggregation model is considered (thus ignoring them), the standard error of the fixed effect b' coefficient is:

$$se(b') = \sqrt{\frac{\frac{P}{(P-1)\tau_I^2} + I\sigma_p^2 + \sigma^2}{IP}} \tag{5}$$

whereas the correct standard error of the estimate, under the presence of random effects, should be

$$se(b) = \sqrt{\frac{P\tau_I^2 + I\sigma_P^2 + \sigma^2}{IP}} \tag{6}$$

Thus, we can take the ratio of the two standard errors to construct, after some algebraic manipulation, an index of *effect variance sensitivity*, which indicates the proportion of variance inflation associated with not considering the random effects and simply aggregate the items. The index, named *Aggregation Standard Error Inflation (ASEI)* for the between-participants case, can be computed as:

$$ASEI_b = 1 - \sqrt{\frac{\frac{P}{(P-1)\tau_I^2} + I\sigma_P^2 + \sigma^2}{P\tau_I^2 + I\sigma_P^2 + \sigma^2}} \tag{7}$$

The $ASEI_b$ indicates the expected inflation in the inferential results of the aggregated model as compared with the complete random coefficients model. It is a relative index varying from 0 to 1 and can be used to understand how different parameters of the model influence the results. $ASEI_b$ is zero if there is no variability of the effects across items ($\tau_I^2 = 0$), indicating that the aggregation model or the mixed model yields the same results, so there is no risk in aggregating the items. However, when τ_I^2 is not null, the aggregability of the measure becomes dubious, and a bias in the results of the aggregation model should be expected. Fig. 1 shows how $ASEI_b$ changes as the parameters vary.

Inspecting $ASEI_b$ clarifies which design and data parameters drive bias in aggregation models compared to effect-variant models. $ASEI_b$ increases with the number of participants (P); a counterintuitive result, as larger samples usually improve accuracy, but here they reduce standard errors and inflate Type I error. Conversely, $ASEI_b$ decreases with more items (I), consistent with reliability theory that fewer items yield less reliable measures (Cronbach, 1951). We also find $ASEI_b$ grows with greater variation in participant intercepts, as it is directly related to the reliability of the measure (cf. (2)), and lower reliability amplifies residual error variance. Finally, weaker effects or less reliable measures further increase $ASEI_b$ because a larger $ASEI_b$ is expected when there is larger residual error variance. These conclusions align with Donnellan et al. (2025)'s findings on standard errors and *random item slope regression*. Similarly, Gilbert et al. (2025) provided a formula to compute the inflation of the standard error of the effect-invariant model based on the items' random variances and their number, and showed that the ratio of standard errors from the item-level heterogeneous treatment effect model to the random intercepts model increases with increasing item variance and decreases with increasing number of items.

1.2. Repeated measures designs

We can apply the same reasoning to the within-participants design, namely, taking the ratio of the standard errors of the regression coefficient under the *aggregation model* and the *effect variant model*. We consider a repeated-measure design where the same set of items is used to measure a sample of participants in two conditions c (0 vs 1). Thus, the effect of the independent variable may also vary across participants. The *effect variant model* is then updated as follows:

$$y_{cip} = a + a_p + a_i + \left(b_w + b_{wi} + b_{wp}\right)\cdot x_c + e_{cip} \tag{8}$$

where x_c is the level of the within-participant variable (0 vs 1), b_w represents the fixed effect of x, b_{wp} represents the random deviation of the effect for participant p, b_{wi} represents the random deviation of the effect for the item i. The random coefficients have variances τ_I^2 for the random slope across items and τ_P^2 for the random slopes across participants.

The *aggregated model* boils down to a paired-sample t-test, which is equivalent to a mixed model without random slopes across items (see Donnellan et al., 2025, Appendix 1):

$$y_{cip} = a + a_p + a_i + \left(b'_w + b_{wp}\right)\cdot x_c + e_{cip} \tag{9}$$

For the full random effects model (8), the standard error of the condition effect is:

$$se(b_w) = \sqrt{\frac{2\tau_I^2 + 2I\tau_P^2 + \sigma}{2PI}} \tag{10}$$

whereas the aggregated model (9), or equivalent the paired-sample t-test, results in:

[2] Donnellan et al. (2025) defined this model as *random item slope regression*. We prefer the term *effects variant model* because it is more general, and can be used also for models with multiple random slopes, or other classes of linear models, such as the logistic or the Poisson mixed models.

[3] The fixed effects estimates do not vary between models when the data is balanced.

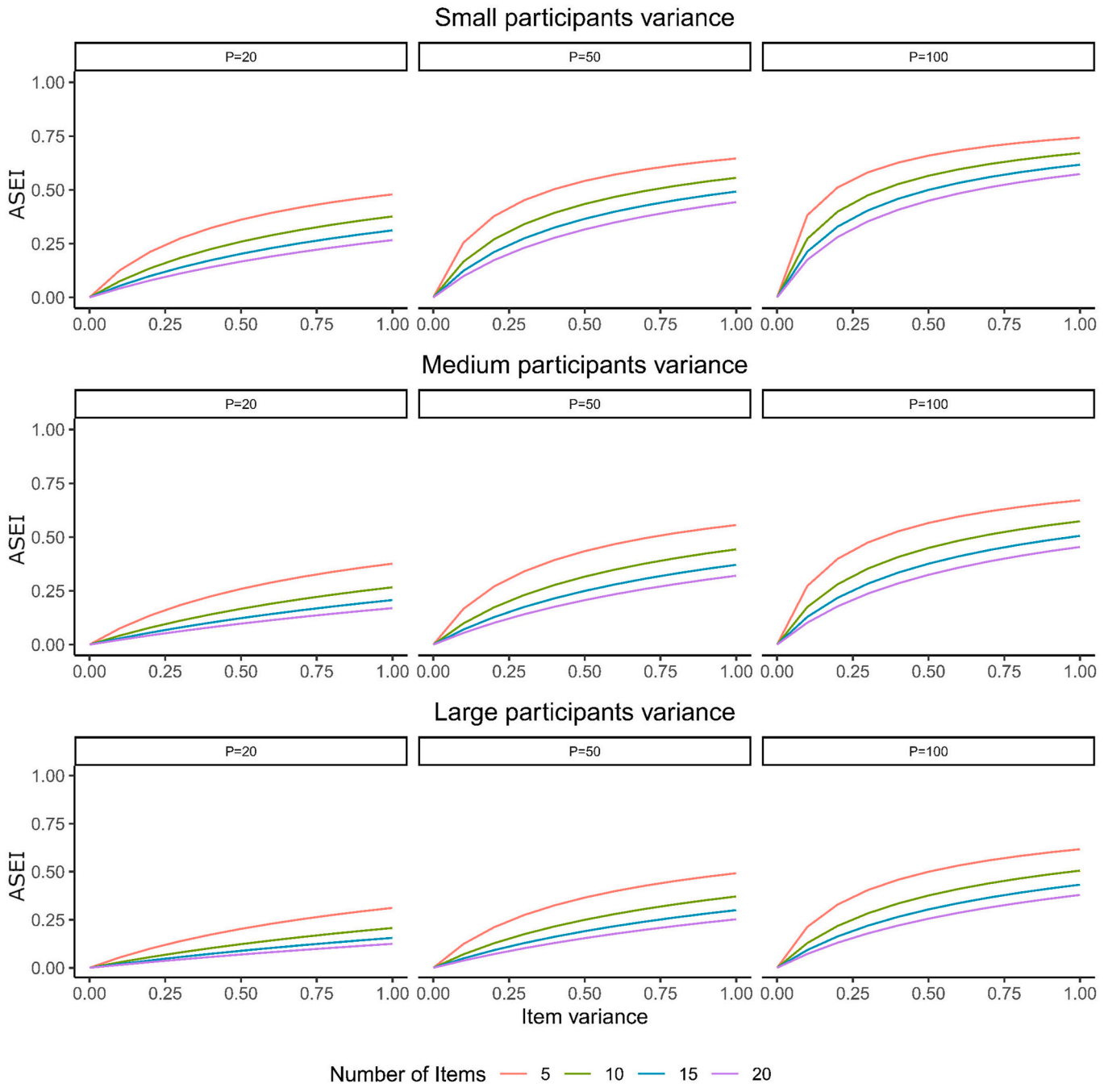

Fig. 1. Values of the $ASEI_b$ by size of the random effects τ_I^2, broken down by number of items, number of participants, and variance of participants intercepts (σ_P^2).

$$se(b_w') = \sqrt{\frac{\frac{\tau_I^2}{(P-1)} + 2I\tau_P^2 + \sigma}{2PI}} \tag{11}$$

Therefore, taking the ratio of the two standard errors, the *Aggregation Standard Error Inflation (ASEI)* for the repeated measure design is:

$$ASEI_w = 1 - \sqrt{\frac{\frac{\tau_I^2}{(P-1)} + 2I\tau_P^2 + \sigma}{2\tau_I^2 + 2I\tau_P^2 + \sigma}} \tag{12}$$

when $\tau_I^2 > 0$, $ASEI_w$ increases when the effect variation across participants is small (τ_P^2), all else equal. This means that if the independent variable's effect varies across items, error inflation is stronger when participants respond similarly to the manipulation. In fact, the more consistent and pervasive the manipulation effect, the greater the bias under standard aggregation. As in between-participant designs, standard errors are equals only when item slopes are absent ($\tau_I^2 = 0$). Fig. 2 illustrates how $ASEI_w$ changes with parameter variation.

From a practical point of view, the computation of the ASEI index entails comparing the standard error of the independent variable coefficient estimated with two mixed models with random intercepts across participants and items, one model with random slopes across items ($\widehat{se(b)}$) and one without random slopes across items ($\widehat{se(b')}$). The ratio of the two standard errors yields the index:

$$ASEI = 1 - \frac{\widehat{se(b')}}{\widehat{se(b)}} \tag{13}$$

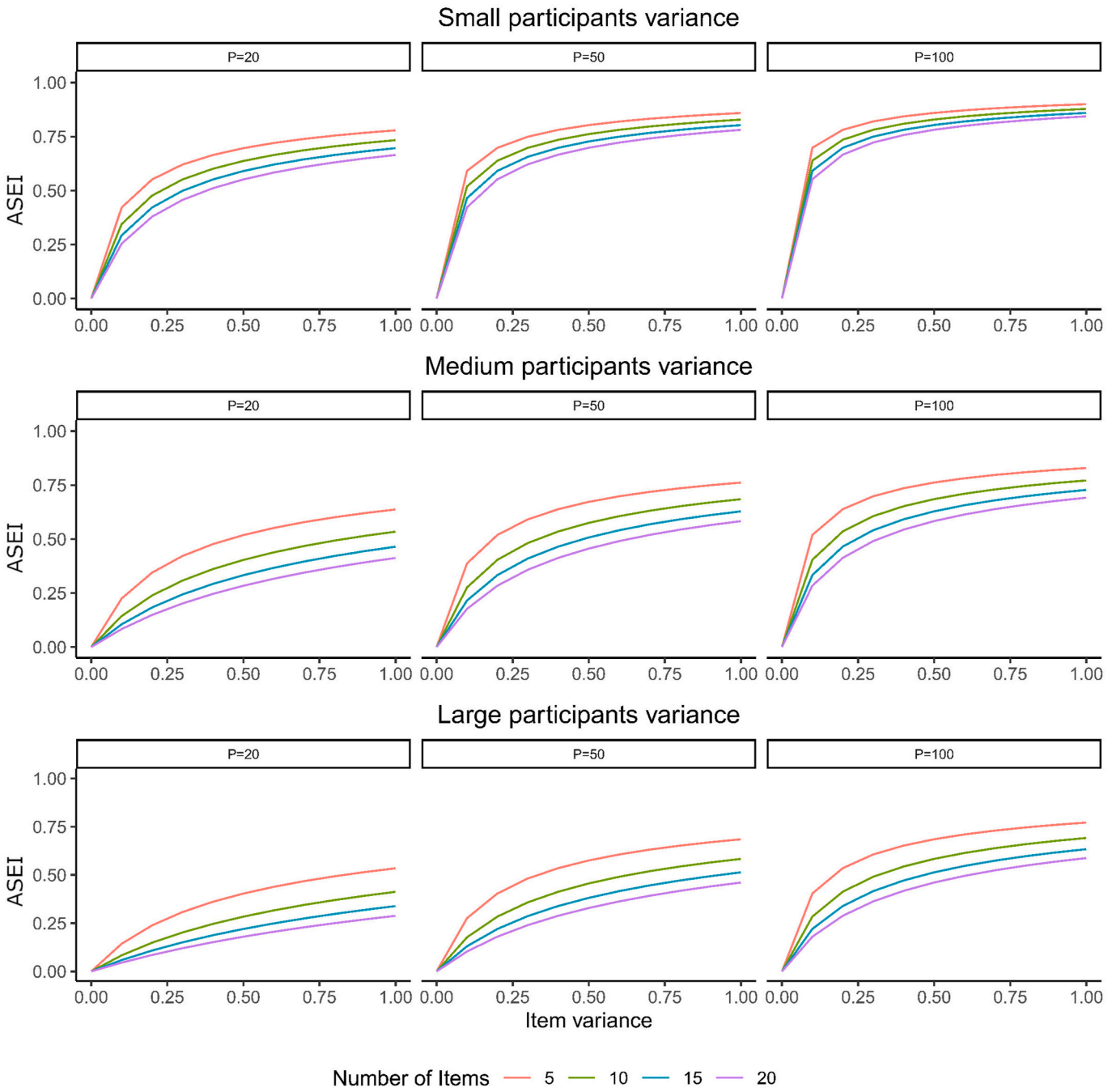

Fig. 2. Values of the $ASEI_w$ by size of the item variance (τ_I^2), broken down by number of items, number of participants, and variance of participants intercepts (σ_P^2). Repeated-measures case.

2. Case study

For the sake of demonstrating the validity of the proposed approach, we focus on studies where two conditions (i.e., experimental and control) are measured on the same participants via multi-item questionnaires (i.e., within-subjects design). This is a typical situation in experimental psychology where, besides performance-based measures, subjective responses are collected via questionnaires (Frisco et al., 2024). Moreover, this design is very frequent in studies evaluating intervention efficacy on psychological symptoms (e.g., Doorn et al., 2023; Harley et al., 2007; Lush et al., 2009). Nonetheless, the principles can be extended to more complex designs, such as longitudinal studies and mixed within-between designs. We focus on a within-participants design because it was not previously tackled by Donnellan et al. (2025) or Judd et al. (2012).

We initially present simulations comparing the inferential performance of the aggregated model, whose effect is tested with a *t*-test, and mixed-effects model under different conditions of effect size, sample size, number of items, and variability across items and participants. We address the Type I error rate when the null hypothesis is true and the Positive Predictive Value (*PPV*) when the alternative hypothesis is true. Then, we consider data from four experiments run in our laboratory (sample size range: 20 - 41) and compare the *t*-test and mixed-effects model results. Finally, we discuss the pros of interpreting random effects.

2.1. Data, materials, and online resources

Data and scripts are available on the Open Science Framework at the following link: https://osf.io/3guq9/?view_only=e066e516a2bf421d82fc48c8294c27a8. Supplemental Materials are available on the journal's website.

3. Simulations design

We compare the inferential performance of the mixed-effects model and the *t*-test in the following design: a set of I items are measured on P participants in two conditions ($cond = \{0,1\}$). Each item is presented in both conditions, and responses are measured at an interval level. The Data Generating Model (DGM) is the following:

$$y_{ip} = a_p + a_i + b_p \cdot cond + b_i \cdot cond + e_{ip} \quad (14)$$

where p refers to the participants $p = [1...P]$ and i to the item $i = [1...I]$. The DGM assumes the coefficients to be randomly drawn from normal distributions defined as follows:

- $a_p \sim N(\overline{a}_p, \sigma_P^2)$
- $a_i \sim N(\overline{a}_i, \sigma_I^2)$
- $b_p \sim N(\overline{b}_p, \tau_P^2)$
- $b_i \sim N(\overline{b}_i, \tau_I^2)$
- $e_{ip} \sim N(0,1)$

We test with simulations that the *t*-test is biased only when $ASEI_w > 0$ and that the error inflation is proportional to $ASEI_w$. See Supplemental Materials 1 for the model definition and the simulation parameters.

3.1. Methods

We simulate the performance of the *t*-test and the mixed model when the null hypothesis is true ($\overline{b} = 0$) and when the alternative hypothesis is true ($\overline{b} = 0.1, 0.5, 1, 2$). We simulate a DGM as described above (14).

A linear mixed model is estimated for each simulation sample, and the p-value of the inferential test associated with the fixed effect of the condition is recorded. The proportion of significant ($p. < .05$) effect is recorded across 1000 repetitions of the same model. In the same sample, the items are aggregated within the condition, and the two conditions' means are compared with a paired *t*-test. The proportion of significant paired t-tests is also recorded. Under the null hypothesis, the ratio of significant results represents the Type I error (i.e., false positives): the larger the proportion, the greater the error. Under the alternative hypothesis, the proportion of significant effects represents the proportion of true positives. The Positive Predictive Value (*PPV*) is then computed as:

$$PPV = \frac{TP}{TP + FP} \quad (15)$$

where *TP* are the true positives, and *FP* are the false positives. The *PPV* gives a value of how much a significant result should be "trusted". In each model the $ASEI_w$ is computed following (12).

3.2. Results

3.2.1. Type I error rate

Table 1 shows the proportion of significant effects at different values of slope variance across items and participants (i.e., 0, and 1.0). When no variance is present (i.e., $\tau_I^2 = 0$ and $\tau_P^2 = 0$), both methods show a low prevalence of significant effects (aggregated *t*-test range: 0.047-0.056; mixed model range: 0.014-0.026). Things change when we consider high slope variances across items and participants (i.e., $\tau_I^2 = 1$ and $\tau_P^2 =$ 1). The aggregated *t*-test shows high error rate (range: 0.209-0.602), which shrinks with an increasing number of items and grows with a larger number of participants. Crucially, the mixed model consistently shows low error with increasing slope variance across items (range: 0.040-0.065). Fig. 3 (panel A) shows the same pattern, with $ASEI_w$ broken down by number of items (I), participants (P), and slope variance across participants (SPV). $ASEI_w$ encompasses all sources of variation (i.e., number of items and participants and slope variance across items and participants), offering a single error inflation index. Fig. 3 (panel B) shows that the *t*-test produces an increasing proportion of false positives (Type 1 error) as $ASEI_w$ increases. On the contrary, the mixed model proportion of significant effect remains around zero, indicating that Type 1 error was not inflated.

Table 1
Proportion of significant effects when the slope variances across items and participants are set at 0 and 1.0, respectively.

Mixed model	Participant variance = 0			Participant variance = 1		
Item variance = 0	P20	P50	P80	P20	P50	P80
I5	0.014	0.016	0.017	0.033	0.032	0.035
I10	0.026	0.021	0.017	0.045	0.043	0.053
I15	0.024	0.025	0.025	0.043	0.048	0.045
Item variance = 1						
I5	0.035	0.051	0.058	0.054	0.062	0.057
I10	0.043	0.049	0.04	0.049	0.045	0.049
I15	0.047	0.05	0.04	0.047	0.04	0.065
Aggregated *t*-test						
Item variance = 0	P20	P50	P80	P20	P50	P80
I5	0.047	0.056	0.037	0.048	0.047	0.051
I10	0.053	0.049	0.038	0.046	0.049	0.058
I15	0.052	0.053	0.046	0.045	0.05	0.051
Item variance = 1						
I5	0.544	0.694	0.762	0.285	0.491	0.602
I10	0.528	0.689	0.754	0.209	0.373	0.493
I15	0.544	0.709	0.758	0.165	0.315	0.429

Note. The table shows the significant effect proportion at different sample sizes and numbers of items. Intercept variances across items and participants are set to 0.25.

3.2.2. Positive Predictive Value (PPV)

The left column of Fig. 4 shows the proportion of significant ($p < .05$) effects in the paired *t*-test (dashed lines) and the mixed model (solid lines) at different values of $ASEI_w$ and condition effects ($\overline{b} = 0.1, 0.5, 1, 2$). The mixed model has lower proportions of true positives than the aggregated *t*-test when there are small to medium differences between conditions (panels A, B, and C). When the condition effect grows (panel D), the mixed model always detects the effect. In general, the proportion of true positives detected by the mixed model tends to decrease with growing values of $ASEI_w$. In the aggregated *t*-test, the proportion of true positives grows with increasing $ASEI_w$ especially when there are small differences between conditions (panels A and B).

Interestingly, the perspective changes when considering the Positive Predictive Value (Fig. 2, right column). At minor differences between conditions (panel A), both methods show low *PPVs*. However, as the condition difference increases, the mixed model becomes progressively more trustworthy. At the same time, the *t*-test always shows decreasing *PPV* as the $ASEI_w$ increases. Table 2 shows *PPVs* when the variances across items and participants are 0.444 (i.e., the median value). The difference between methods increases as the effect size rises. Moreover, the mixed model *PPV* increases with the number of participants and items. Still, in the aggregated *t*-test, the *PPV* increases as items augment but shrinks as sample size increases, consistently with what we found in the H0 simulation.

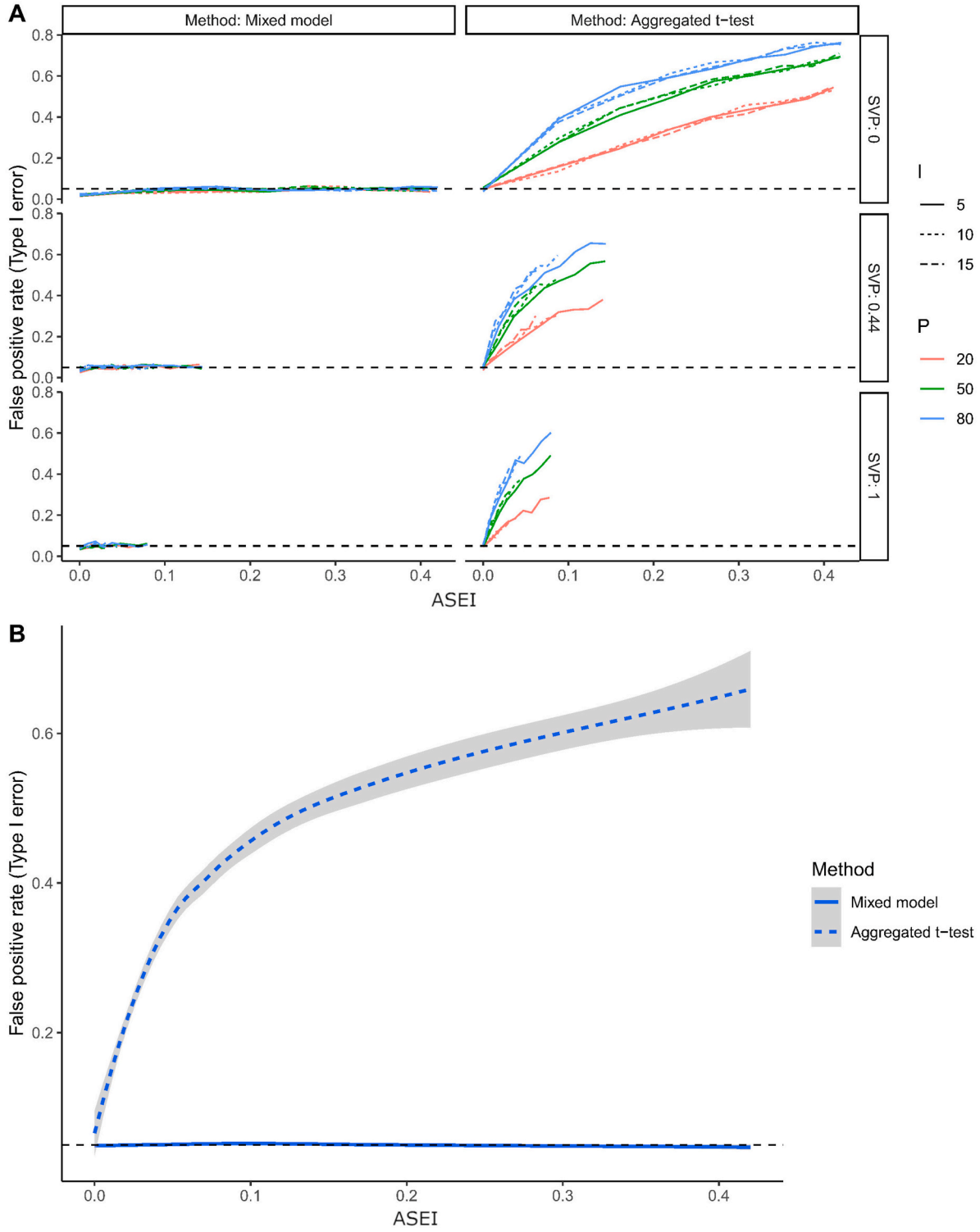

Fig. 3. Simulation results. Panel A shows the relationship between the proportion of significant ($p < .05$) effects and *ASEI* broken down by number of items (I), number of participants (P), and variance of participants intercepts (SVP). Panel B shows the proportion of significant ($p. < .05$) effects in the paired *t*-test and the mixed model at different values of *ASEI*. The line type represents the method used to perform the test. The black dashed line shows the conventional significance level of 0.05.

4. Real data analyses

4.1. Data description

We re-analysed data from four different studies run in our laboratory ((Tosi et al., 2020); Tosi et al., 2022; (Tosi et al., 2025)). In particular, we included participants from previous studies assessing the embodiment sensations after body illusion manipulations. All participants were exposed to two conditions in a within-subjects design. At the end of each condition, participants completed a six-item questionnaire (Q1 to Q6), which is the target of the current analysis. See Supplemental Materials 1 for the complete data description.

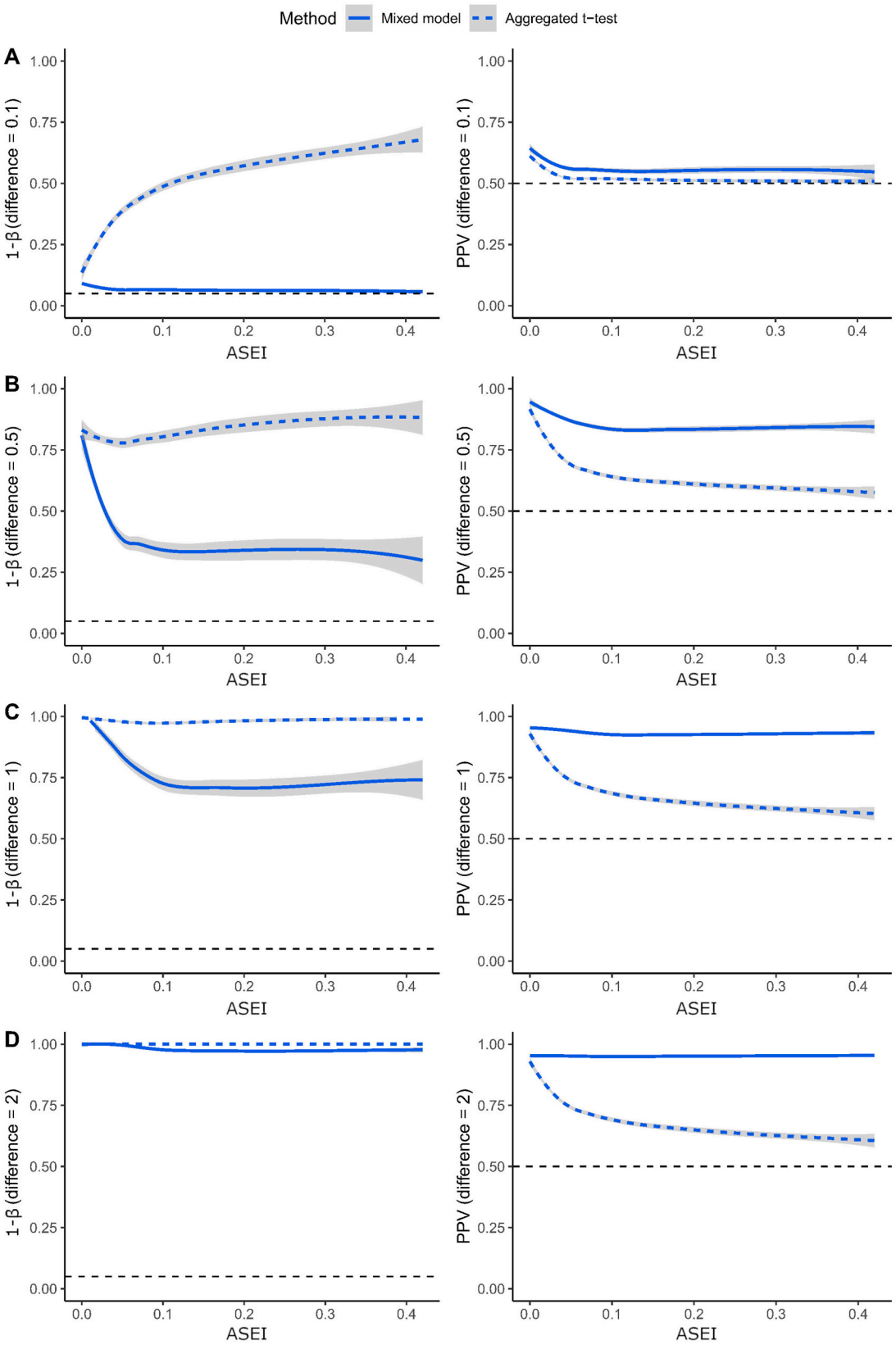

Fig. 4. Simulation results. The left column of the figure shows the proportion of significant ($p < .05$) effects in the paired *t*-test and the mixed model at different values of *ASEI*. The right column of the figure shows the Positive Predictive Value at different values of *ASEI*. The line type represents the method used to perform the test. Panel A: difference between condition = 0.1; panel B: difference between condition = 0.5; panel C: difference between condition = 1.0; panel D: difference between condition = 2.0. β = Type II error.

Table 2
Positive Predictive Value at different condition effects when the slope variances across items and participants are set to 0.44.

	Mixed model			Aggregated *t*-test		
	P20	P50	P80	P20	P50	P80
Condition difference = 0.1						
I5	0.509	0.496	0.531	0.48	0.516	0.522
I10	0.608	0.578	0.581	0.567	0.541	0.533
I15	0.518	0.62	0.652	0.518	0.549	0.557
Condition difference = 0.5						
I5	0.804	0.793	0.811	0.69	0.636	0.624
I10	0.912	0.909	0.908	0.797	0.73	0.684
I15	0.903	0.934	0.935	0.819	0.76	0.722
Condition difference = 1.0						
I5	0.92	0.916	0.919	0.78	0.694	0.661
I10	0.96	0.952	0.95	0.844	0.754	0.702
I15	0.946	0.956	0.953	0.861	0.776	0.733
Condition difference = 2.0						
I5	0.949	0.94	0.942	0.791	0.696	0.662
I10	0.963	0.953	0.951	0.847	0.754	0.702
I15	0.948	0.956	0.953	0.862	0.776	0.733

Note. The table shows the PPV for varying numbers of participants (P) and items (I). Intercept variances across items and participants are set to 0.25.

4.2. Methods

Typically, participants' responses to items Q1 to Q6 are averaged. The mean score is used as the dependent variable to assess the difference between the experimental (anatomical) and the control (non-anatomical) conditions. Here, we compare the *t*-test results from the averaged responses and the mixed model method. Item Q4 was reversed before the analyses.

Analyses were performed in r; the *t*-test was implemented using the *t.test* function, and the mixed model method was implemented using the *lme4::lmer* function.

4.3. Results

The comparison between the *t*-test and mixed model on real data shows that the approaches coincide on the same result in one out of four cases (Table 3). Specifically, a significant difference between experimental conditions was detected in Experiment 1 (*t*-test $p < .01$; mixed model $p < .05$), which also has small *ASEI* (*ASEI* = 0.03). When looking at Experiments 3 and 4, the *t*-test resulted in a significant difference ($p < .05$). Crucially, the mixed models show a tendency towards significance but do not reach it ($p > .06$). We can see a positive bias of the t-tests, which is confirmed by medium values of *ASEI* ($0.03 < ASEI < 0.05$). Experiment 2 shows a higher error inflation (*ASEI* = 0.10), indeed the result of the *t*-test is biased towards significance (*t*-test $p < .05$; mixed model $p = .25$). Moreover, the mixed models (conditional R^2 range = 0.49 – 0.61) always explain more variance than the t-tests (conditional R^2 range = 0.19 – 0.44). The complete results are reported in Supplemental Materials 1 (Table SM1).

If we reduce the random effect terms and compare the reduced model to the original one, we obtain significant effects of item random slope in all the experiments (Table 4). Significant effects of participant random slope emerge in experiments 1, 3, and 4. These results suggest that the experimental condition has different effects across individuals and items. Indeed, the correlations between participants' random slopes and the difference between participants' mean responses in the two conditions range between $r = 0.977$ and $r = 0.999$. Random coefficients are reported in Supplemental Materials 1 (Table SM2).

5. Discussion

Using multiple questions to measure a construct increases reliability by capturing different facets, but averaging responses overlooks the variability across items. This practice is common in experimental psychology, such as studies on embodiment and intervention effectiveness (Botvinick and Cohen, 1998; Doorn et al., 2023; Lush et al., 2009; Serino et al., 2023; Tosi et al., 2022). However, ignoring the random selection of items assumes *effects invariance* (i.e., that variable effects do not differ across stimuli), which, like ignoring participant variability, biases F-tests and inflates Type I error (Clark, 1973; Judd et al., 2012; Locker et al., 2007). Psycholinguistic and social psychology research recommends mixed models treating both participants and items as random factors, which corrects this bias (Clark, 1973; Judd et al., 2012; Locker et al., 2007). Building on this, Donnellan et al. (2025) proposed the idea of *effect variance* as a general measurement model, showing that aggregation underestimates standard errors and increases Type I error when random effects exist. We extend this by introducing the *Aggregation Standard Error Inflation (ASEI)*, the ratio of standard errors from aggregation and mixed models, to quantify error inflation. Focusing on within-participant designs and small samples typical in psychology, we demonstrate that *ASEI* relates to inflated Type I error, and that mixed models yield more reliable, significant results than aggregation.

Table 3
Comparison between paired *t*-test and mixed model in real data.

	Aggregated *t*-test						Mixed model					
	estimate	95% CI	t	df	p		estimate	95% CI	t	df	p	ASEI
Experiment 1						(intercept)	0.15	−1.00; 1.30	0.28	12.84	0.78	
Orientation	1.28	[0.59; 1.97]	3.89	19.00	<0.01**	Orientation	1.28	[0.29; 2.28]	2.82	11.96	<0.05*	0.03
R^2	0.44					marginal R^2	0.09					
						conditional R^2	0.61					
Experiment 2						(intercept)	−0.09	−1.11; 0.93	−0.20	7.58	0.84	
Orientation	0.44	[0.04; 0.83]	2.29	23.00	<0.05*	Orientation	0.44	[-0.39; 1.27]	1.25	6.78	0.25	0.10
R^2	0.19					marginal R^{-2}	0.01					
						conditional R^2	0.53					
Experiment 3						(intercept)	0.16	−1.08; 1.40	0.29	8.56	0.78	
Orientation	0.88	[0.33; 1.42]	3.30	23.00	<0.05*	Orientation	0.88	[-0.16; 1.91]	1.90	9.15	0.09	0.05
R^2	0.32					marginal R^2	0.05					
						conditional R^2	0.59					
Experiment 4						(intercept)	0.17	−0.71; 1.05	0.43	9.91	0.68	
Orientation	0.74	[0.22; 1.15]	3.63	40.00	<0.01**	Orientation	0.74	[-0.03; 1.50]	2.17	9.12	0.06	0.03
R^2	0.25					marginal R^2	0.03					
						conditional R^2	0.49					

Note. The table shows the variance explained (R^2) and the condition effect resulting from the *t*-test and the mixed model. Significance code: * = p < .05; ** = p < .01; *** = p < .001.

Table 4
Likelihood Ratio tests.

	Npar	logLik	AIC	LRT	df	Pr(>Chisq)
Experiment 1						
full model	9	−456.76	931.51			
orientation in (orientation|item)	7	−467.52	937.64	10.13	2	<0.01**
orientation in (orientation|id)	7	−467.52	949.05	21.53	2	<0.001***
Experiment 2						
full model	9	−535.45	1088.90			
orientation in (orientation|item)	7	−541.80	1097.60	12.70	2	<0.01**
orientation in (orientation|id)	7	−535.96	1085.90	1.03	2	0.06
Experiment 3						
full model	9	−531.57	1081.20			
orientation in (orientation|item)	7	−542.29	1098.60	21.43	2	<0.001***
orientation in (orientation|id)	7	−541.51	1097.00	19.87	2	<0.001***
Experiment 4						
full model	9	−948.53	1915.10			
orientation in (orientation|item)	7	−956.14	1926.30	15.21	2	<0.001***
orientation in (orientation|id)	7	−958.79	1931.60	20.52	2	<0.001***

Note. The table compares the original model and models where each random effect term is reduced. Npar = number of parameters; logLik = log-likelihood for the model; AIC = AIC for the model; LRT = likelihood ratio test statistic; df = degrees of freedom for the likelihood ratio test; Pr(>Chisq) = p-value. Significance code: * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

The first simulation shows that the *t*-test becomes increasingly biased (higher Type I error) as the *ASEI* grows, reflecting, above all, the item slope variance (i.e., the variation across items of the independent variable effect), the number of items, and the number of participants. Error rates rise when the independent variable effect varies across items, especially if participants show similar sensitivity to the independent variable (i.e., low slope variance across participants). When effects vary across both items and participants, *ASEI* decreases with more items but increases with more participants, as larger samples amplify distortions from aggregation. This pattern, consistent with Judd et al. (2012) and Donnellan et al. (2025), highlights that aggregating items can worsen the error inflation with bigger samples. In contrast, mixed models maintain Type I error near the nominal alpha regardless of slope variance, item count, or sample size.

In the second set of simulations, we tested four effect sizes to evaluate true positive rates and Positive Predictive Value (*PPV*). Looking at the proportions of true positives, the mixed model appears underpowered compared to the aggregated *t*-test, as it accounts for variability across items and participants. Indeed, the proportion of true positives decreases with growing *ASEI*. For the *t*-test, the relationship between true positives and *ASEI* reverses from small to large effects: with small effects, power grows with *ASEI*, suggesting greater sensitivity. However, this sensitivity is misleading because it stems from biased estimates. *PPV* (i.e., the ratio between true positives and the sum of true and false positives) clarifies this result: at small effects, both methods show low precision (i.e., low *PPV*), but as effect size increases, the mixed model becomes progressively more accurate than the *t*-test. The mixed model *PPV* also rises with sample size and item number, indicating that testing more participants on multiple items improves the statistical power. Conversely, the *t*-test *PPV* increases with item number but declines with larger samples, consistent with the pattern observed in the first set of simulations and by Judd et al. (2012). Overall, while aggregated t-tests yield more significant results, these are less trustworthy than those from mixed models, which provide fewer but more valid findings.

When we test the mixed model on real data, the results show that this approach explains more variance than the *t*-test in small samples. Moreover, a *t*-test positive bias emerges in three cases (Experiments 2, 3, and 4). The higher the difference between the models' results, the higher the result *ASEI* value, in line with the first simulation results. Combined, simulations and real data point out that when the independent variable affects the items (i.e., high variation in item slope) and the measure is less reliable (i.e., high variation in participant intercepts), the *t*-test erroneously attributes the variance. On the contrary, the mixed model controls for different sources of variation and reliably estimates their effects. A secondary benefit of not aggregating items is the attenuation bias mitigation (Gilbert, 2025). When aggregated scores are standardized, the effect coefficient is attenuated due to measurement error (Hedges, 1981; Shear and Briggs, 2024). Not aggregating items into a single score or classical errors-in-variables correction can remove this bias (Gilbert, 2025).

Another important point is that random coefficients provide additional insights into experimental results. Item random slopes can quantify variability in the observed condition effect across items. For instance, our data showed that the Q5 item (i.e., Did it seem like the legs in the video were in the location where your legs were?) has the highest random coefficients across all experiments, suggesting that it is the most sensitive to the orientation difference between conditions. This variability might indicate differences in item sensitivity, content, or measurement features (like factor loadings). Indeed, in a latent-variable framework, differences in observed condition effects across items are an inevitable consequence of heterogeneous factor loadings. However, previous studies have already noted that the random slope is not due to model misspecification, as the item variance remains similar across models with equal and variable factor loadings, and SEM models' performance with equal factor loadings is nearly identical to that allowing variable loadings (Gilbert et al., 2025). Additionally, item slope variance could be influenced by systematic factors and associated with item covariates, such as how the item aligns with the intervention or condition (Kim et al., 2023). Whether the item variability originates from differences in factor loadings, item content, measurement characteristics, or other item-specific properties is not essential for the inferential goal of the model. The mixed-effects model remains appropriate even when factor loadings differ across items, as it explicitly models heterogeneity in observed effects.

To examine this, we generated data from a congeneric one-factor model, in which the experimental condition influences a single latent variable, and each item reflects that latent variable through its own loading. Under this structure, an item's response to the condition is the underlying latent effect passed through that item's loading, so items with stronger loadings necessarily show larger condition effects. So, even when the dependency between loadings and item effects is substantial, the validity of the effect-variant model is not compromised. The spread of the item-specific effects is jointly determined by the magnitude of the latent effect and the variability of the loadings: the item effects can differ

only to the extent that there is a genuine effect for the loadings to scale. When the latent effect is absent, every item's effect is zero, no matter how widely the loadings vary, so loading heterogeneity alone cannot generate variance across the item slopes and, consequently, cannot produce false positives. Item-slope variance arises only when a real effect is present, and in that case, it reflects authentic differences in how strongly the manipulation registers on each item, which is exactly the variability the effect-variant model is designed to accommodate. Our simulations confirmed this reasoning: when the loadings varied but no effect was present, both methods controlled the error rate, and the estimated item-slope variance was essentially zero, whereas when a real effect was scaled by heterogeneous loadings, the aggregated *t*-test grew increasingly anticonservative as the sample size increased while the effect-variant model retained the nominal Type I error rate throughout (Supplemental Materials 2, Fig. SM1). The safeguard the model provides, therefore, does not depend on the data literally arising from a random-slope process; it holds even when item heterogeneity originates in the factor-loading structure, whereas aggregation offers no such protection.

Identifying item heterogeneity can inform questionnaire design and interpretation by highlighting which items contribute more strongly to the observed condition effect. This perspective relates to Differential Item Functioning (*DIF*), which assesses whether items on a questionnaire perform differently for different groups of individuals, even if they have the same level of trait being studied (Woods, 2009). In particular, the Multiple Indicators, Multiple Causes (*MIMIC*) approach uses Structural Equation Models (*SEM*) to measure *DIF* by regressing the item response onto a grouping variable (Finch, 2005). *DIF* undercover situations where respondents at the same level of the latent trait have different probabilities of answering a specific item correctly based on their group membership. Similarly, the random slopes capture differences in the condition effect across items. While technically a type of *DIF*, our model has a different purpose. Unlike *DIF*, which usually aims to detect item bias while controlling for the latent trait, our approach focuses on modeling variability in observed responses across items and participants in experimental settings. Similarly, participants' random slopes capture individual differences in how they respond to the experimental manipulations, offering valuable information about participants' responsiveness to the manipulation and differentiating responders from non-responders.

In contrast, random intercepts capture consistency across participants and questionnaire items. Specifically, an item's random intercept reflects its expected average score across participants, while a participant's intercept represents their overall mean across all items. This distinction from random slopes is important because it allows us to understand not only the overall treatment effect but also how individual participants deviate from the average response pattern. Since the variances of these intercepts correspond to variability in scores across participants and items, they can be used to compute reliability measures such as the intraclass correlation coefficient (*ICC*; i.e., the proportion of variance explained by intercept variability) or Cronbach's alpha (see (2)). By estimating reliability within a mixed-model framework, we gain a more detailed picture of variance decomposition, identifying systematic sources of variation and improving the precision of our measurement model.

This work advances the study of *effect variant models* by introducing the *Aggregation Standard Error Inflation (ASEI)*, a psychometric measure of error inflation from item aggregation when random effects are ignored. We show that *ASEI* relates to increased Type I error and reduced *PPV*, with larger samples amplifying this error. Beyond controlling random variance, we propose using random effects to assess reliability, participant responsiveness, and item variability. Our findings demonstrate that when effects vary across items, mixed models outperform aggregated t-tests by avoiding F-test bias, improving accuracy, and providing richer insights into participants and the questionnaire structure. We recommend using the *ASEI* to quantify potential error inflation from aggregation.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Giorgia Tosi: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Marcello Gallucci:** Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Daniele Romano:** Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.metip.2026.100275.

Data availability

Data and scripts are available on the Open Science Framework at the following link: https://osf.io/3guq9/?view_only=e066e516a2bf421d82fc48c8294c27a8.

References

Ahmed, I., Bertling, M., Zhang, L., Ho, A.D., Loyalka, P., Xue, H., Rozelle, S., Benjamin, W., 2025. Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. J. Res. Educational. Effect. 18 (4), 854–877. https://doi.org/10.1080/19345747.2024.2361337.

Algina, J., Penfield, R.D., 2009. Classical Test Theory. *the Sage Handbook of Quantitative Methods in Psychology*, pp. 93–122.

Angoff, W.H., 1981. Use of difficulty and discrimination indices for detecting item bias. In: Berk, R.A. (Ed.), Handbook of Methods for Detecting Test Bias. The Johns Hopkins Press, pp. 96–116.

Bogdan, P.C., 2025. One decade into the replication crisis, how have psychological results changed? Adv. Methods Pract. Psychol. Sci. 8 (2), 25152459251323480.

Botvinick, M., Cohen, J., 1998. Rubber hands 'feel' touch that eyes see. Nature 391 (6669), 756–756.

Brennan, R.L., 2001. Variability of statistics in generalizability theory. Generalizability Theor. 179–213. https://doi.org/10.1007/978-1-4757-3456-0_6.

Clark, H.H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. J. Verb. Learn. Verb. Behav. 12 (4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16 (3), 297–334. https://doi.org/10.1007/BF02310555.

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

Donnellan, E., Usami, S., Murayama, K., 2025. Random item slope regression: an alternative measurement model that accounts for both similarities and differences in association with individual items. Psychol. Methods 30 (4), 744–769. https://doi.org/10.1037/met0000587.

others Doorn, M. van, Monsanto, A., Wang, C.L., Verfaillie, S.C., Amelsvoort, T. A. van, Popma, A., et al., 2023. The effects of a digital, transdiagnostic, clinically and peer-moderated treatment platform for young people with emerging mental health complaints: repeated measures within-subjects study. JMIR mHealth uHealth 11 (1), e50636. https://doi.org/10.2196/50636.

Finch, H., 2005. The MIMIC model as a method for detecting DIF: comparison with mantel-haenszel, SIBTEST, and the IRT likelihood ratio. Appl. Psychol. Meas. 29 (4), 278–295. https://doi.org/10.1177/0146621605275.

Frisco, F., Bruno, V., Romano, D., Tosi, G., 2024. I am where i believe my body is: the interplay between body spatial prediction and body ownership. PLoS One 19 (12), e0314271. https://doi.org/10.1371/journal.pone.0314271.

Gilbert, J.B., 2025. How measurement affects causal inference: attenuation bias is (usually) more important than outcome scoring weights. Methodology 21 (2), 91–122. https://doi.org/10.5964/meth.15773.

Gilbert, J.B., Himmelsbach, Z., Soland, J., Joshi, M., Domingue, B.W., 2025. Estimating heterogeneous treatment effects with item-level outcome data: insights from item

response theory. J. Pol. Anal. Manag. 44 (4), 1417–1449. https://doi.org/10.1002/pam.70025.
Gilbert, J.B., Kim, J.S., Miratrix, L.W., 2023. Modeling item-level heterogeneous treatment effects with the explanatory item response model: leveraging large-scale online assessments to pinpoint the impact of educational interventions. J. Educ. Behav. Stat. 48 (6), 889–913. https://doi.org/10.3102/10769986231171710.
Green, B.F., Tukey, J.W., 1960. Complex analyses of variance: general problems. Psychometrika 25 (2), 127–152. https://doi.org/10.1371/journal.pone.0314271.
Halpin, P., Gilbert, J., 2024. Testing whether reported treatment effects are unduly influenced by item-level heterogeneity. Retrieved from. https://osf.io/preprints/psyarxiv/9ru45v1. (Accessed 15 September 2024).
Harley, R.M., Baity, M.R., Blais, M.A., Jacobo, M.C., 2007. Use of dialectical behavior therapy skills training for borderline personality disorder in a naturalistic setting. Psychother. Res. 17 (3), 351–358. https://doi.org/10.1080/10503300600830710.
Hedges, L.V., 1981. Distribution theory for Glass's estimator of effect size and related estimators. J. Educ. Stat. 6 (2), 107–128. https://doi.org/10.3102/10769986006002107.
Holland, P.W., Thayer, D.T., 1986. Differential item functioning and the mantel-haenszel procedure. ETS Res. Rep.Ser. 1986 (2), 1–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x.
Holland, P.W., Wainer, H., 2012. Differential Item Functioning. Routledge.
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. J. Personality Soc. Psychol. 103 (1), 54. https://doi.org/10.1037/a0028347.
Kim, J.S., Burkhauser, M.A., Relyea, J.E., Gilbert, J.B., Scherer, E., Fitzgerald, J., Mosher, D., McIntyre, J., 2023. A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: transfer effects on students' reading comprehension. J. Educ. Psychol. 115 (1), 73–98. https://doi.org/10.1037/edu0000751.
Locker, L., Hoffman, L., Bovaird, J.A., 2007. On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. Behav. Res. Methods 39, 723–730. https://doi.org/10.3758/BF03192962.
Longo, M.R., Schüür, F., Kammers, M.P., Tsakiris, M., Haggard, P., 2008. What is embodiment? A psychometric approach. Cognition 107 (3), 978–998.
Lush, E., Salmon, P., Floyd, A., Studts, J.L., Weissbecker, I., Sephton, S.E., 2009. Mindfulness meditation for symptom reduction in fibromyalgia: psychophysiological correlates. J. Clin. Psychol. Med. Settings 16, 200–207. https://doi.org/10.1007/s10880-009-9153-z.
McDonald, R.P., 2013. Test Theory: a Unified Treatment. psychology press. https://doi.org/10.4324/9781410601087.
Mehta, P.D., Neale, M.C., 2005. People are variables too: multilevel structural equations modeling. Psychol. Methods 10 (3), 259. https://doi.org/10.1037/1082-989X.10.3.259.
Meredith, W., 1993. Measurement invariance, factor analysis and factorial invariance. Psychometrika 58 (4), 525–543. https://doi.org/10.1007/BF02294825.
Miratrix, L.W., Weiss, M.J., Henderson, B., 2021. An applied researcher's guide to estimating effects from multisite individually randomized trials: estimands, estimators, and estimates. J. Res. Educational. Effect. 14 (1), 270–308. https://doi.org/10.1080/19345747.2020.1831115.
Rijmen, F., Tuerlinckx, F., De Boeck, P., Kuppens, P., 2003. A nonlinear mixed model framework for item response theory. Psychol. Methods 8 (2), 185. https://doi.org/10.1037/1082-989X.8.2.185.
Romano, D., Maravita, A., Perugini, M., 2021. Psychometric properties of the embodiment scale for the rubber hand illusion and its relation with individual differences. Sci. Rep. 11 (1), 5029.
Serino, S., Sansoni, M., Di Lernia, D., Parisi, A., Tuena, C., Riva, G., 2023. 360-degree video-based body-ownership illusion for inducing embodiment: development and feasibility results. Virtual Real. 27 (3), 2665–2672. https://doi.org/10.1007/s10055-023-00836-6.
Shavelson, R.J., Webb, N.M., 1981. Generalizability theory: 1973–1980. Br. J. Math. Stat. Psychol. 34 (2), 133–166. https://doi.org/10.1111/j.2044-8317.1981.tb00625.x.
Shear, B.R., Briggs, D.C., 2024. Measurement issues in causal inference. Asia Pac. Educ. Rev. 25, 719–731. https://doi.org/10.1007/s12564-024-09942-9.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420.
Tosi, G., Maravita, A., Romano, D., 2022. I am the metre: the representation of one's body size affects the perception of tactile distances on the body. Q. J. Exp. Psychol. 75 (4), 583–597. https://doi.org/10.1177/17470218211044488.
Tosi, Giorgia, Frisco, Francesca, Maravita, Angelo, Romano, Daniele, 2025. The exposure to body size distortions affects allocentric distance perception in extra-personal space. Cortex 185, 50–63. https://doi.org/10.1016/j.cortex.2025.01.004. https://linkinghub.elsevier.com/retrieve/pii/S001094522500019X.
Tosi, Giorgia, Parmar, Jassleen, Dhillon, Inderpreet, Maravita, Angelo, Iaria, Giuseppe, 2020. Body illusion and affordances: the influence of body representation on a walking imagery task in virtual reality. Experimental Brain Research 238 (10), 2125–2136. https://doi.org/10.1007/s00221-020-05874-z. https://link.springer.com/10.1007/s00221-020-05874-z. (Accessed 13 July 2020).
Winer, B.J., 1962. Statistical principles in experimental design. https://doi.org/10.2307/3172747.
Woods, C.M., 2009. Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. Multivariate Behav. Res. 44 (1), 1–27. https://doi.org/10.1080/00273170802620121.
Yarkoni, T., 2022. The generalizability crisis. Behav. Brain Sci. 45, e1. https://doi.org/10.1017/S0140525X20001685.