Type-1 error inflation in the traditional by-participant analysis to metamemory accuracy: a generalized mixed-effects model perspective

Article

Accepted Version

Type-1 error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective

Kou Murayama Michiko Sakaki
School of Psychology and Clinical Language Sciences, University of Reading, UK

Veronica X. Yan
Department of Psychology, University of California, Los Angeles

Garry M. Smith
School of Systems Engineering, University of Reading, UK.

Accepted by Journal of Experimental Psychology: Learning, Memory, & Cognition

Correspondence address:
Kou Murayama
School of Psychology and Clinical Language Sciences
University of Reading
Earley Gate, Whiteknights, Reading RG6 6AL, UK
Tel: +44 (0)118 378 5558 Email: k.murayama@reading.ac.uk

Author notes

Kou Murayama and Michiko Sakaki, School of Psychology and Clinical Language Sciences, University of Reading, Reading, UK; Veronica X. Yan, Department of Psychology, University of California, Los Angeles; Garry M. Smith, School of Systems Engineering, University of Reading, Reading, UK. This research was partly supported by the Marie Curie Career Integration Grant (CIG618600; to Michiko Sakaki) and the EPSRC NeuroCloud project (via Garry Smith) under grant EP/I016856/1. We gratefully acknowledge the support of Japan Society for the Promotion of Science (research fellowship of Kou Murayama). We thank Alan Castel and other members of CogFog for helpful comments regarding this research. Correspondence concerning this article should be addressed to Kou Murayama, School of Psychology and Clinical Language Sciences, University of Reading, Earley Gate, Whiteknights, Reading, UK, RG6 6AL. E-mail: k.murayama@reading.ac.uk.

Abstract

In order to examine metacognitive accuracy (i.e., the relationship between metacognitive judgment and memory performance), researchers often rely on by-participant analysis, where metacognitive accuracy (e.g., resolution, as measured by the gamma coefficient or signal detection measures) is computed for each participant and the computed values are entered into group-level statistical tests such as the *t*-test. In the current work, we argue that the by-participant analysis, regardless of the accuracy measurements used, would produce a substantial inflation of Type-1 error rates, when a random item effect is present. A mixed-effects model is proposed as a way to effectively address the issue, and our simulation studies examining Type-1 error rates indeed showed superior performance of mixed-effects model analysis as compared to the conventional by-participant analysis. We also present real data applications to illustrate further strengths of mixed-effects model analysis. Our findings imply that caution is needed when using the by-participant analysis, and recommend the mixed-effects model analysis.

Key words: metamemory, resolution, multilevel model, generalized linear model, signal detection theory, type 1 error

While people often believe that they can accurately predict their later memory performance, a large body of research indicates that this is often not the case (e.g., Schwartz, Benjamin, & Bjork, 1997; Bjork, Dunlosky, & Kornell, 2013). To examine the quality of people's memory predictions, research on metamemory typically examines relative accuracy of metacognitive monitoring (i.e., resolution)—the degree to which a person's metamemory judgments (e.g., judgments of learning; feeling of knowing, etc.) are associated with the actual likelihood of correct recall of to-be-remembered items (Nelson & Narens, 1990). That is, resolution quantifies whether those items with higher judgments are more likely to be recalled on a later test than those items with lower judgments.

To date, a number of indices have been proposed to measure the relative accuracy of a learner's metamemory judgments (Benjamin & Diaz, 2008; Gonzalez & Nelson, 1996; Maniscalco & Lau, 2012; Masson & Rotello, 2009; Rotello, Masson, & Verde, 2008; Schraw, 1995), although the gamma correlation is the most popular and widely used in the literature (Nelson, 1984). Each of the proposed measures has both strengths and weaknesses, but one a common feature is that they all focus on estimating relative accuracy for *individual* participants. In contrast, surprisingly little attention has been paid to the *group-level* statistical inference of relative accuracy, despite the fact that making group-level inferences is usually the primary concern in empirical research (e.g., "Is the mean relative accuracy significantly different from chance?" or "Does the mean relative accuracy statistically differ between conditions?").

The current article aims to address this under-examined issue of group-level analysis of relative accuracy measurement in metamemory research. The organization of the article is as follows: First, we begin with a brief review of relative accuracy measures proposed in previous literature, and describe how past studies typically make a group-level inferences with these measures. We call these traditional, widely-used approaches "by-participant analyses". Second, we argue that, regardless of the specific relative accuracy measurement used, the traditional by-participant analysis to make group-level inferences could inflate Type-1 error rate. This inflation of Type-1 error rate is produced by the presence of random item effects (a topic that will be discussed later in the paper) in memory performance, which are common in standard metamemory research. Third, we present a mixed-effects model analysis to effectively resolve the inflation of Type-1 error rate. Mixed-effects modeling (see Baayen, Davidson, & Bates, 2008; Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006; Raudenbush & Bryk, 2002; Searle, Casella, & McCulloch, 1992) has become increasingly popular in various fields of psychology, but remains less common within experimental psychology. Our work serves as one of the first attempts to highlight the importance and usefulness of this approach in the context of metamemory research. Fourth, we report a set of simulation studies with a variety of experimental designs to quantify the danger of the traditional by-participant analysis—Type 1 error rate inflation, sometimes to an extraordinary degree (e.g., 30%)—and to illustrate the effectiveness of the mixed-effects model analysis. Finally, we apply mixed-effects modeling to actual data to further clarify the strengths of this approach.

Measuring Relative Accuracy of Metamemory Monitoring

In a typical metamemory experiment, participants study and make metamemory judgments (e.g., judgments of learning; JOLs) for each to-be-remembered item during the course of studying, and are later asked to recall these items. Researchers then obtain relative accuracy measures to quantify the degree to which the participants' metamemory judgments (i.e., the predictor variable) predict their actual recall performance (i.e., the criterion variable). While metamemory judgments are typically continuous (e.g., rate your likelihood of later recall this item on a scale of 0 to 100%) or ordinal polychotomous (e.g.,

rate the likelihood of later recall on a scale of 1 to 7), memory performance is mostly dichotomous (i.e., the item is either recalled or not). Different numerical measurements have been proposed to summarize relative accuracy, and fall, broadly speaking, into one of three families: Gamma, signal detection, and correlation.

Gamma Family

In a very influential paper, Nelson (1984) proposed that the gamma coefficient (*G*; Goodman & Kruskal, 1954) as the best measurement of relative accuracy (see also Nelson, 1986, 1996; Schwartz & Metcalfe, 1994). The gamma coefficient is a nonparametric measurement of association between two binary or ordered variables. To compute the gamma coefficient from a set of *K* items, all $K(K\text{-}1)/2$ item pairs are categorized into 1) concordant pairs, where the ordering between the two items on the predictor variable is consistent with the ordering of the same two items on the criterion variable (i.e., when a recalled item is given a higher judgment than an unrecalled item), 2) disconcordant pairs, where the ordering between the two items on the predictor variable is opposite from the ordering of the same two items on the criterion variable, and 3) tied pairs, where the ordering on the predictor or the criterion variable is tied (i.e., the two items have the same value either on the predictor, the criterion variable, or both). Designating the total numbers of concordant item pairs, disconcordant pairs, and tied pairs as *C*, *D*, and *T* [i.e., $K(K\text{-}1)/2 = C + D + T$], *G* is defined as

$$G = \frac{(C - D)}{(C + D)} \tag{1}$$

G, like Pearson's correlation coefficients, takes on values between -1 and 1, with values larger than zero indicating a positive association between the predictor and criterion variables. Nelson (1984) made several arguments to support the effectiveness of *G* in assessing the relative accuracy of metamemory judgments. Since then, *G* has been by far the most commonly used measurement in metamemory research, although not as frequently used in other fields.

One feature of Equation 1 is that the tied pairs are completely discarded from the computation of *G*. Some other researchers have proposed to correct the computation of *G* for the presence of ties (for reviews, L. C. Freeman, 1986; Gonzalez & Nelson, 1996). Specifically, Kim (1971) suggested adding the number of pairs that are tied only in the predictor variable to the denominator of Equation 1. On the other hand, Wilson (1974) proposed to add the number of pairs that are tied only in the criterion variable to the denominator. Somers (1968) alternatively suggested adding the number of pairs that are tied only in the predictor variable or only in the criterion variable (but not in both) to the denominator. In this paper, these alternative measures are denoted as G_k, G_w, and G_s, respectively (see Gonzalez and Nelson, 1996, for situations where the alternative measures are preferred).

Benjamin and Diaz (2008) noted that *G* suffers from interval-level inferences because of the boundaries at -1 and 1. That is, they indicated that intervals close to the boundary (e.g., $G = 0.98$ and $G = 0.99$) and intervals further from the boundary (e.g.,$G = 0.50$ and $G = 0.51$) do not have the same meaning. To address this problem, Benjamin and Diaz (2008) proposed an alternative measurement called *G**.

$$G^* = \log\frac{(G + 1)}{(1 - G)} \tag{2}$$

*G** is a simple transformation of *G,* but does not have lower or upper boundaries.

Signal Detection Family

In cases where metamemory judgments have two categories (typical, for example, in research on the "feeling of knowing"; Hart, 1965), the resulting data can be summarized

with two numbers: The hit rate (*H*), which is the probability of saying "yes" to the items that are later recalled, and the false alarm rate (*F*), which is the probability of saying "yes" to the items that are later forgotten. This data structure fits well with signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005). Indeed, several researchers have recommended the application of SDT to assess relative accuracy of metamemory judgments (Benjamin & Diaz, 2008; Masson & Rotello, 2009; Rotello et al., 2008; Swets, 1986; see also (Barrett, Dienes, & Seth, in press; Higham, 2007; Maniscalco & Lau, 2012). The SDT-based approach assumes that the perceived memory strength of items varies on a single dimensional scale, and metamemory judgments are made based on whether perceived memory strength for each item exceeds or falls below a threshold(s). In the simplest model, perceived memory strength for recalled and forgotten items is assumed to follow two separate normal distributions with equal variance. The accuracy of judgments is defined as the distance between these two distributions, which is computed as follows:

$$d' = z(H) - z(F) \tag{3}$$

where *z* is the inverse cumulative normal distribution function with a mean of 0 and a standard deviation of 1. When metamemory judgments are made on an ordinal polychotomous scale, we can relax the assumption of equal-variance distributions between recalled and forgotten items to estimate a more general relative accuracy measure called d_a (Simpson & Fitter, 1973):

$$d_a = \sqrt{\frac{2}{1+m^2}} \times y_0 \tag{4}$$

where y_0 and m represent the y-intercept and slope of the normal deviate isosensitivity function (i.e., a linear function that relates *z*-transformed hit and false alarm rates; see Macmillan & Creelman, 2005), which can be estimated from the data. An alternative to d_a was also proposed and called A_z (Swets & Pickett, 1982):

$$A_Z = \Phi\left(\frac{d_a}{\sqrt{2}}\right) \tag{5}$$

where Φ is the cumulative normal distribution function.

Several simulation studies indicated that these SDT measures are superior to *G* in certain respects (Benjamin & Diaz, 2008; Masson & Rotello, 2009; Rotello et al., 2008), and more recent research has begun employing the SDT measures as alternatives to the conventional *G* (e.g., Luna, Higham, & Martin-Luengo, 2011). An overlooked, but potential caution one should employ with the use of SDT measures is that these measures posit separate (mostly normal) distributions for target and distractor items, and previous simulation studies have generated simulated datasets based on this assumption. In many metamemory experiments, however, such as those using JOLs (the main focus of the present study), participants study and make metamemory judgments for all studied items; there are no distractor items. In these cases, the perceived memory strength for all the items is most likely to follow a single normal distribution, rather than a mixture of two separate distributions. In these situations, it may not be appropriate to use SDT-based accuracy measures to assess relative accuracy of metamemory judgments[1].

Correlation Family

One straightforward way to quantify the relative accuracy is to compute the correlation between metamemory judgments and memory performance. Correlation measures have been typically used in eyewitness memory research (Krug, 2007), and are implicitly implemented in a recent stochastic model on the accuracy of JOLs (Jang, Wallsten, & Huber, 2012). The simplest index is a Pearson product-moment correlation between metamemory judgments and recall memory performance. As memory

performance is dichotomous (recalled or forgotten), it is also equivalent to the point-biserial correlation (r_{pb}). Some researchers may view a biserial correlation (r_b) as a more appropriate measure of metacognitive accuracy. In the context of metamemory research, the biserial correlation represents a correlation between metamemory judgments and latent continuous memory strength (see Adams, 1960), but sets an assumption that the observed memory performance is dichotomized based on a fixed threshold. Similarly, we can extend the biserial correlation to a polychoric correlation (r_{pc}), which represents a correlation between two latent continuous variables that are observed as ordered categories. The assumption behind this idea is that observed categorical metamemory judgments (e.g., JOLs with a Likert-type scale) reflect thresholded latent continuous judgments (see Jang et al., 2012). Both biserial correlation and polychoric correlation can be computed by a two-step maximum likelihood procedure (Olsson, 1979).

Although rarely used, a regression coefficient from logistic regression analysis that predicts memory performance from metamemory judgments may be a viable alternative to index metamemory accuracy. Logistic regression posits that a logit—the natural log of the odds—is linearly related to the independent variable(s). In the context of JOL accuracy, the model takes the following form.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \times JOL_i \quad (6)$$

where π_i is a probability that item *i* is recalled, and JOL_i is the JOL rating for the *i* th item. β_1 is the critical coefficient that represents the relationship between metamemory judgment and memory performance (in the current study, we call it *logistic B*). One advantage of *logistic B* is that it, or rather, the exponential coefficient, Exp (β_1), provides a more interpretable and intuitive metric than other accuracy measurements (e.g., *G*)—exponential beta is interpreted as the effect of the independent variable on the odds ratio of successful memory recall (i.e., the probability of recalling items divided by the probability of forgetting them). For example, if Exp (β_1) = 1.1, it means that the odds that an item is recalled is increased by 1.1 times as the JOL value is increased by one unit. Note that Exp (β_1) = 1 when JOLs do not predict memory performance at all (i.e., $\beta_1 = 0$) and $0 <$ Exp (β_1) < 1 when JOLs negatively predict memory performance (i.e., $\beta_1 < 0$).

Other Measures

Schraw (1995) suggested that the Hamann coefficient (*HC*) provides a better measure of relative metamemory accuracy. *HC* can be computed when both predictor and criterion variables are dichotomous using the following formula (Romesburg, 1984):

$$HC = \frac{(N_C - N_{IC})}{N_T} \quad (7)$$

where N_T is the total number of items, N_C is the number of correct judgments (the number of recalled items that are judged to be recalled and the number of forgotten items that are judged to be forgotten), and N_{IC} is the number of incorrect judgments (the number of recalled items that are judged to be forgotten, and the number of forgotten items that are judged to be recalled). Although this measure has been criticized as a measure of metamemory accuracy (Cheng, 2010; Nelson, 1996; Wright, 1996), it is sometimes used as a supplement of *G* (e.g., Reggev, Zuckerman, & Maril, 2011). Another measure of metacognitive accuracy is Hart's difference score, *D*, which is calculated as the difference in mean metacognitive judgments between recalled and unrecalled items (Hart, 1965; this is also conceptually equivalent to corrected hit rate in the case where both criterion and predictor variables are dichotomous). This measure is intuitively easy to understand, and

still practically reported (e.g., Merritt, Hirshman, Hsu, & Berrigan, 2005).

By-participant Approach to the Group-level Inference of Relative Accuracy

When data are collected from a group of participants and items, researchers are mostly interested in performing statistical tests (e.g., "Is the averaged gamma statistically different from zero?") to make group-level inferences. Group-level inference refers to making a statistical inference regarding whether an obtained result in a given sample can be generalized to the population of participants and items. Being able to make this inference is obviously a critical aspect of empirical research in psychology, and research in metamemory accuracy is no exception.

Importantly, although our review revealed a variety of measures that quantify the relative accuracy of metacognitive judgments, there is surprisingly little variation with regard to the group-level statistical inference in the literature. Specifically, researchers take two steps to obtain group-level inferential statistics: Researchers compute a measure of relative accuracy (e.g., *G*) for each participant, and then enter a set of these values into a *t*-test or an analysis of variance (ANOVA) using participants as the unit of analysis. Following Baayen (2008), we shall call this approach the *by-participant analysis*. When making a group-level inference on the relationship between two variables, this by-participant analysis has long been advocated in the literature of experimental psychology (Lorch & Myers, 1990; Monin & Oppenheimer, 2005), as it has some desirable characteristics over other approaches (e.g., pooling across participants, aggregation over participants or items, etc.).

The traditional by-participant analysis, however, considers only random variation across participants (i.e., random participant effect), and overlooks an important effect inherent in metamemory experiments: a random item effect. Typically, items used in a specific experiment are considered as a (ideally random) sample from an infinite population of items, and the same set of items is used across all the participants. Due to variation in random sampling, the relative metacognitive accuracy in the population of items should be different from that in the sampled items, and the relative accuracy in the sampled items should be be similar across participants as long as metamemory judgments (e.g., JOLs) for individual items have some similarities across participants (this is very likely, as metamemory judgments are influenced by item characteristics; see Koriat, 1997). For example, let's think about a situation where *G* between metacognitive judgments and memory performance is zero in an infinite population of items. In any given study, however, only a sample of items is included, and because of the random sampling variation of items, it is almost impossible that *G* computed from the included items is also exactly zero. Imagine that you have collected data from 1,000 participants in a metamemory experiment using one set of items, and compute *G* for each participant. Despite the fact that the relative accuracy (*G*) at the population level is zero, the averaged *G*s across participants approach a certain non-zero value (denoted as γ'), not zero. Although γ' may be very small, the statistical test (i.e., a one-sample *t*-test) is then very likely to show that the averaged value is significantly different from zero. In other words, statistical tests that do not account for a random item effect can seriously inflate the Type-1 error rate, producing significant effects even when the population *G* is zero.

Generalized Mixed-effects Model

The issue of random item effect is not new: Clark (1973) raised the issue nearly 40 years ago, and it has recently attracted revived attention in multiple fields of psychology (Baayen et al., 2008; E. Freeman, Heathcote, Chalmers, & Hockley, 2010; Judd, Westfall, & Kenny, 2012). The impact of random item effect in metamemory research, however, has rarely been discussed. One naïve approach to account for the random item effect is to compute relative accuracy for each item (across participants) and conduct a *t*-test using

items as the unit of analysis (called "by-item analysis"; Raaijmakers, 2003; see also Clark, 1973). This analysis, however, does not consider random variation across participants, and therefore the generalizability of the results to the population of participants is limited (Raaijmakers, Schrijnemakers, & Gremmen, 1999). The limitation of the conventional by-participant analysis and the by-item analysis is that both deal with only one type of random effect—either the random participant effect or random item effect. In recent years, however, statistical techniques, called *mixed-effects models,* have been developed that easily and effectively account for both random effects simultaneously (Baayen, 2008; Jaeger, 2008; Littell et al., 2006; Ozechowski, Turner, & Hops, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

The mixed-effects model with a dichotomous dependent variable (i.e., memory performance) is called the *generalized mixed-effects model*, and actually a simple extension of the logistic regression model explained earlier. Specifically, in a typical metamemory experiment where there is one continuous independent variable (i.e., metamemory judgments such as JOLs), the model can be specified as follows:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{00} + u_{0j} + u_{0i} + \beta_{10} \times JOL_{ij} \quad (8)$$

where π_{ij} represents the probability that the *i* th item of the *j* th participant is recalled. JOL_{ij} is metamemory judgment of the *i* th item of the *j* th participant (preferably mean-centered for each participant, to account for individual differences in average JOL ratings; Enders & Tofighi, 2007), and the association between memory strength and metamemory judgments (i.e., the effect of primary interest) is represented by β_{10}. As in logistic regression model, β_{10} is typically interpreted after being converted into Exp (β_{10}). Importantly, the equation takes into account both item (u_{0i}) and participant (u_{0j}) random effects that are added to the overall intercept term β_{00}. These multiple random effect components with different variances can be estimated based on maximum likelihood methods. Software for mixed-effects models is now widely available in specialized packages such as *HLM* (Raudenbush, Bryk, Cheong, Congdon, & Toit, 2011) or *Mplus* (Muthen & Muthen, 1998-2012) and in general statistical packages such as *SAS* (PROC MIXED or PROC GLIMMIX; Littell et al., 2006), *SPSS* (MIXED), or *R* (lme4 package; Bates, Maechler, & Bolker, 2011). It should be noted that the generalized mixed-effects models specified here include participants and items as crossed, independent random effects, as opposed to hierarchical or multilevel models in which random effects are assumed to be nested (Hox, 2002; see also Van den Noortgate & Onghena, 2006, for the simiralities and differences between the by-participant analysis and multilevel modeling analysis). The model specified in Equation 8 is considered an extension of these multilevel models, as it accounts for the fact that items are crossed with (rather than nested within) participants and accordingly, explicitly incorporates a random item effect (Quene & van den Bergh, 2008). Therefore, although recent metamemory research has begun to use multilevel modeling to model metamemory accuracy (e.g., Castel, Murayama, Friedman, McGillivray, & Link, in press; Hines, Touron, & Hertzog, 2009; Tauber & Rhodes, 2012), it is important to realize that the specification of a random item effect makes the mixed-effects model proposed here critically different from standard multilevel modeling approaches used in past research. Except for two remarkable papers specifically focusing on signal detection theory (Rouder & Lu, 2005; Rouder et al., 2007), the issue of random item effect has never been explicitly addressed in the context of metamemory research.

Mixed-effects models have several additional advantages. First, mixed-effects models have great modeling flexibilities. Equation 8 is a basic model and can be modified

according to individual experimental contexts and designs. For example, if researchers are interested in the difference in metacognitive accuracy across two groups, we can include an interaction term between metacognitive judgments and experimental conditions. Researchers can also specify *random slopes of participants* to account for the possibility that the association between metacognitive judgments and memory performance is different across participants This random slopes model can be specified as follows:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = (\beta_{00} + u_{0j} + u_{0i}) + (\beta_{10} + u_{1j}) \times JOL_{ij} \quad (9)$$

Equation 9 is different from Equation 8 in that random slope component (u_{1j}) is added to the model. β_{10} still represents the overall association between memory strength and metamemory judgments (i.e., the effect of primary interest), but this model also allows for the variation in the associations (i.e., slopes) across participants by incorporating u_{1j}.

Second, mixed-effects models allow us to test whether hypothesized random effects are actually present in the data. The random item effect in Equation 8 or 9 (u_{0i}) represents the inter-item variation of recall difficulty (i.e., u_{0i} is large when, on average, some items are difficult to remember and other items are easy to remember)[2]. We believe that such random item effects are relatively common in memory research (as discussed later), but regardless of one's standpoint, mixed-effects model can objectively evaluate the presence and the magnitude of different kinds of random effects in the data. Third, researchers can also include as many additional predictors as they want to control for the effects of these variables. This is an attractive feature of the mixed-effect model analysis, as computing partial associations is very cumbersome (if not impossible; see Goodman & Kruskal, 1954; Nelson, Narens, & Dunlosky, 2004) for the conventional *G* and SDT measures. We will revisit these issues in our simulation and empirical studies.

Monte Carlo Simulations

To illustrate the inflation of Type-1 error rate in the traditional by-participant analysis and the effectiveness of mixed-effects modeling analysis, we performed a set of Monte Carlo simulations with several experimental settings used in metamemory research to compare these analyses. The experimental designs were selected for illustrative purpose and simplicity, but there are clearly many other alternative designs that are more complex. Mixed-effects model analysis can be easily adapted for such complex experimental designs. The Supplementary Materials Online includes *R* script (using the *lme4* package; Bates et al., 2011) to perform mixed-effects model used in each simulation.

Simulation 1: A Single-Group Case with Typical JOL ratings

In Simulation 1, we investigated the Type-1 error rate for the simplest case where researchers are interested in whether metamemory accuracy (i.e., the JOL-memory association) is significantly different from chance based on a dataset of multiple participants and items. The relationship between metamemory judgments (i.e., JOLs) and memory strength was assumed to be independent at the population level (i.e., null relationship). Then, we computed and compared the proportion of false-positive effects (with alpha = 0.05) by various by-participant measures and the mixed-effect model.

Method. Hypothetical JOL experiment data were simulated by systematically varying the number of simulated participants (N = 20, 40, 60, or 80) and the number of items (K = 10, 30, 50, 70). In the simulation, for each trial of each participant, we randomly sampled continuous JOL values from a normal distribution with mean = 0 and SD = 1, and computed the corresponding memory strength by considering a random participant effect, a random item (intercept) effect, a random slope connecting JOLs and memory strength, and random noise (mean = 0 and SD = 1). Critically, the population

mean of the random slope is zero, meaning that the simulation assumed no overall relationship between JOLs and memory strength at the population level.

Details of the pre-defined parameters are described in the Supplementary Materials Online. Among other parameters, the foremost interest of our simulation is the effect of random item effect on Type-1 error rates. We manipulated the absence or the presence of random item effect by setting the SD of random item effect to 0 (i.e., no item effect variance) or 0.6 (i.e., item effect variance is about one-third of the variance of random noise). It should be added that, to account for possible variation in the slopes across participants, the SD of the random slopes was set to 0.3 in our simulation. The simulation also assumed that the simulated continuous JOL values from the same item were weakly correlated ($r = 0.3$) across participants, in order to account for the findings that JOLs are influenced by item characteristics (i.e., intrinsic cues; Koriat, 1997).

For each item, we set a threshold value of zero on the memory strength dimension such that any item with a strength value falling above the threshold was classified as recalled, and all other items were classified as forgotten. We also set five equal-interval threshold values on the JOL dimension such that the continuous JOLs are mapped onto a 6-point discrete scale, as is frequently done with JOL research (e.g., Dunlosky & Connor, 1997; Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002).

For each simulated experimental dataset with N participants of K items, a set of measures of metamemory accuracy were computed (i.e., G, G_w, G^*, r_{pb}, r_b, r_{pc}, *logistic B*, d_a, A_z, and D)[3] for each participant, and these values were entered into a one-sample t-test to test whether the average values were statistically different from chance. In addition, the same dataset was applied to a generalized mixed-effect model. The tested model was equivalent to Equation 9, and the independent variable (i.e., JOLs) was centered within participants (Enders & Tofighi, 2007; Hoffman & Stawski, 2009).

The main focus of this model is the statistical significance of the fixed slope value of JOL (β_{10} in Equation 9). There are primary two ways to obtain p-values from generalized mixed-effects models[4]. First, we divide the estimated coefficient by its standard error to obtain a z value, and judge the effect to be significant if the z value surpasses 1.96. Previous studies indicated that such z tests tend to be lenient in the context of mixed-effects model analysis (Baayen et al., 2008). The second option is to test the fixed slope value by using a log-likelihood ratio test (LRT; Baayen, 2008). Specifically, we applied a mixed-effects model twice to the same data, once with and once without the fixed slope. We then compared the fit statistics between these two models with a LRT. When a significant improvement of model fit was observed by including the fixed slope effect, we considered the fixed slope as significant (for problems with this procedure, see Pinheiro & Bates, 2000). The total number of replications (i.e., simulated experiments) was 5,000 for each combination of the parameters. Alpha was set to 0.05 throughout the simulation.

Results. Without a random item effect, the simulation showed that all the metamemory measurements and the mixed-effects model closely kept the nominal Type-1 error rates (see Figure S1 in Supplementary Materials Online), except for a very slight anticonservatism of z test in mixed-effects modeling with a small number of participants or items. These results suggest that when there truly is no random item effect, by-participant analysis yields tests of metamemory accuracy that have appropriate Type-1 error rates.

When a random item effect is present, however, the picture changes dramatically. Figure 1 plots the Type-1 error rates as a function of number of participants and items. In these simulated experiments, the mixed-effects model still shows reasonable Type-1 error rate (except for z test showing a slight inflation when the numbers of participants and items are small). In contrast, the conventional by-participant analysis shows a remarkable degree of positive bias, regardless of the accuracy measure used. Generally, Type 1 error rates

increased when 1) number of items decreased, and 2) number of participants increased. In the worst cases, Type-1 error rates exceeded .30 in our simulation. Larger number of participants exacerbated Type-1 error rates because, as described earlier, a random item effect produces an artefactual non-zero association between metamemory judgment and memory performance in sampled items. Thus, the chance of detecting this non-zero association increases as the number of participants increase. Increasing the number of items, on the other hand, directs the artefactual non-zero association in sampled items toward zero (i.e., the population value). Accordingly, we found in our simulation that using large number of items inhibited the inflation of Type-1 error rates. We also observed that the regression coefficient of logistic regression analysis (*logistic B*) showed slightly lower Type 1 error rates than other by-participant analyses. These results are, however, the consequences of some inappropriate solutions in logistic regression analysis with small samples (see Supplementary Materials Online for further the discussion).

In sum, these results indicate that the traditional by-participant analysis bears the potential danger to yield statistically significant results in the absence of true metamemory accuracy (and especially so with large sample sizes). The mixed-effects model analysis (especially when LRT is used), however, always showed the appropriate Type-1 error rates, irrespective of the number of items or participants, providing the strong evidence that this model is more appropriate in the presence of random item effects.

Statistical power. Although our main focus is on Type-1 error rates, we also investigated the statistical power of the metamemory measures and mixed-effects model. For that purpose, we ran the same set of simulations with the true slope of 0.2 (with the current simulation parameters, a JOL-memory slope of 0.2 corresponds to a correlation of about 0.17), and computed the rate of correctly detecting statistically significant relations between JOL ratings and memory performance. Figure 2 summarizes the findings. Given the statistical fact that Type-1 error rates and Type-2 error rates are negatively related, one may expect higher statistical power for by-participant analysis. Figure 2 indicates that this is not really the case. In most of the cases, statistical power is comparable between by-participant analysis and mixed-effects modeling. The mixed-effects model actually has greater statistical power when the numbers of participants and item are large. These results indicate that lower Type I error rate in mixed-effects model was not simply due to a general conservatism (see Supplementary Materials Online on adjusted power analysis).

Simulation 2: A Single-Group Case with varied JOL thresholds

Nelson (1984) argued that Goodman and Kruskal's gamma correlation (*G*) is preferable because it is insensitive to the placements of the thresholds of metacognitive judgments. In Simulation 1, however, we placed fixed-interval thresholds for metacognitive judgments across participants (i.e., we set five equal-interval threshold values on the JOL dimension such that the continuous JOL are mapped onto a 6-point discrete scale), and these fixed thresholds may have underestimated the usefulness of *G*. Simulation 2 addressed this issue by adopting varied thresholds for JOL ratings across participants.

Method. The simulation was identical to Simulation 1 with a random item effect, except for one setting. Specifically, rather than using fixed threshold values to determine categorical JOL ratings, we randomly sampled (from a uniform distribution between -1.5 to 1.5) five threshold values for each participant, and used these thresholds to map a continuous JOL values onto categorical JOL ratings on a 1-6 scale.

Results. The varied thresholds simulation revealed that *G* and other metamemory measures still exhibited Type-1 error rate inflation (Figure 3). On the other hand, mixed-effects model analysis kept Type-1 error rates close to 0.05, despite the fact that the mixed-effects model assumes an interval scale of measurement. These findings indicate the

robustness of mixed-effects modeling to threshold variation in metamemory judgment.

Simulation 3: A Single-Group Case with Dichotomous JOL ratings

In some metamemory paradigms, such as "feeling of knowing", metamemory judgments are made on a dichotomous scale (e.g., Hart, 1965). Accordingly, literature on measurement of metamemory accuracy has often focused on dichotomous metamemory judgments (Nelson, 1984; Rotello et al., 2008). As such, it is important to examine the performance of the by-participant analysis with a dichotomous independent variable. Because dichotomous judgments carry less information with regard to participants' true state (MacCallum, Zhang, Preacher, & Rucker, 2002), a mixed-effects model analysis may not exhibit better performance than other metamemory measures specifically adapted for dichotomous variables. In Simulation 3, we investigated the effectiveness of the by-participant and the mixed-effects model analyses in the case of dichotomous independent and dependent variables.

Method. The simulation setup was identical to that of Simulation 1 with the presence of a random item effect, except for the following two aspects: First, instead of using fixed multiple thresholds to determine multiple discrete JOL ratings, the current simulation set a single threshold value that varies across participants. Simulated continuous JOL values were categorized as "yes" when the values were above the threshold and as "no" when the values are below the threshold, resulting in a 2 (metamemory judgment; yes vs. no) X 2 (memory recall; recalled vs. forgotten) observed data structure for each participant. Second, we computed measures of metamemory accuracy that are adapted to the dichotomous independent variable. We kept *G*, G_w, *G**, as these measurements can be used for dichotomous independent cases. From the signal detection family, we computed *d'* as well as G_c, which was proposed by Masson & Rotello (2009) as an alternative measure of metamemory accuracy in dichotomous metamemory judgment:

$$G_C = 2(H - F) - (H - F)^2$$

From the correlation family, we computed Pearson product-moment correlation (r_{pb}), polychoric correlation (r_{pc}), and logistic regression coefficients (*logistic B*) as in Simulation 1. In the case of dichotomous independent and dependent variables, the first two correlations are called the Phi coefficient and tetrachoric correlation, respectively. We also computed Hamann coefficient (*HC*) and Hart's difference score *D*.

Results. Figure 4 reports the observed Type-1 error rates as a function of number of participants and items. Although Type-1 error rates for measurements using the by-participant analysis are somewhat smaller than Simulation 1, they are still well beyond the nominal alpha = 0.05. On the other hand, despite the loss of information due to dichotomization, the mixed-effects model retained appropriate Type-1 error rates, further indicating the robustness of this approach (see also Supplementary Materials Online on the discussion of slightly lower Type-1 error rates of *logistic B* and *G**).

Simulation 4: A Case with Comparing Two Within-participant Conditions with a Between-Item Manipulation

Metamemory research is sometimes interested in comparing metamemory accuracy between two (or more) within-person conditions with a between-item manipulation. For example, Zimmerman and Kelley (2010, Experiment 1; see also Tauber & Dunlosky, 2012) asked participants to make JOLs for emotionally negative and neutral word pairs (22 pairs each). Although not the main focus of the article, they compared the JOL-recall gamma coefficients across the emotionality conditions, and found better metamemory accuracy (*G*) for negative word pairs than neutral word pairs. Note that this example examined the effect of emotional valence with a between-item manipulation (i.e., each condition has different items). In Simulation 4, we examined whether the inflation of Type-1 error rates would be observed with such an experimental design.

Method. Hypothetical JOL experiment datasets were simulated by systematically varying the number of simulated participants (N = 20, 40, 60, or 80) and the total number of items (K = 20, 40, 60, 80). The basic parametric settings were almost the same as that of Simulation 1 with a random item effect. The main difference was that the items were split into one of the two conditions. Importantly, we posited that the true slope relating JOL and memory strength is identical (β = 0.3) between these conditions. That is, we assumed no effects of the experimental condition on metamemory accuracy at the population level. We also arbitrarily set the effects of experimental condition on memory to 0 and the SD of random participant effect of experimental condition to 0.3.

For each simulated experimental dataset, the measurements of metacognitive accuracy (i.e., G, G_w, G^*, r_{pb}, r_b, r_{pc}, *logistic B*, d_a, A_z, and D) were computed for each condition of each participant. These values were entered into a paired *t*-test to test whether the average values were statistically different between the two experimental conditions. In addition, the same dataset was applied to a generalized mixed-effect model, with experimental condition as a predictor. Importantly, we included the interaction between metamemory judgment (JOL) and experimental condition. This interaction term was the primary focus of this mixed-effects model, as the term represents different JOL-memory relations between the two experimental conditions. Due to the computational load, the total number of replications was 3,000 for each combination of parameter values.

Results. The results (Figure 5) revealed that the conventional by-participant analysis still produces substantial inflation of Type-1 error rates, even when the primary focus is on the difference between conditions. Like in Simulations 1-3, Type-1 error rates increase as the number of participants increases. An interesting observation is that, unlike the previous simulations, increasing the number of items did not prevent the inflation of Type-1 error rates in this experimental design (see Supplementary Materials Online "Effects of Random Slope Variance on Type1 error Rate" for further discussion).

Simulation 5: A Case with Comparing Two Between-participant Groups with a Within-Item Manipulation

Another typical comparison in metamemory research is that of metamemory accuracy between two independent groups with the same items. For example, research on aging has been concerned with age-related differences (e.g., younger adults vs. older adults) in metacognitive accuracy (Hertzog & Dunlosky, 2011). In these studies, younger and older adults make metacognitive judgments on the same learning materials, and their metacognitive accuracy (typically, gamma correlation) is compared.

Simulation 5 examined the Type-1 error rates when two independent groups are compared (i.e., a between-participants design). Importantly, in this particular paradigm, we do not anticipate the inflation of Type-1 error rates with the presence of a random item effect as defined in the previous simulations. Because the same learning materials are used for both groups, the artefactual non-zero association between metacognitive judgment and memory performance in sampled items is expected to be identical between the groups. Therefore, as far as researchers focus on the between-groups difference, statistical inferences are not biased. However, this expectation rests on the critical assumption that the random item effect (intercept) is identical across the groups. This is not a realistic assumption in many situations. For example, dated words may be easier to older adults, and words with higher-present-day frequency may be easier for younger adults to remember (the word frequency cohort effect; Worden and Sherman-Brown, 1983). Therefore, item effects may be different between younger and older adults. In the presence of such a random item effect X group interaction (in the context of mixed-effects model, this is equivalent to a random item slope of group), we may still expect inflation of Type-1 error rates with the traditional by-participant analysis.

Method. We systematically varied the number of simulated participants for each group (N = 20, 40, 60, or 80) and the total number of items (K = 10, 30, 50, 70) to generate hypothetical JOL experiments. The basic parametric settings were the same with Simulation 4, with the following exceptions: First, because we were interested in the between-groups difference, we did not split data of individual participants into two conditions. Instead, for each replication, two sets of independent data (number of participants = N) were generated to represent each group. Crucially, we posited that the true slope relating JOL and memory strength is identical ($\beta = 0.3$) between these groups, assuming no group difference in metamemory accuracy at the population level. Second, we manipulated the presence or absence of the random item effect X group interaction, by setting the SD of the random effect to 0.6 or 0.

For each replication, the measurements of metacognitive accuracy (i.e., G, G_w, G^*, r_{pb}, r_b, r_{pc}, *logistic B*, d_a, A_z, and D) were computed for each participant. These values were entered into an independent samples *t*-test to test whether the average values were statistically different between the groups. The same dataset was applied to a generalized mixed-effect model, with group as a participant-level predictor. Importantly, we included the [cross-level] interaction between metamemory judgment (JOL) and group—the primary focus of this mixed-effects model analysis, as the term represents different JOL-memory relations between the groups.

Results. Type-1 error rates without the random item effect X group interaction showed that, although this simulation included a random item effect, as expected, the by-participant analysis measures all exhibited appropriate Type-1 error rates, regardless of the number of items or participants (see Figure S3 in Supplementary Materials Online). When the item effect X group interaction is considered, however, the same pattern as our previous simulations was observed in the by-participant analysis—Type-1 error rates were inflated (Figure 6), especially when the experimental design involves a small number of items or large number of participants. On the other hand, the mixed-effects model analysis kept the correct Type-1 error rates.

Simulation 6: A Case with Comparing Two Within-participant Groups with a Within-Item Manipulation

Some metamemory studies employ counterbalancing procedures to assign the same items to different within-participant conditions. For example, Sungkhasettee, Friedman, and Castel (2011) assessed the JOLs of the items presented in an upright or inverted format. The word format (upright vs. inverted) was manipulated within participants: Half of the items were assigned to the upright condition, the other half of the items were assigned to the inverted condition, and the assignment of the items were counterbalanced across participants. Like the previous simulation using a between-participants design (Simulation 5), this paradigm is another example where the effect of a random item effect inflating Type-1 error rate in by-participant analysis is minimized: As the same items are used between the conditions (across participants), the artificial non-zero association between metacognitive judgment and memory performance in sampled items is expected to be the same across the conditions, preventing false significant effects (Raaijmakers et al., 1999). It is still possible, however, to conceive of scenario where this design is subject to the inflation of Type-1 error rate. Specifically, some types of experimental manipulations may alter the pattern of the random item effect between the conditions. Taking an example from Sungkhasettee et al. (2011), inverted words were overall recalled better than upright words on the final memory test, perhaps due to their more elaborate encoding process. Such an elaborative encoding process could have a differential impact on different items. If this random item effect X condition interaction is present, the artificial non-zero association between metacognitive judgment

and memory performance is different between the conditions, resulting in the inflation of Type-1 error rate. Simulation 6 examined the effects of random item effect X condition interaction in a within-participant design.

Method. We systematically varied the number of simulated participants for each group (N = 20, 40, 60, or 80) and the total number of items (K = 20, 40, 60, 80) to generate hypothetical JOL experiments. The basic parametric settings were the same with Simulation 4, with the two exceptions. First, the assignment of the items was counterbalanced across participants with two counterbalanced lists, so that the same item was assigned to the two conditions equally often. Second, we manipulated the presence or absence of the random item effect X condition interaction, by setting the SD of the random effect to 0.6 or 0.

By-participant analysis was the same with Simulation 4. The same dataset was applied to a generalized mixed-effect model, with condition as a within-participant predictor. Importantly, we included the interaction between metamemory judgment (JOL) and condition, which was the primary focus of this mixed-effects model analysis, as the term represents different JOL-memory relations between the conditions.

Results. In accordance with Simulation 5, Type-1 error rates without the random item effect X group interaction, but with the random item effect (intercept), showed appropriate Type-1 error rates, regardless of the analysis methodology used (see Figure S4 in Supplementary Materials Online). When the item effect X group interaction is considered, however, the conventional by-participant analysis showed increased Type-1 error rates, especially when the experimental design involves a small number of items or large number of participants. On the other hand, the mixed-effects model analysis kept the correct Type-1 error rates. It should be noted, however, that the false-positive rates were smaller than the other simulations, indicating the relative robustness of the current experimental design to Type-1 error rate inflation.

Real Data Applications

The Monte Carlo simulations show that the traditional by-participant analysis tends to produce increased Type-1 error rates, regardless of specific type of measurements used. In contrast, the mixed-effects model analysis effectively prevented the inflation of Type-1 error rates. In the following sections, we present several real data applications to further complement these findings, and highlight additional strengths that are unique to the mixed-effects model analysis.

Real Data Example 1: Test of Random Item effect

Our simulations revealed that a random item effect is a critical factor to determine whether the by-participant analysis produces increased Type-1 error rates. Without a random item effect, the by-participant analysis was able to hold Type-1 error rates at the set alpha level of 5%. This raises an important question: Is the random item effect common in practice? A random item effect in the context of memory research refers to the variation of memory performance (e.g., recall rate) between items. In other words, it is the inter-item variability in memorability (or recall difficulty) of items. We believe that such a random item effect is common in memory research (see also E. Freeman et al., 2010), because previous studies have revealed a myriad of factors that influence memorability of individual items, such as word frequency (Hall, 1954), word length (Watkins, 1972), imagery (Paivio & Smythe, 1971), emotionality (Kleinsmith & Kaplan, 1963), etc. Unless all of these factors are held completely constant, a random item effect is likely to exist.

To illustrate this point, the current example utilized a real metamemory dataset with standardized word stimuli (i.e., word stimuli used in a previous study), and empirically examined the presence of random item effect in that experiment. The mixed effects model is particularly useful even in this context, because unlike the by-participant

analysis, mixed-effects modeling can statistically test any random components included in the model. By taking advantage of this feature in the mixed-effects model analysis, we examined whether the random item effect is statistically significant for real experimental data using standardized word stimuli.

Method. Fifty participants were recruited from Amazon.com's Mechanical Turk. The learning materials were 40 related word-pairs taken from Experiment 3 of Connor, Dunlosky, and Hertzog (1997). These target-cue pairs were highly associated on the basis of the University of South Florida associability norms. In the experiment, each word pair was presented for three seconds in a randomized order and participants were asked to make predictions (JOLs) about the likelihood that the target item would be recalled in a later cued recall test on scale of 0% to 100%. This study phase was followed by a 60 second distractor task, and then a cued recall test.

Results. To examine whether there was a statistically significant random item effect, we conducted a mixed-effects model, with memory performance as the dependent variable. This model was the same with Equation 8 but the independent variable (JOLs) was omitted from the model. The results (from a LRT) showed a highly significant random item effect, χ^2 (1) = 78.80, $p < .01$. These findings indicate that the variation in memory performance with this specific learning material includes random item effects, suggesting the importance of incorporating random item effects with this learning material. The random participant effect was also significant, χ^2 (1) = 306.16, $p < .01$.

The averaged *G*s between JOLs and memory performance was .17 (*SD* = .37), which was significantly greater than zero, $t(45) = 3.10$, $p < .01$. Note that we needed to exclude four participants for whom we could not compute *G*, because they did not have variation in JOLs or memory performance. We also conducted a mixed effect model analysis that predicts memory performance from JOLs (Equation 9). JOLs were transformed into units of 10% (i.e., 1 point interval represents 10%) and group-mean centered. The results showed that the effects of JOL on memory performance was statistically significant with both *z* test, $z = 4.86$, $p < .01$ and LRT, χ^2 (1) = 20.37, $p < .01$. The estimated exponential beta value, Exp (B), was 1.18, meaning that the odds that an item is recalled is increased by 1.18 times as the JOL value is increased by 10%. The random slope variance was not statistically significant, χ^2 (1) = 1.33, $p = .25$.

The mixed effect model confirmed that the association between JOLs and memory performance was statistically significant in this experiment. However, the random item effect observed in this experiment may have inflated Type-1 error rate, had there not been any associations between JOLs and memory performance at the population level. Accordingly, to evaluate Type-1 error rates that could have happened with this memory data set, we further conducted a Monte Carlo simulation. Specifically, for the obtained memory dataset, we replaced the actual JOLs with hypothetical JOL values that were randomly generated in the same manner with Simulation 1. Then, we conducted a statistical test on the relationship between the generated JOLs and memory performance once with a by-participant analysis with gamma correlation and again with a mixed effect model analysis. This procedure was repeated 3,000 times. Because JOLs were randomly generated without considering actual memory performance, the JOL-memory relation should be null at the population level, and the rate of significant effects in this simulation can be considered as Type-1 error rate.

Consistent with findings from previous simulation studies, the mixed effects model produced a Type-1 error rate of 0.05 using the z-test and of 0.04 using LRT. On the other hand, the by-participant analysis with gamma correlation showed a Type-1 error rate of 0.15, comparable to what was observed in the previous simulations (see Figure 1). Although we need to be careful in generalizing our findings to other learning stimuli, these

findings indicate that our simulation studies reflect realistic situations that are existent in real memory data, at least to some extent.

Real Data Example 2: Data with a Large Number of Missing Gammas

When a participant's data does not vary in either metacognitive judgments (i.e., the same JOL value for all the items) or memory performance (i.e., perfect recall or zero recall), most measurements of metacognitive accuracy cannot be computed for that person. This is one of the issues frequently encountered in metamemory research. For example, Experiment 1 of Zimmerman and Kelley (2010), which was mentioned earlier, excluded 12 out of 48 participants from their analysis because gamma could not be computed for these participants. For the same reason, Experiment 3 of Schwartz and Metcalfe (1992) excluded 11 out of 32 participants for their analysis of metacognitive accuracy. This problem is particularly likely to happen when the number of items is small or memory performance exhibits ceiling or floor effects. The omission of participants, however, may bias the results, unless the pattern of missing/non-computable data takes place completely at random (Schafer & Graham, 2002).

On the other hand, the mixed-effects model analysis is a one-step procedure that does not directly compute metacognitive accuracy for each individual. Accordingly, a mixed-effects model can make full use of the information in the data, without the need to exclude participants that lack within-person variation. We will illustrate this point in the next example using the dataset from Experiment 2 of Kornell and Bjork (2008)[5].

Method and Procedure. In Experiment 2, Kornell and Bjork (2008) asked 54 participants to study 20 Swahili-English translations (Nelson & Dunlosky, 1994), each printed on a card. The procedure was similar to studying flashcards: Participants cycled through the cards repeatedly. On each trial, the front of the card (the Swahili cue) was shown first, before the card appeared to flip and the back (the English target) was shown. After the presentation of each target, participants were allowed to drop the item from the stack, if they did not want to study that word again in the next study cycle. When they decided to drop an item, they were asked to make a JOL on that item by selecting one of six categories (0%, 20%, 40%, 60%, 80%, and 100%). A cued-recall test was administered after a 5-minute distractor task.

Results. This illustration aims to examine the metamemory accuracy for the dropped items that received JOLs. Two participants did not drop any items and thus made no JOL ratings, resulting in the final set of 52 participants. We first computed *G* for each participant. As reported in Kornell and Bjork (2008), we could not compute *G* for as many as 48% of participants (i.e., 25 participants). This is mostly because of the fact that many participants showed perfect memory performance for the items they decided to drop. Another reason is that some participants had only a few JOL ratings because JOLs are made only on the dropped items. With the remaining 27 participants, the average gamma correlation between JOLs and recall performance (for the dropped items) was very high, $M = .59$ ($SD = .51$) and significantly greater than zero, $t(26) = 6.03$, $p < .01$.

Consistent with the previous application, our preliminary analysis on the data from Kornell and Bjork (2008) showed a significant random item effect, $\chi^2(1) = 4.44$, $p < .05$. We then conducted a mixed-effects model predicting memory performance from JOLs (Equation 9). JOLs were transformed in the units of 10% (i.e., 1 point interval represents 10%) and group-mean centered. Despite the large number of participants with zero variance, the model converged successfully. The effect of JOL on memory performance was statistically significant for both *z* test, $z = 5.16$, $p < .01$, and LRT, $\chi^2(1) = 18.2$, $p < .01$. The estimated exponential beta value Exp (B) was 1.63, meaning that the odds that an item is recalled are 1.63 times greater as the value of JOL is increased by 10%. The random slope variance was not statistically significant, $\chi^2(1) = 0.86$, $p = .35$. Although

both analyses showed the statistically significant effect of JOLs on memory performance, the present example provides a nice illustration of usefulness of a mixed-effects model analysis when there are a large number of participants with zero variance. In such cases, even if the results are consistent, the mixed-effects model is preferable given that it makes full use of the information in the data.

Real Data Example 3: Controlling for Other Independent Variables

The mixed-effects model can be viewed as an extension of regression or logistic regression analyses, and accordingly, researchers may include multiple independent variables. The mixed-effects model analysis produces estimated beta values (called partial regression coefficients), which can be interpreted as the effect of a given independent variable after partialing out for the other independent variables (Cohen, Cohen, West, & Aiken, 2003). As such, by including multiple predictor variables, mixed-effects model enables researchers to statistically tease apart independent contributions of each independent variable to the dependent variable. Most of the conventional measurements in the by-participant analysis (e.g., G or A_z), in contrast, have difficulty in addressing the effects of multiple independent variables or partial relations.

To illustrate the usefulness of a multiple-predictor model, this example employed data from unpublished work (Yan, Murayama, & Castel, 2013). This work aimed to investigate the effects of personal preference judgment on memory performance. Thus, though the study does not specifically examine metamemory judgments, it still highlights an important strength of the mixed-effects model analysis.

Method and Procedure. Ninety-one participants were recruited from Amazon.com's Mechanical Turk (for details of the experiment, see Supplementary Materials Online). The learning materials were 16 popular ice cream flavors (e.g., strawberry, coconut). Participants were shown the 16 flavors one at a time, and asked to rate their preference for that flavor on a 1 (I really don't like this flavor) - 10 (I love this flavor) scale. Following a distractor task, participants were asked to recall the flavors they had studied. After recall, participants were asked to rate the familiarity of each flavor in their daily experience from 1 (not at all familiar) to 10 (very familiar).

Results. We first computed G between preference ratings and memory performance for each participant. Excluding 10 participants for whom G could not be computed, the average G was positive and statistically significant, $M = 0.08$, $t\,(80) = 2.13$, $p < .05$, suggesting that preference drives better memory performance. However, memory performance was also positively related to flavor familiarity, $M = .20$, $t\,(87) = 4.48$, $p < .01$, and preference ratings and familiarity ratings were also strongly related, $M = .52$, $t\,(80) = 14.32$, $p < .01$. These findings suggest the possibility that the positive relationship between preference and memory performance may be caused by the fact that familiar flavors are more preferred and memorable.

Our preliminary mixed-effects model showed a significant random item effect, $\chi^2\,(1) = 64.72$, $p < .01$. When we tested a mixed-effects model that included preference ratings as a predictor, the results showed significant effect of preference on memory performance, Exp (B) = 1.06, for both z test, $z = 2.88$, $p < .01$, and LRT, $\chi^2\,(1) = 8.00$, $p < .01$. The estimated coefficient indicated that the odds that an item is recalled are 1.06 times likely as the value of preference rating is increased by 1. When we added familiarity ratings as the second predictor, however, the familiarity ratings showed a significant effect on memory performance for both tests, Exp (B) = 1.07, $z = 3.10$, $p < .01$; $\chi^2\,(1) = 9.24$, $p < .01$, and the effects of liking was no longer significant for both tests, $z = 1.34$, $p = 0.18$; $\chi^2\,(1) = 1.74$, $p = .19$. These results indicated that preference is not associated with memory performance, when familiarity ratings are held constant, a conclusion that could not have been drawn from the G measures. The random slope of either preference or

familiarity was not significant, $p = 1.00$ and $p = 1.00$, respectively.

General Discussion

Since the 1980s, measurements of metacognitive accuracy have been a recurring issue in metamemory literature. Researchers have proposed a variety of measurements, including the most widely used gamma coefficient (*G*). The strengths and weaknesses of different measurements have also been discussed at length (Benjamin & Diaz, 2008; Maniscalco & Lau, 2012; Masson & Rotello, 2009; Nelson, 1984; Rotello et al., 2008; Schwartz & Metcalfe, 1994). Despite the number of proposed measurements, previous research has almost unanimously relied on the so-called by-participant analysis, where metacognitive accuracy measurements are computed for each participant and then entered into a statistical test (e.g., *t*-test) to make a group-level inference. In contrast, little attention has been paid to Type-1 error rates when making group-level statistical inferences. Although by-participant analysis is a well-accepted procedure in practice, the current work showed that this conventional analysis can inflate Type-1 error rates when a random item effect is present. Our simulation showed that, in certain circumstances, this inflation can be considerably high, calling for attention to this important issue. Our Monte Carlo simulation illustrated that the mixed-effects model can be an effective way to address this issue, across many different experimental designs and even in situations where the common metacognitive accuracy measure (i.e., *G*) allegedly has advantage (e.g., varied thresholds in JOLs; Simulation 2).

It should be emphasized that we by no means argue that many of the findings on metacognitive accuracy are false-positives. The relationship between metacognitive judgments and memory performance is mostly small to moderate, if not large (Dunlosky & Metcalfe, 2009), and it is quite implausible to explain this large body of consistent observations by increased false-positive rates with the by-participant analysis. Nevertheless, our findings have potential to provide a useful alternative view to interpret some past findings in metamemory research. For example, as described earlier, Experiment 1 of Zimmerman and Kelley (2010) found better metamemory accuracy (assessed by *G*) for negative word pairs than neutral word pairs using the by-participant analysis. However, this result was not replicated in their subsequent studies (Experiments 2 -4). Although the authors did not provide explanations, our work could suggest that these inconsistent results may be simply due to a Type-1 error. Another example comes from aging literature, where the majority of research has shown that there is no age-related decline in terms of metacognitive accuracy (Hertzog, Sinclair, & Dunlosky, 2010; Robinson, Hertzog, & Dunlosky, 2006; for reviews, see Hertzog & Dunlosky, 2011; Rhodes & Tauber, 2011). There are, however, a few studies that show a statistically significant difference in metacognitive accuracy between younger and older adults (e.g., Daniels, Toth, & Hertzog, 2009; Souchay, Moulin, Clarys, Taconnat, & Isingrini, 2007). These inconsistencies were attributed to methodological differences in experiment (see Hertzog, Dunlosky, & Sinclair, 2010). However, our findings suggest that part of these inconsistent findings might be caused by the inflation of Type-1 error rates in using the by-participant analysis.

When By-participant Analysis is Resistant to the Type-1 Error Inflation

Our primary recommendation is to use mixed-effects model whenever possible, but it is also practically useful to know when the conventional by-participant analysis does not inflate Type-1 error rates (especially when we read or review a paper that uses the by-participant approach). Our simulations suggested some answers to this question.

The foremost important factor is clearly the presence of a random item effect. As indicated earlier, it is difficult to nullify random item effect in practice, as there are numerous factors that contribute to the memorability of individual items. Careful choice of items that hold several important characteristics (e.g., word frequency, word familiarity)

constant, however, may reduce the random item effect to a certain extent. These homogeneous items, on the other hand, cannot be regarded as a random sample from a broader item population. Thus, this selective approach should limit the generalizability of the findings to a broader set of items with different item characteristics. The second factor we observed is the number of items in an experiment. Our simulations showed that using a large number of items (per condition) typically reduces Type-1 error rates. Although this may appear to be a realistic solution, in practice it is often difficult to increase the number of items per condition, especially when there are multiple within-participant factors.

Another important factor is the nature of experimental design. As shown in Simulations 5-6, random item effect does not necessarily inflate Type-1 error rate when two groups/conditions use the same items across participants (see Figures S3-S4; see Raaijmakers et al., 1999). Our simulations indicated that a random item effect X condition interaction could increase the Type-1 error rate even in those cases, but the impact was less than in the other simulations. A similar alternative design that can be applied to most experimental designs is to prepare a large number of items and assign a small set of different items to each participant/condition, either by using different lists or randomly sampling items. Although not standard in previous experiments (see Son & Kornell, 2009, for an exception), such item selection procedure could minimize the problem of by-participant analysis. Finally, our simulations also showed that the number of participants is another consistent factor that influences Type-1 error rates. That is, increasing the number of participants exacerbated false-positive errors. This finding, however, does not mean that studies with a smaller number of participants are desirable. As shown in our simulation (Figure 2), small numbers of participants yield poor statistical power, which makes it difficult to make inferences when statistically significant effects are not obtained (Wagenmakers, 2007).

Issue of Model Specification

One of the strengths of mixed-effects modeling is the flexibility of model construction. That is, researchers can flexibly incorporate different kinds of random effects, such as a random participant effect or random participant slopes based on the experimental design, and statistically evaluate the presence of these effects. But the flip side of this flexibility is the risk of model misspecification. For example, in the current paper, our mixed effects model mainly tested the model in Equation 9 (and data are basically generated by this model). It is possible, however, to add another type of random item effects to the model—a random item slope (u_{1j} in the following equation).

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = (\beta_{00} + u_{0j} + u_{0i}) + (\beta_{10} + u_{1j} + u_{1i}) \times JOL_{ij} \quad (10)$$

In this model, metacognitive accuracy can vary across different items as well as different participants. We did not consider the random item slopes in our models because 1) the complexities of the model would be likely to a produce non-convergence problem in estimation process and 2) in our own experiences, the variance of random item slopes is much less likely to be statistically significant[6]. If random item slopes exist, however, specifying a model in Equation 9 would be likely to produce incorrect parameter estimates and Type-1 error rate. Although the field has not come to an agreement, two other strategies to handling the issue of model misspecification are worth mentioning.

First, we specify all the possible random effects afforded by the design of the experiment in focus in any statistical analyses. Barr et al. (2013) showed that this strategy performs better than other modeling strategies when the dependent variable is continuous. He also noted, however, that the complexity of the full model may hinder the estimation

process when the dependent variable is categorical like in our simulations and examples. Second, we can decide upon the inclusion or exclusion of a random effect based on the statistical test (i.e., LRT) of that random effect. This stepwise procedure has the advantage in that it can avoid model nonconvergence and possibly preserves statistical power by minimizing the number of estimated parameters. On the other hand, this strategy capitalizes on the sampling variation and thus runs the risk of model overfitting, endangering the generalizability of results (Babyak, 2004). In addition, this strategy may lead to different final models depending on how random item effects are sequentially tested (Barr et al., 2013). Future work is needed to delineate the best way to avoid model misspecification in the context of generalized mixed-effects model.

Additional Strengths of Mixed-effects Model and Expansion

Missing data (participants who do not allow us to compute metacognitive accuracy measure) present a practical challenge to researchers in metamemory. This happens, because most of the measurements in metacognitive accuracy cannot be computed when the variance of one variable is zero. The situation becomes even worse for the measurements that do not have lower or upper boundaries such as *G**, *logistic B*, or *d*'. For these measurements, "perfect" association produces an infinite value, and researchers are forced to drop these participants or add some ad-hoc modifications to re-compute the value (Stanislaw & Todorov, 1999). On the other hand, our second application illustrated that the mixed-effects model can easily address this issue, making full use of the information from the data. It should be added that the mixed-effects model is also resistant to item-level missing data (i.e., missing observations within each participant). This is typically the case in "feeling-of-knowing" research, where participants make feeling-of-knowing judgments only for those items for which they could not recall a correct answer (Gruneberg, Monks, & Sykes, 1977). In such situations, the reliability of computed metamemory-memory association is different between participants, as the number of items is different across participants (e.g., *G* computed from a participant with fewer items is relatively unstable). Mixed-effects model analysis effectively takes into account the reliability information and provides a precise estimate of metamemory-memory association.

The final empirical application showed the capacity of the mixed-effects model to include multiple predictors. In metamemory literature, it is not uncommon to ask participants to make multiple metacognitive judgments (e.g., Leonesio & Nelson, 1990; Wahlheim, Finn, & Jacoby, 2012). Furthermore, typical metamemory experiments inherently contain a variety of information that may influence memory performance or metamemory judgments such as reaction time or item order. In these situations, the mixed-effects model is particularly a powerful tool, as it computes the unique contribution of each independent variable to memory performance. This way, researchers can eliminate possible confounding variables or understand the factors that contribute to a specific metamemory-memory relation[7]. In addition, recent statistical developments allow researchers to test more complicated path models (e.g., mediation model) using a multilevel structural equation model framework (Asparouhov & Muthen, 2012; see also Preacher, Zyphur, & Zhang, 2010).

As the equal-variance signal detection model can be considered as a submodel of generalized mixed-effects model (DeCarlo, 1998), it is possible to expand our model to encompass the signal detection measure of metamemory accuracy. Similarly, Rouder and his colleagues (Rouder & Lu, 2005; Rouder et al., 2007) recently provided a hierarchical Bayes framework of the signal detection model that takes into account both random participant and item effects. This expansion of the model, though the analysis is not yet easily accessible to researchers who are not expert in statistics, would open a venue for a

more fine-grained analysis on metacognitive accuracy. A Bayesian framework (e.g., Fong, Rue, & Wakefield, 2010) can also effectively loosen the assumptions inherent in the mixed-effects models (e.g., normality of error term). Future studies would do well to examine and compare the performance of such advanced statistical models.

References

Adams, J. F. (1960). The effect of non-normally distributed criterion scores on item analysis techniques. *Educational and Psychological Measurement, 20*, 317-320.

Asparouhov, T., & Muthen, B. (2012). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters.* Retrieved from http://www.statmodel.com/download/NCME_revision2.pdf

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, England: Cambridge University Press.

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. doi: 10.1016/j.jml.2007.12.005

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regressiontype models. *Psychosomatic Medicine*, *66*, 411-421. doi: 10.1097/01.psy.0000127692.23278.a9

Barr, D. J., Levy, R., Scheepers, C. and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.

Barrett, Adam B.; Dienes, Zoltan; Seth, Anil K. (in press). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes (Version R package version 0.999375-39). Retrieved from http://CRAN.R-project.org/package lme4

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73-94). New York: Psychology Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417-444. Doi: 10.1146/annurev-psych-113011-143823

Bruner, J. & Austin, P. (2009). Inflation of Type 1 error rate in multiple regression when independent variables are measured with error. *Canadian Journal of Statistics*, *37*, 33-46.

Castel, A. D., Murayama, K., Friedman, M. C., McGillivray, S., & Link, I. (in press). Selecting valuable information to remember: Age-related differences and similarities in self-regulated learning. *Psychology and Aging*.

Cheng, C. M. (2010). Accuracy and stability of metacognitive monitoring: A new measure. *Behavior Research Methods, 42*(3), 715-732. doi: 10.3758/brm.42.3.715

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

Clarke, F. R., Birdsall, T. G., & Tanner, W. P. (1959). Two types of ROC curves and definitions of parameters. *Journal of the Acoustical Society of America, 31*, 629-630.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*(1), 50-71.

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods*, *16*, 166-178. doi: 10.1037/a0023355

Daniels, K. A., Toth, J. P., & Hertzog, C. (2009). Aging and recollection in the accuracy of

judgments of learning. *Psychology and Aging, 24*(2), 494-500. doi: 10.1037/a0015269.
DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186-205. doi: 10.1037/1082-989X.3.2.186
Dixon, P. (2008). Models of accuracy in repeated-measures designs. [Article]. *Journal of Memory and Language, 59*(4), 447-456. doi: 10.1016/j.jml.2007.11.004
Dunlosky, J., & Connor, L. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory & Cognition, 25*(5), 691-700. doi: 10.3758/bf03211311
Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA, US: Sage Publications, Inc, Thousand Oaks, CA.
Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138. doi: 10.1037/1082-989x.12.2.121
Fong, Y. Y., Rue, H., & Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics, 11*(3), 397-412. doi: 10.1093/biostatistics/kxp053
Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory and Language, 62*(1), 1-18. doi: 10.1016/j.jml.2009.09.004
Freeman, L. C. (1986). Order-Based Statistics and Monotonicity: A Family of Ordinal Measures of Association. *The Journal of Mathematical Sociology, 12*(1), 49-69.
Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10*(4), 843-876. doi:
Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119*, 159-165.
Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732-769.
Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. Oxford, England: John Wiley, Oxford.
Gruneberg, M. M., Monks, J., & Sykes, R. N. (1977). Some methodological problems with feeling of knowing studies. *Acta Psychologica, 41*, 365-371.
Hall, J. F. (1954). Learning as a function of word frequency. *The American Journal of Psychology, 67*, 138-140.
Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*(4), 208-216. doi: 10.1037/h0022263
Hertzog, C., & Dunlosky, J. (2011). Metacognition in later adulthood: Spared monitoring can benefit older adults' self-regulation. *Current Directions in Psychological Science, 20*(3), 167-173. doi:
Hertzog, C., Dunlosky, J., & Sinclair, S. M. (2010). Episodic feeling-of-knowing resolution derives from the quality of original encoding. *Memory & Cognition, 38*(6), 771-784. doi: 10.3758/mc.38.6.771
Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging, 17*(2), 209-225. doi: 10.1037//0882-7974.17.2.209
Hertzog, C., Sinclair, S. M., & Dunlosky, J. (2010). Age differences in the monitoring of learning: Cross-sectional evidence of spared resolution across the adult life span. *Developmental Psychology, 46*(4), 939-948.
Higham, P. A. (2007). No Special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General,*

136(1), 1-22.
Hines, J. C., Touron, D. R., & Hertzog, C. (2009). Metacognitive Influences on Study Time Allocation in an Associative Recognition Task: An Analysis of Adult Age Differences. *Psychology and Aging, 24*(2), 462-475. doi: 10.1037/a0014417
Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development, 6*, 97-120.
Hox, J. (2002). *Multilevel analysis techniques and applications*: Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434-446. doi: 10.1016/j.jml.2007.11.007
Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review, 119*(1), 186-200.
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.
Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997. doi:10.2307/2533558
Kim, J.-O. (1971). Predictive measures of ordinal association. *American Journal of Sociology, 76*(5), 891-906.
Kleinsmith, L. J., & Kaplan, S. (1963). Paired-associate learning as a function of arousal and interpolated interval. *Journal of Experimental Psychology, 65*(2), 190-193.
Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology-General, 126*(4), 349-370. doi: 10.1037//0096-3445.126.4.349
Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits--and costs--of dropping flashcards. *Memory, 16*(2), 125-136.
Krug, K. (2007). The relationship between confidence and accuracy: Current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice, 3*(1), 7-41.
Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 464-470.
Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.
Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 149-157. doi: 10.1037/0278-7393.16.1.149
Luna, K., Higham, P. A., & Martin-Luengo, B. (2011). Regulation of Memory Accuracy With Multiple Answers: The Plurality Option. *Journal of Experimental Psychology-Applied, 17*(2), 148-158. doi: 10.1037/a0023276

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19-40. doi: 10.1037/0033-295x.92.3.317.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, Mahwah, NJ.
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition: An*

International Journal, 21(1), 422-430.
Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 509-527.
Merritt, P., Hirshman, E., Hsu, J., & Berrigan, M. (2005). Metamemory without the memory: are people aware of midazolam-induced amnesia? *Psychopharmacology, 177*(3), 336-343. doi: 10.1007/s00213-004-1958-8
Monin, B., & Oppenheimer, D. M. (2005). Correlated Averages vs. Averaged Correlations: Demonstrating the Warm Glow Heuristic Beyond Aggregation. *Social Cognition, 23*(3), 257-278. doi: 10.1037/0022-3514.64.3.431
Muthen, L. K., & Muthen, B. (1998-2012). *Mplus user's guide*. Los Angeles, CA: Author.
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109-133. doi: 10.1037/0033-2909.95.1.109
Nelson, T. O. (1986). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin, 100*(1), 128-132.
Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10*(3), 257-260.
Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2*(3), 325-335. doi: 10.1037/0096-3445.122.1.4710.1037/0096-3445.122.1.471993-20317-001
Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125-173.
Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods, 9*, 53-69.
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*(4), 443-460.
Ozechowski, T. J., Turner, C. W., & Hops, H. (2007). Mixed-effects logistic regression for estimating transitional probabilities in sequentially coded observational data. *Psychological Methods, 12*(3), 317-335. doi: 10.1037/1082-989x.12.3.317
Paivio, A., & Smythe, P. C. (1971). Word imagery, frequency, and meaningfulness in short-term memory. *Psychonomic Science, 22*(6), 333-335.
Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-Plus*. New York: Springer.
Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A General Multilevel SEM Framework for Assessing Multilevel Mediation. *Psychological Methods, 15*(3), 209-233. doi: 10.1037/a0020141
Quene, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413-425. doi: 10.1016/j.jml.2008.02.002
Raaijmakers, J. G. W. (2003). A further look at the "language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 57*(3), 141-151. doi: 10.1037/0278- 7393.21.3.785
Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "The language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*(3), 416-426. doi: 10.1006/jmla.1999.2650

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & Toit, M. (2011). *HLM7: Hierarchical linear and non-linear modeling.* Chicago: Scientific Software International.

Reggev, N., Zuckerman, M., & Maril, A. (2011). Are all judgments created equal? An fMRI study of semantic and episodic metamemory predictions. *Neuropsychologia, 49*(5), 1332-1342. doi: 10.1016/j.neuropsychologia.2011.01.013

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131-148.

Robinson, A. E., Hertzog, C., & Dunlosky, J. (2006). Aging, Encoding Fluency, and Metacognitive Monitoring. *Aging, Neuropsychology, and Cognition, 13*(3-4), 458-478. doi:

Romesburg, H. C. (1984). *Cluster analysis for researchers.* London: Wadsworth, Inc.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70*(2), 389-401. doi: 10.3758/PP.70.2.389

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573-604. doi:

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika, 72*(4), 583-600. doi: 10.1037/0096-3445.133.2.189

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology, 9*(4), 321-332. doi: 10.1002/acp.2350090405

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, *6*, 132-137.

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 1074-1083. doi: 10.1037/0278-7393.17.5.897

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition *Metacognition: Knowing about knowing.* (pp. 93-113). Cambridge, MA, US: The MIT Press, Cambridge, MA.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* New York: John Wiley.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models using SAS Proc MIXED. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*, 512–524. doi:10.1198/108571102726

Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin, 80*(6), 481-488. doi: 10.1037/h0035203

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. London: Sage Publications.

Somers, R. H. (1968). On the measurement of association. *American Sociological Review, 33*(2), 291-292.

Son, L. K., & Kornell, N. (2009). Simultaneous decisions at study: Time allocation,

ordering, and spacing. *Metacognition and Learning, 4*(3), 237-248. doi:
Souchay, C., Moulin, C. J. A., Clarys, D., Taconnat, L., & Isingrini, M. (2007). Diminished episodic memory awareness in older adults: Evidence from feeling-of-knowing and recollection. *Consciousness and Cognition: An International Journal, 16*(4), 769-784. doi:
Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavioral Research Methods, Instruments, & Computers, 31*, 137-149.
Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99*(2), 181-198.
Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York: Academic Press.
Tauber, S. K., & Dunlosky, J. (2012). Can older adults accurately judge their learning of emotional information? *Psychology and Aging, 27*(4), 924-933. doi: 10.1037/a0028447
Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging, 27*(2), 474-483. doi: 10.1037/a0025246
Van den Noortgate, W., & Onghena, P. (2006). Analysing repeated measures data in cognitive research: A comment on regression coefficient analyses. *European Journal of Cognitive Psychology, 18*(6), 937-952. doi: 10.1080/09541440500451526
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.
Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition, 40*(5), 703-716. doi:
Watkins, M. J. (1972). Locus of the modality effect in free recall. *Journal of Verbal Learning & Verbal Behavior, 11*(5), 644-648.
Wilson, T. P. (1974). Measures of association for bivariate ordinal hypotheses. In H. M. Blalock (Ed.), *Measurement in the social sciences* (pp. 327-342). Chicago: Aldine.
Worden, P. E., & Sherman-Brown, S. (1983). A word-frequency cohort effect in young versus elderly adults' memory for words. *Developmental Psychology, 19*(4), 521-530.
Wright, D. B. (1996). Measuring feeling of knowing: Comment. *Applied Cognitive Psychology, 10*(3), 261-268. doi:
Yan, V. X., Murayama, K., & Castel, A. D. (2013, November). The role of preference versus familiarity in incidental and intentional learning. *Poster presented at the 54th annual scientific meeting of the Psychonomic Society*, Toronto, Ontario.
Zimmerman, C. A., & Kelley, C. M. (2010). "I'll remember this!" Effects of emotionality on memory predictions versus memory performance. *Journal of Memory and Language, 62*(3), 240-253. doi: 10.1016/j.jml.2009.11.004

Footnotes

1. Some researchers argue that Type-2 signal detection tasks—where people judge the correctness of their own decisions (Clarke, Birdsall, & Tanner, 1959; Galvin, Podd, Drga, & Whitmore, 2003)—provides a better measure of metacognitive accuracy (Barrett et al., in press; Higham, 2007; Maniscalco & Lau, 2012). Although Type-2 SDT fits well with some specific metacognitive judgment tasks (e.g., confidence judgment of perceptual discrimination task), it is comparable to the standard SDT (also called the Type-1 SDT) in that the model assumes separate distributions for target and distractor items. Therefore, Type-2 SDT also may not be applicable when there are no distractor items, as in common in JOL experiments. In addition, these measures are also subject to the inflation of Type-1 error rates that we will demonstrate.

2. It is possible to further incorporate *random item slopes* (i.e., the inter-item variation of metacognitive accuracy) in our models. In the current manuscript, we limited the term random item effect to refer to random item intercept (u_{0i}) in Equation 8 or 9 (i.e., the inter-item variation of overall recall difficulty), following the conventional use of this term (e.g., Judd et al., 2012; but see Barr, Levy, Scheepers, & Tily, 2013). Random item slopes will be discussed in General Discussion.

3. We also computed G_k and G_w. As these measures showed very similar results with G_s, in this and in the following simulations, for display purposes, we only report the results from G_s, but not from G_k or G_w. The results from these measures can obtained from the authors upon request.

4. When the dependent variable is continuous, there are a few more options to derive *p*-values from mixed-effects models. For example, recent simulation studies showed that the Kenward–Roger approximation (Kenward & Roger, 1997) controls for Type-1 error rate relatively well (Schaalje, McBride, & Fellingham, 2002). Baayen et al. (2008) recommended Markov Chain Monte Carlo methods. These options are, however, not yet well implemented in models with a categorical dependent variable (i.e., generalized mixed-effects model).

5. We thank Dr. Nate Kornell for sharing the data. In their Experiment 2, there were four experimental conditions, but the current manuscript focuses only on one of the conditions (i.e., Drop – JOL condition) as this is the only condition that includes JOLs.

6. In fact, none of the real data used in our examples showed significant random item slopes in our preliminary analyses.

7. As indicated in the literature (Brunner & Austin, 2009; Culpepper & Aguinis, 2011), when there is more than one correlated independent variables, measurement error of one independent variable can inflate the Type-1 error rates of the other predictors in regression analysis. We believe this issue should also apply to the generalized mixed-effects model. In addition, measurement error in the independent variable(s) typically produces attenuated parameter estimates. Although beyond the scope of the paper, multilevel structural equation model (Asparouhov & Muthen, 2012) may potentially address these problems by controlling for measurement error of independent variables.

Figure Captions

Fig. 1. Type 1 error rates as a function of number of participants and number of items in Simulation 1, when random item effect is present. G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. *D* = Hart's difference score. Mixed model z test = *z* value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value (α = .05) is highlighted by the dotted line.

Fig. 2. Statistical power as a function of number of participants and number of items in Simulation 1, when random item effect is present (true slope = 0.2). G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. *D* = Hart's difference score. Mixed model z test = *z* value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model.

Fig. 3. Type 1 error rates as a function of number of participants and number of items in Simulation 2, when random item effect is present. G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. *D* = Hart's difference score. Mixed model z test = *z* value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value (α = .05) is highlighted by the dotted line.

Fig. 4. Type 1 error rates as a function of number of participants and number of items in Simulation 3, when random item effect is present. G_w = corrected gamma correlation proposed by Wilson (1974). Phi = Phi coefficient. Tetrachoric Cor. = tetrachoric correlation. *Logistic* = logistic regression coefficient. d' = a signal detection d prime computed by Equation 3. G_c = a signal detection measure computed by Equation 8. *HC* = Hamann coefficient. *D* = Hart's difference score. Mixed model z test = *z* value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value (α = .05) is highlighted by the dotted line.

Fig. 5. Type 1 error rates as a function of number of participants and number of items in Simulation 4, when random item effect is present. G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. *D* = Hart's difference score. Mixed model z test = *z* value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value (α = .05) is highlighted by the dotted line.

Fig. 6. Type 1 error rates as a function of number of participants and number of items in Simulation 5, when random item effect X group interaction is present. G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. D = Hart's difference score. Mixed model z test = z value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value ($\alpha = .05$) is highlighted by the dotted line.

Fig. 7. Type 1 error rates as a function of number of participants and number of items in Simulation 6, when random item effect X condition interaction is present. G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. D = Hart's difference score. Mixed model z test = z value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value ($\alpha = .05$) is highlighted by the dotted line.

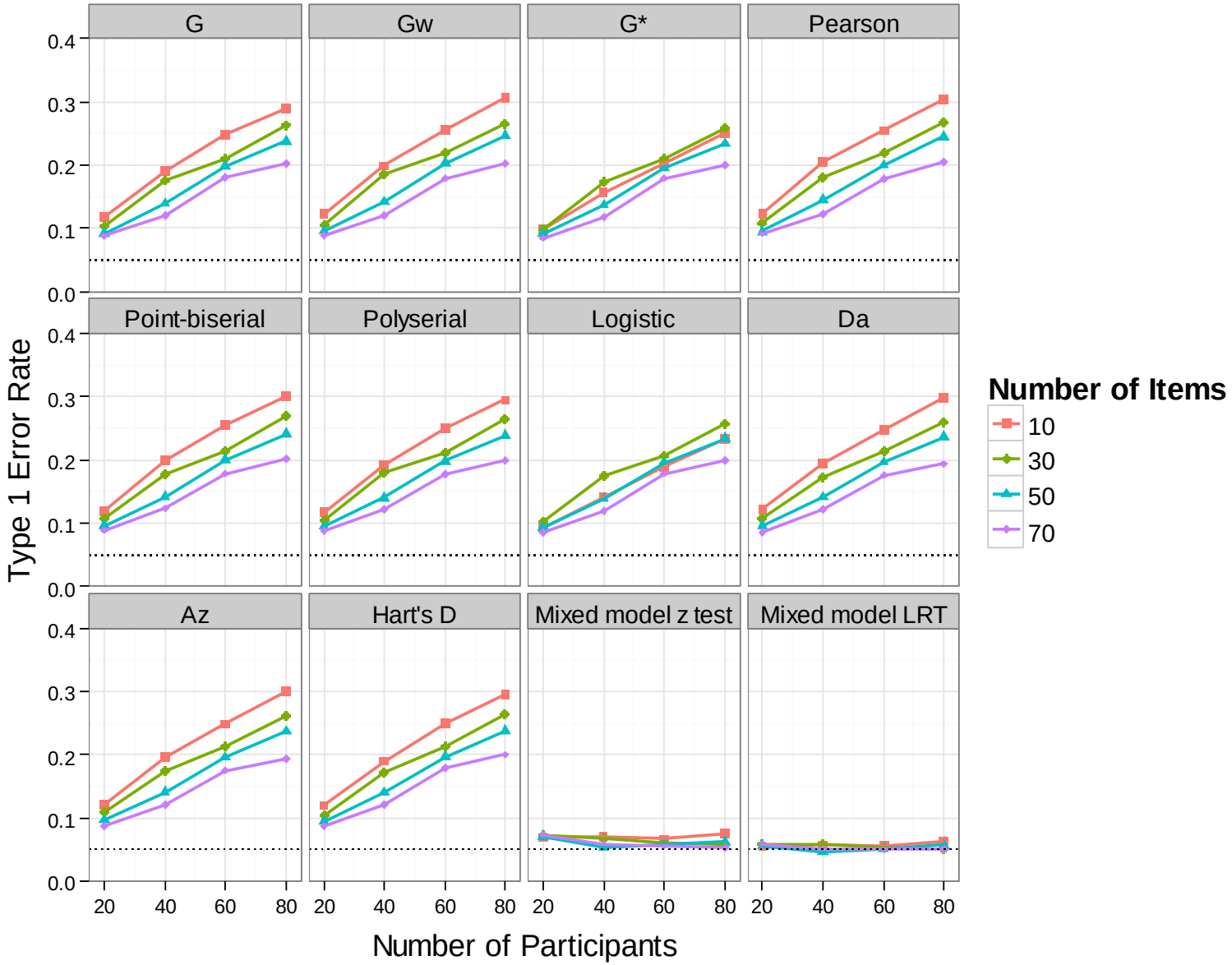

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
20
40
60
80
Number of Participants
Number of Items
10
30
50
70

Figure 1

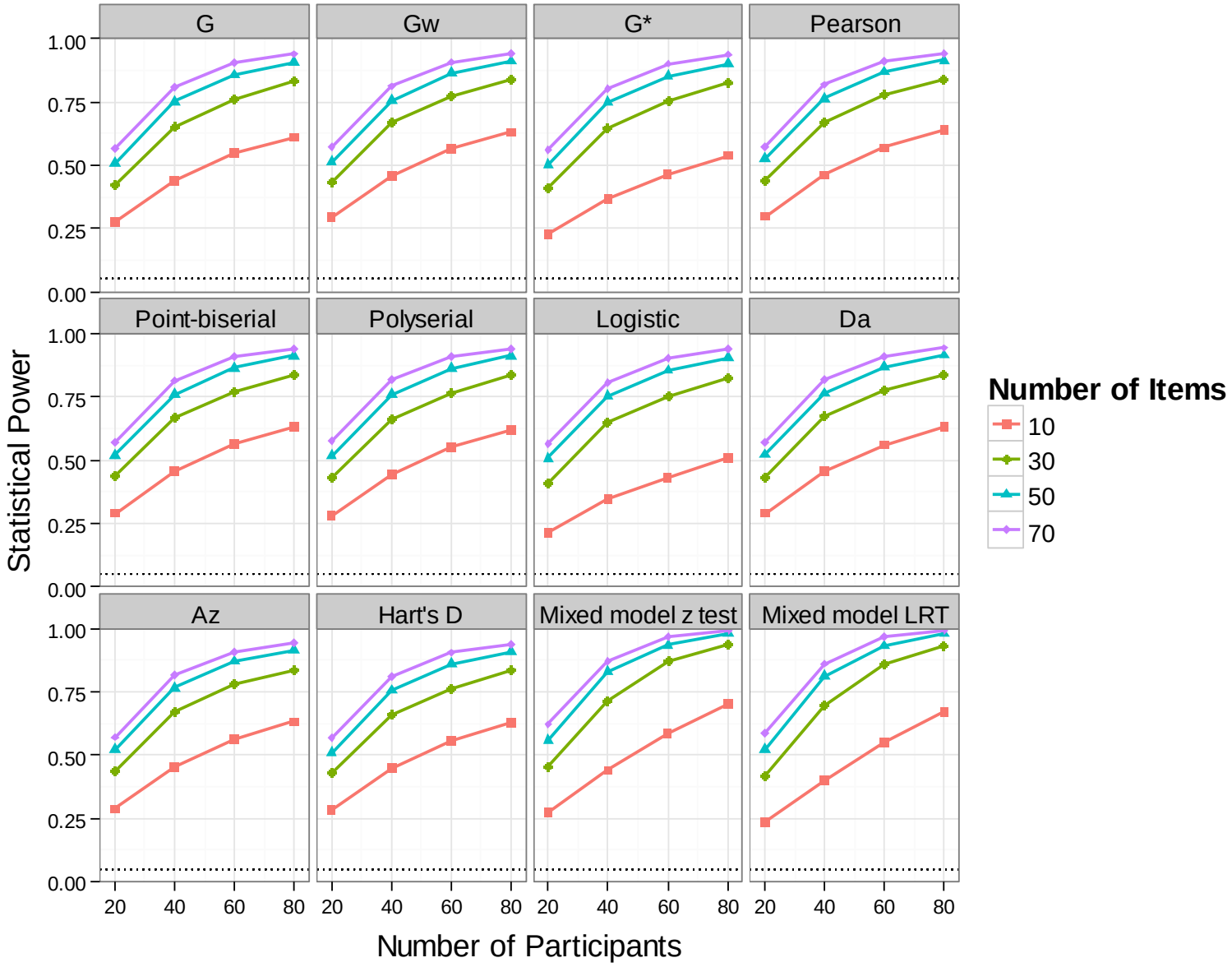

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Statistical Power
0.00
0.25
0.50
0.75
1.00
20
40
60
80
Number of Participants
Number of Items
10
30
50
70

Figure 2

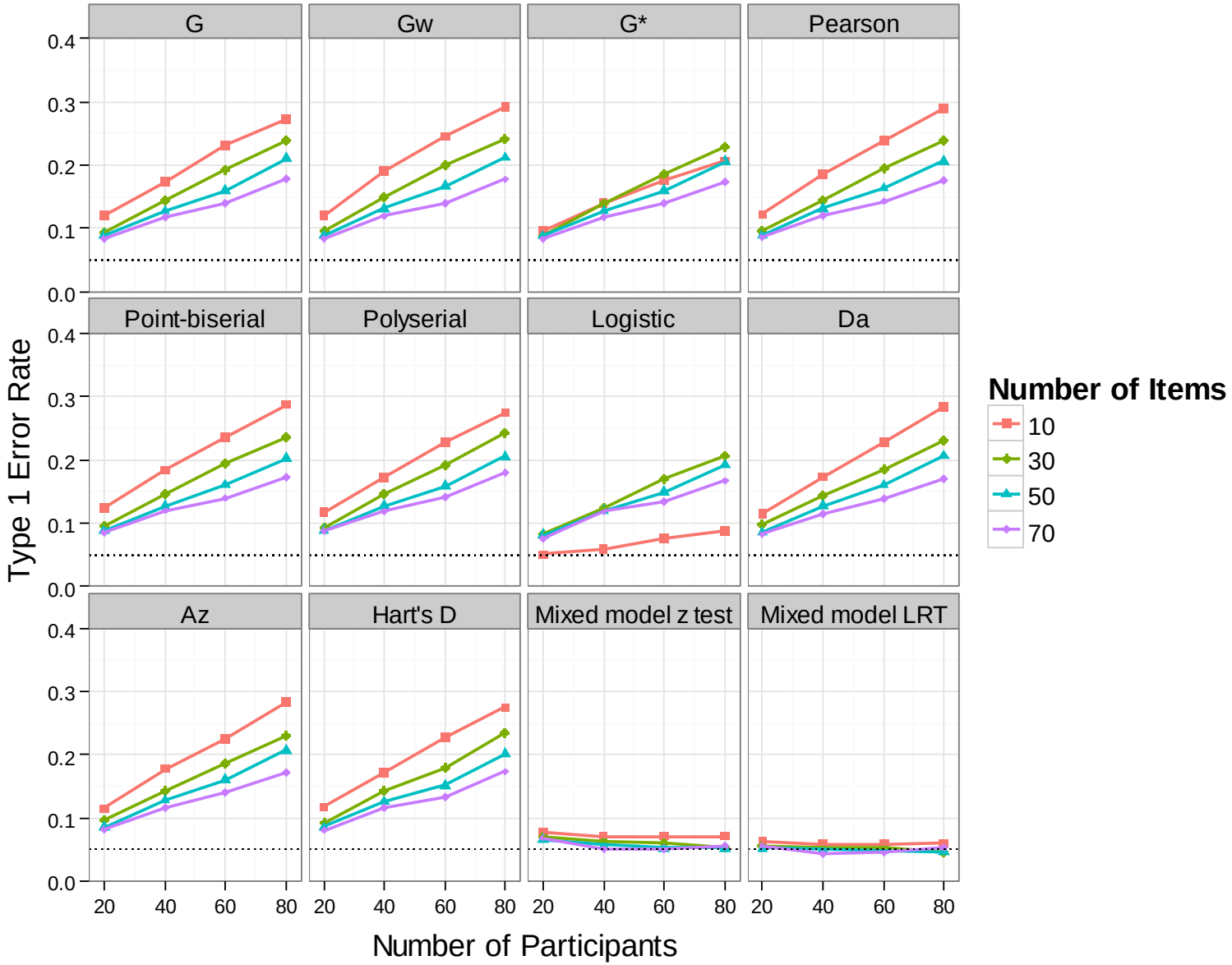

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
20
40
60
80
Number of Participants
Number of Items
10
30
50
70

Figure 3

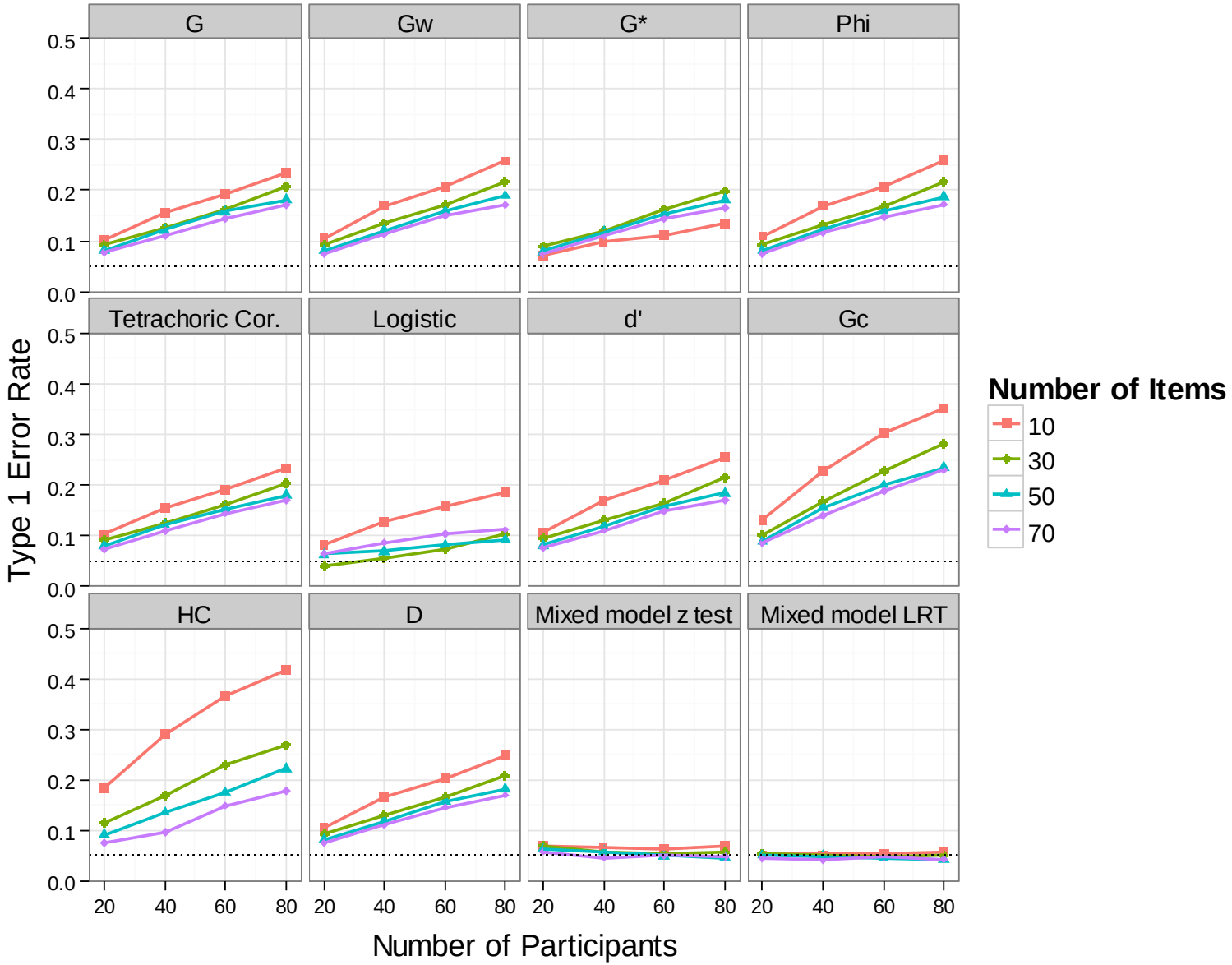

G
Gw
G*
Phi
Tetrachoric Cor.
Logistic
d'
Gc
HC
D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
0.5
20
40
60
80
Number of Participants
Number of Items
10
30
50
70

Figure 4

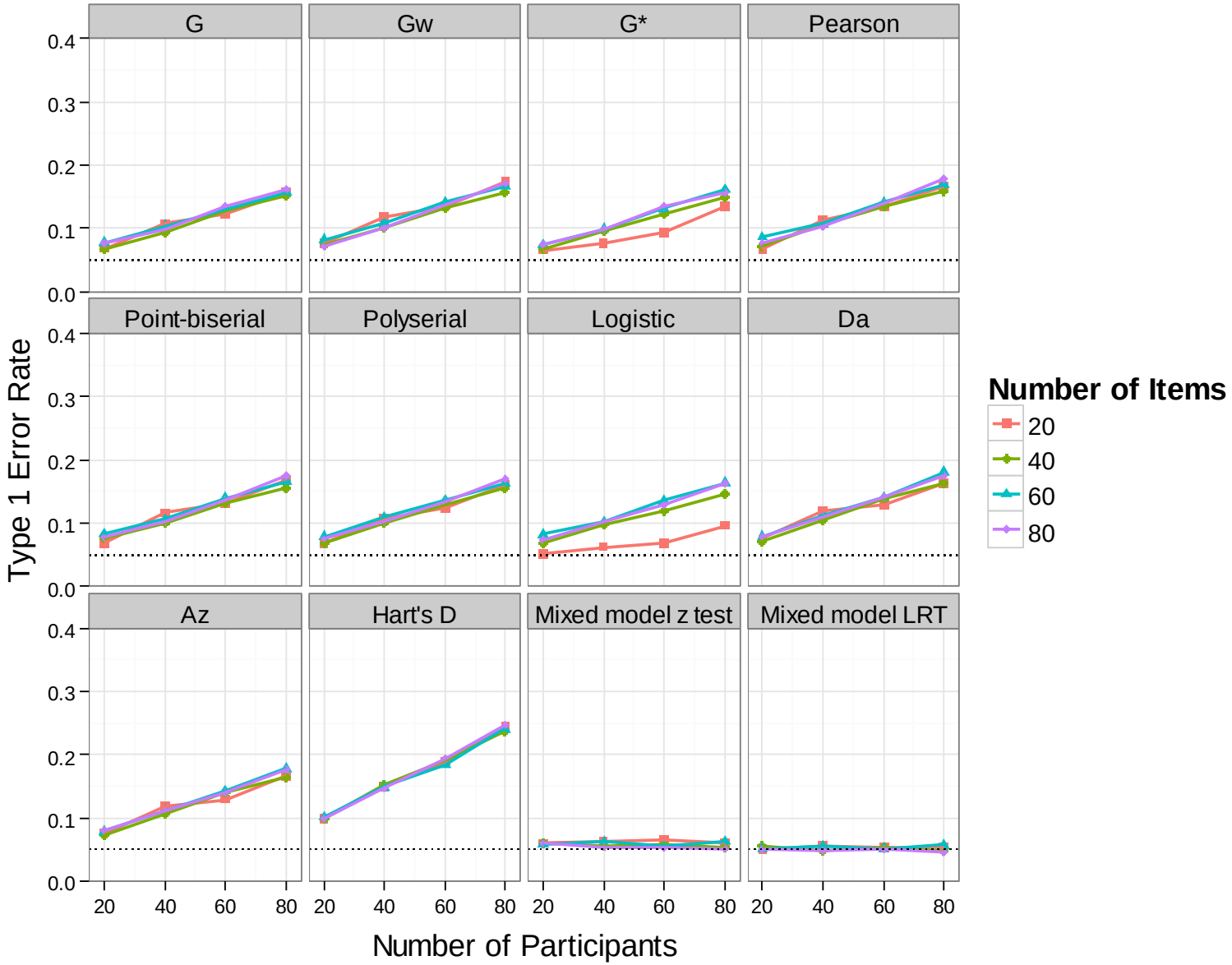

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
20
40
60
80
Number of Participants
Number of Items
20
40
60
80

Figure 5

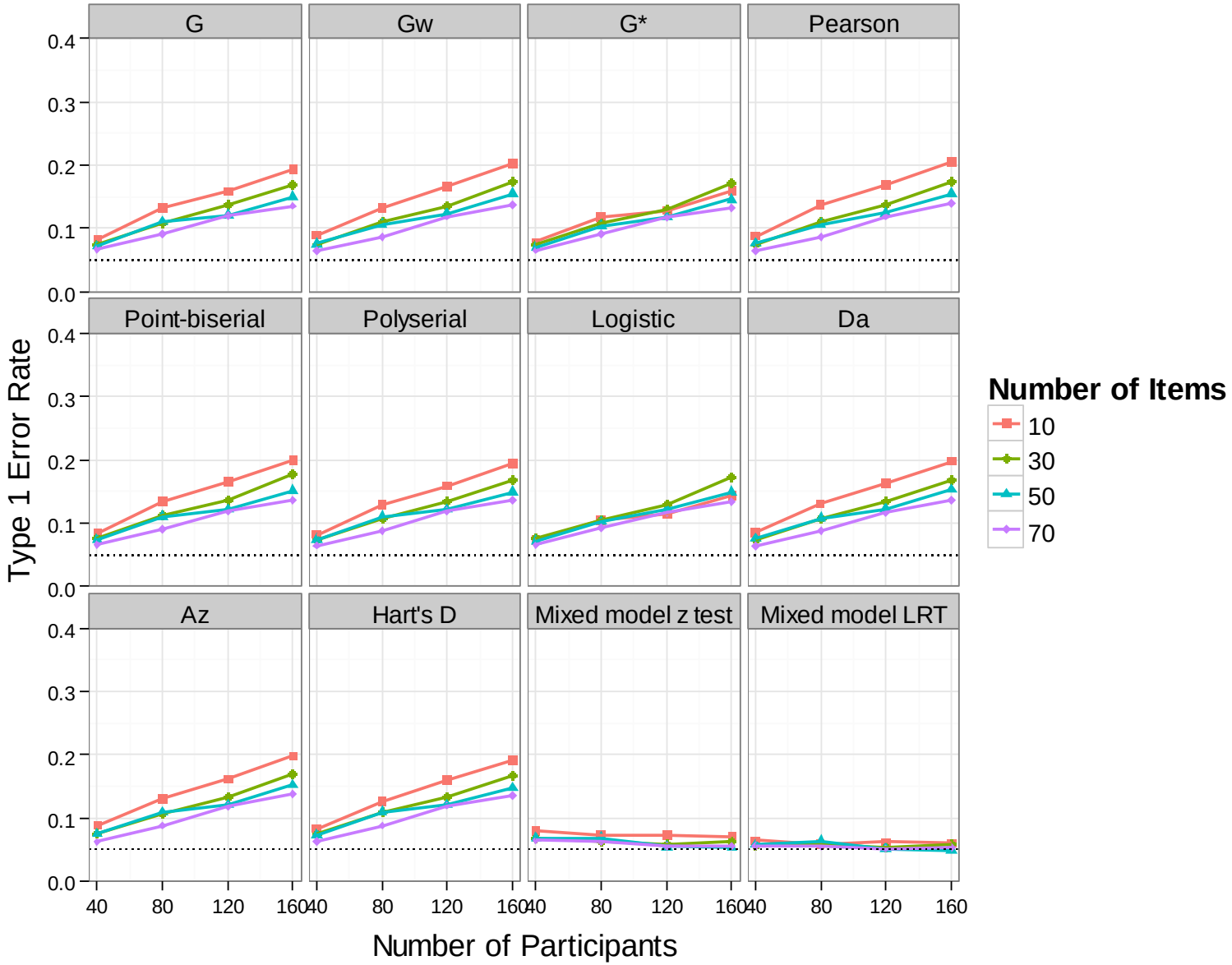

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
40
80
120
160
Number of Participants
Number of Items
10
30
50
70

Figure 6

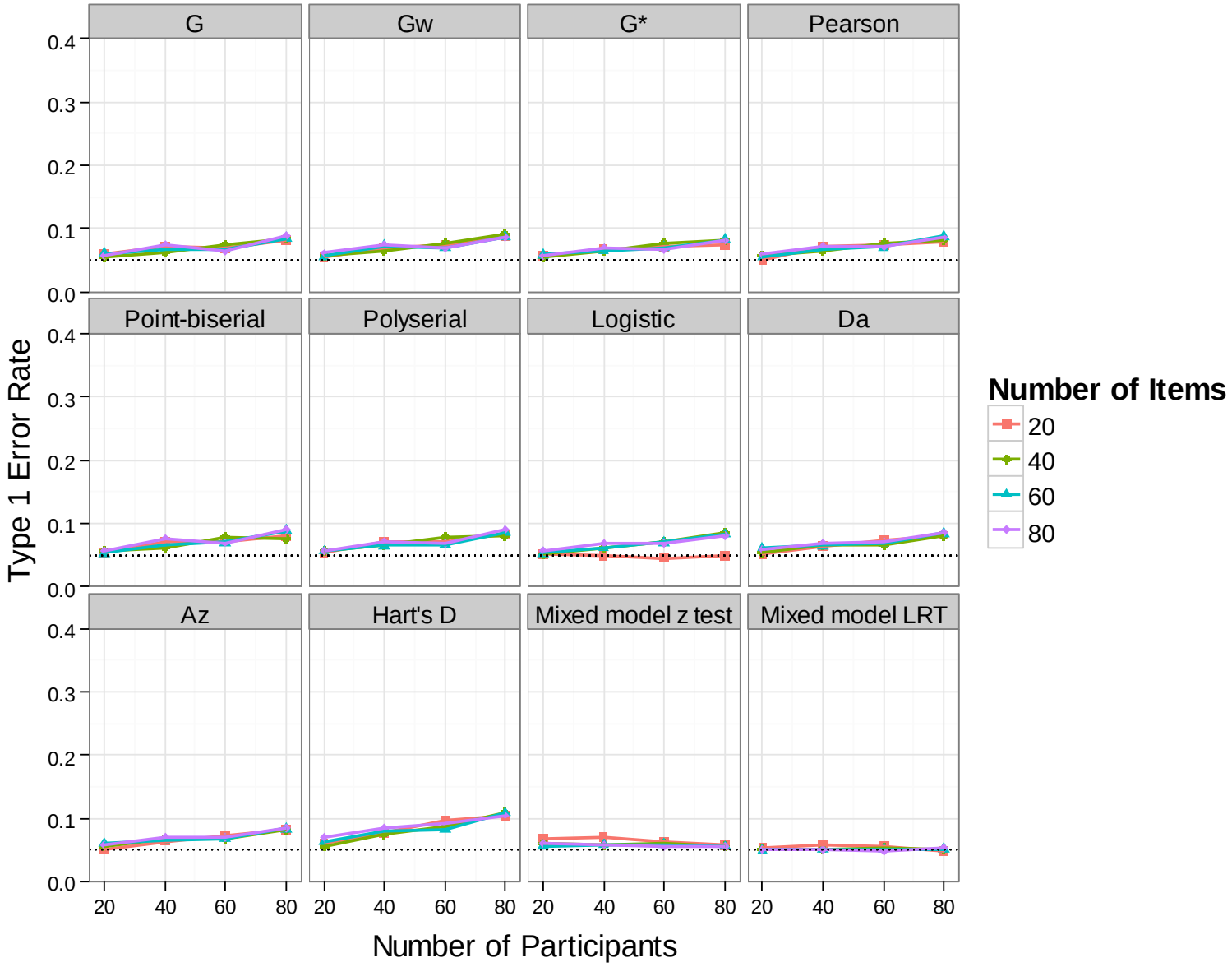

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
20
40
60
80
Number of Participants
Number of Items
20
40
60
80

Figure 7

Supplementary Materials Online

Details of Simulation 1

Hypothetical JOL experiment data were simulated by systematically varying the number of simulated participants (N = 20, 40, 60, or 80) and the number of items (K = 10, 30, 50, 70). In the simulation, for each trial of each participant, we randomly sampled continuous JOL values from a normal distribution with mean = 0 and SD = 1, and computed the corresponding memory strength by considering a random participant effect, a random item effect, a random slope connecting JOLs and memory strength, and random noise. All the random effects were simulated from a normal distribution with mean = 0. Accordingly, the population mean of the random slope is zero, meaning that the simulation assumed no overall relationship between JOLs and memory strength at the population level. The SD of the random noise was held constant across the simulations at 1. We manipulated the absence or the presence of a random item (intercept) effect by setting the SD of the random item effect to 0 (i.e., no item effect variance) or 0.6 (i.e., item effect variance is about one-third of the variance of random noise). Our pilot simulation indicated that the variance of the random participant effect does not influence any of the simulation results. This is a logical consequence of the fact that the by-participant analysis computes relative accuracy based only on covariation within each participant. That is, by computing relative accuracy measurements separately for each participant, between-participants variation in memory performance is effectively eliminated in the by-participant analysis. The mixed-effects model analysis also takes into account random participant effect. Therefore, we arbitrarily set the SD of the random participant effect to 0.6.

The SD of the random slopes was set to 0.3. With the current simulation parameters, a JOL-memory slope of 0.3 approximately corresponds to a correlation of 0.25. By setting random slope SD = 0.3, therefore, the simulation posits that the correlation between JOL and memory strength vary mostly between -.50 to .50 across participants. Note that this is the variation in slope/correlation when there are unlimited number of items—with a limited number of items, the observed variance in the relation between JOLs and memory would be even larger. Finally, many studies found that JOLs are influenced by item characteristics (i.e., intrinsic cues; Koriat, 1997). Accordingly, in order to simulate realistic experiments, the simulation also assumed that the simulated continuous JOL values from the same item were weakly correlated (r = 0.3) across participants.

For each item, we set a threshold value of zero on the memory strength dimension such that any item with a strength value above the threshold was classified as recalled, and all other items were classified as forgotten. It should be noted that the possible fluctuation of the threshold value (i.e., threshold variance) across participants, items or trials is reflected in the random effects in our simulation. In other words, our simulation took the threshold variance into consideration by incorporating different types of random effects. We also set five equal-interval threshold values on the JOL dimension such that the continuous JOLs are mapped onto a 6-point discrete scale, as is frequently done with JOL research (e.g., Dunlosky & Connor, 1997; Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002).

For each simulated experimental dataset with N participants of K items, all the possible measures of metamemory accuracy were computed (i.e., G, G_w, G^*, r_{pb}, r_b, r_{pc}, d_a, Az, and D) for each participant, and these values were entered into a one-sample t-test to test whether the average values were statistically different from chance. In addition, the same dataset was applied to a generalized mixed-effect model using the *lme4* package in *R* (Bates et al., 2011). We used a standard logit-link function to handle the dichotomous

dependent variable. The tested model was equivalent to Equation 9, including three independent random components (i.e., random participant intercept, random item intercept, and random participant slope), and the independent variable (i.e., JOLs) was centered within participants following recommendations in the past literature (Enders & Tofighi, 2007; Hoffman & Stawski, 2009). The primary focus of this model is the statistical significance of the fixed slope value of JOL (β_{10} in Equation 9). There are primary two ways to obtain *p*-values from generalized mixed-effects models (see footnote 4 in the main text). First, we divide the estimated coefficient by its standard error to obtain a *z* value, and judge the effect to be significant if the *z* value surpasses 1.96. Previous studies indicated that such *z* test tends to be lenient in the context of mixed-effects model analysis (Baayen et al., 2008). The second option is to test the fixed slope value by using a log-likelihood ratio test (LRT; Baayen, 2008). Specifically, we applied a mixed-effects model twice to the same data, once with and once without the fixed slope (i.e., a baseline model). We then compared the fit statistics between these two models with a LRT. When a significant improvement of model fit was observed by including the fixed slope effect, we considered the fixed slope as significant. One possible weakness of this approach is that the model has an appropriate, nested baseline model. However, the appropriate baseline is not always available. For example, if a model has two main effects and one interaction between them, the main effect of this model is difficult to test using LRT, because dropping the main effect while keeping the interaction term yields results in an inappropriate model.

The total number of replications (i.e., simulated experiments) was 5,000 for each combination of the parameters. Alpha was set to 0.05 throughout the simulation.

Previous studies indicated that the conventional *z*-test tends to be lenient in the context of mixed-effects model analysis (Baayen et al., 2008). Accordingly, we additionaly tested the fixed slope value by using a log-likelihood ratio test (Baayen, 2008). Specifically, we applied a mixed-effects model twice to the same data, once with and once without the fixed slope. We then compared the fit statistics between these two models with a log-likelihood ratio test. When a significant improvement of model fit was observed by including the fixed slope effect, we considered the fixed slope value as significant. The total number of replications (i.e., simulated experiments) was 5,000 for each combination of the number of participants (N = 20, 40, 60, or 80), the number of items (K = 10, 30, 50, 70), and the random item effects (SD = 0 or 0.6). Alpha was set to 0.05 throughout the simulation.

The simulations presented in this paper were computed in parallel, using 120 CPU cores provided by a dedicated hybrid CPU-GPU Infiniband compute cluster as well as 13 high-performance analysis laboratory workstations; both computer platforms are hosted at the Centre for Integrative Neuroscience and Neurodynamics, University of Reading, UK.

Logistic Regression Coefficients and Type 1 Error Rates

Throughout the simulations, the regression coefficient of logistic regression analysis (*logistic B*) consistently showed slightly lower Type 1 error rates than other measures using by-participant analysis. These results do not mean that *logistic B* is resistant to the Type-1 error rate inflation caused by random item effects. The lower Type 1 error rates actually came from the fact that logistic regression model cannot uniquely estimate the regression coefficient when predictors can perfectly separate the occurrence and absence of a binary outcome (called “linear separation problem). This issue is not so common when we have many observations. In the context of metamemory research and our simulation, however, logistic regression is applied to each individual with a relatively small number of items, possibly resulting in the omission of relatively large number of

participants. As a consequence, *logistic B* showed artificially deflated Type 1 rates in our simulations.

Adjusted Power Analysis in Simulation 1

Although Figure 2 presented the statistical power of different methodologies, it is misleading to directly compare the power of different approaches that differ in Type 1 error rate, because the power of anticonservative approaches will be inflated. An ideal statistical analysis method maximizes statistical power while keeping Type 1 error nominal. To make a fair comparison across the methodologies, we also calculated an adjusted power (Barr, Levy, Scheepers, & Tily, 2013), a power rate corrected for the difference in Type-1 error rates. Specifically, for each methodology, we first obtained the *p*-value at the 5% quantile of the empirical *p*-value distribution yielded in Simulation 1 (i.e., simulation with the null hypothesis). We used this *p*-value as the cutoff for rejecting the null hypothesis for the given methodology in the statistical power simulation (i.e., true slope = 0.2). As illustrated in Figure S2, the adjusted power analysis showed much higher statistical power for mixed-effects model than for by-participant analyses, further indicating the advantage of mixed-effects modeling.

G and Type 1 Error Rates in Simulation 3*

One anomalous observation in Simulation 3 (Figure 2) is that Type-1 error rates of *G** inflate at a slower rate than the other measures as the number of participants increased. This is caused by a special property of computing *G**. As shown in Equation 2, *G** cannot be computed (i.e., it is treated as missing) when $G = -1$ or $G = 1$. With a small number of categories like in the current simulation, *G* is likely to take such extreme values especially when the number of items is small, and this increases number of missing data points. Accordingly, *G** has a smaller number of participants contributing to the group-level inferences in comparison to other metamemory measurements, resulting in smaller Type-1 error rates. Given that participants with $G = -1$ or 1 have certain meaningful information about their metacognitive accuracy, such omissions are not considered a desirable characteristic of *G**.

Effects of Random Slope Variance on Type1 error Rate

As a supplementary analysis, we examined the effects of random slope variance (i.e., the variation of JOL-memory relations across participants). All the simulations we conducted posited a variance in true slopes (across participants), and this simulation aimed to examine the impact of this assumption. If the slope variance were smaller, the computed metamemory accuracy measures (e.g., *G*) would vary less across participants. Accordingly, we could expect an even higher chance of finding a (false) significant effect. To confirm this point, we simulated the same set of experiments in Simulation 1 with random slope SD = 0, 0.15, and 0.30 (in our original simulation, we used 0.30). Table S1 reports the observed Type-1 error rates with $N = 40$. The results with $N = 20$, 60, and 80 are available from the authors upon request. Consistent with our prediction, the results indicated that the traditional by-participant analysis produces higher Type-1 error rates when there is a smaller random slope variance. Another interesting observation is that, when random slope variance is small, increasing the number of items does not mitigate the inflation of Type-1 error rates. This may be because increasing the number of items would also decrease the sampling variation of metacognitive accuracy measurements (e.g., *G* becomes stable with many items), which ironically, enhances the chance of detecting small non-zero artefactual association. In other words, increasing the number of items drives two opposite forces, and these two effects are somewhat balanced out when random slope variance is zero.

These findings can explain why increasing the number of items did not alleviate the inflation of Type-1 error rate in Simulation 4 and Simulation 6. In Simulations 4 and 6,

although the simulations posited random slopes across participants, each participant had the same random slope across the conditions. Accordingly, in these particular experimental setups, the two random slopes within each participant cancel each other out when examining the difference between the two conditions. As our supplementary simulation above indicated, without variance in random slopes across participants, the number of items tends to have no influence on Type-1 error rates. The weak effect of the number of items in Simulations 4 and 6 can be explained by such factors.

Interestingly, none of our real data examples (Examples 1-3) showed statistically significant random slopes of participants. This supplementary simulation suggests that such situation would exacerbate the inflation of Type-1 error rates caused by the by-participant analysis in the presence of random item effect.

Details of the Real Data Example 3

Yan, Murayama, and Castel (2013) examined how personal preference of to-be-remembered items contributes to subsequent memory performance with an intentional learning paradigm. Ninety-one participants were recruited from Amazon.com's Mechanical Turk. The learning materials were 16 popular ice cream flavors (e.g., strawberry, coconut). At the beginning of the study, participants were told that they would be presented with ice cream flavors that they would later be asked to recall. Participants were then shown the 16 flavors one at a time, for seven seconds each. The order of the flavors was randomized for each individual. In each trial, one of the flavors appeared in the middle of the screen, with a textbox underneath and a "Liking (1-10)?" prompt. Participants were asked to rate each flavor during its presentation on a scale of 1 (I really don't like this flavor/least favorite) to 10 (I love this flavor/one of my favorites). This was followed by a 30 second distractor task, and then a free recall test. Participants were asked to recall the flavors they had studied, regardless of whether they liked or disliked them for 90 seconds. After recall, participants were shown the 16 ice cream flavors sequentially again (in a randomized order), and were asked to rate the familiarity of each flavor in their daily experience from 1 (not at all familiar) to 10 (very familiar). Participants were reminded that familiarity was not the same as liking. Below each flavor was a prompt, "Familiarity (1-10)?" and a text box in which they entered their responses.

***R* Code**

All the mixed-effects models conducted in this paper were performed using the *lme4* package in *R* (Bates et al., 2011). The codes used in our simulations are described below. Note that we dropped the covariance between random components from all the models (see Barr et al., 2013). We also did not estimate all the possible random components, but focused on main random components that are likely to be present. These decisions were made in order to avoid non-convergence of parameter estimates in our research, but researchers should be careful in specifying their random components, taking into account both the nature of experimental designs and the number of observations (see General Discussion). It is also possible to statistically test the presence of these random components. In all the codes, `m`, `sub`, `item`, and `JOL` are the variables that represent memory performance (0 = forgotten, 1 = recalled), participants, items, and metamemory judgments (e.g., JOLs), respectively.

Simulations 1 - 3: A Single-Group Case. In this simple model, three random components were specified: random participant effect (intercept), random item effect (intercept), random JOL slope of participants.

```
glmer(m ~ 1 + JOL + (1 | sub) + (-1 + JOL | sub) + (1 | item), family=binomial
```

```
(link="logit"))
```

Simulation 4: A Case with Comparing Two Within-participant Conditions with a Between-Item Manipulation. In this model, `condition` represents two experimental conditions (-1 or 1) and `JOLXcondition` represents the interaction effect between condition and metamemory judgments on memory performance. `JOLXcondition` is the focal effect that represents the difference in metacognitive accuracy (i.e., relationships between metamemory judgments and memory performance) between the conditions. The model now included four independent random components: random participant intercept, random item intercept, random participant slope of JOL, and random participant effect of experimental condition.

```
glmer(m ~ 1 + JOL + condition + JOLXcondition + (1 | sub) + (-1 + JOL | sub) + (-1 + condition | sub) + (1 | item), family=binomial (link="logit"))
```

Simulation 5: A Case with Comparing Two Between-participant Groups with a Within-Item Manipulation. In this model, the critical component in the context of current paper is the random item effect X condition interaction, which is a random item slope of the condition effect (`condition | item`). The model included four independent random components: random participant intercept, random item intercept, random participant slope of JOL, and random item slope of group.

```
glmer(m ~ 1 + JOL + condition + JOLXcondition + (1 | sub) + (-1 + JOL | sub) + (1 | item) + (-1 + condition | item), family=binomial (link="logit"))
```

Simulation 6: A Case with Comparing Two Within-participant Groups with a Within-Item Manipulation. Again, in this model, the critical component in the context of current paper is the random item effect X condition interaction, which is a random item slope of the condition effect (`condition | item`). The model now included five independent random components: random participant intercept, random item intercept, random participant slope of JOL, random participant slope of condition, and random item slope of condition.

```
glmer(m ~ 1 + JOL + condition + JOLXcondition + (1 | sub) + (-1 + JOL | sub) + (-1 + condition | sub) + (1 | item) + (-1 + condition  | item), family=binomial (link="logit"))
```

References: Supplementary Materials Online

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, England: Cambridge University Press.

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. doi: 10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C. and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes (Version R package version 0.999375-39). Retrieved from http://CRAN.R-project.org/package lme4

Dunlosky, J., & Connor, L. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory & Cognition*, *25*, 691-700. doi: 10.3758/bf03211311

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138. doi: 10.1037/1082-989x.12.2.121

Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging*, *17*, 209-225. doi: 10.1037//0882-7974.17.2.209

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development, 6*, 97-120.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology-General, 126*, 349-370. doi: 10.1037//0096-3445.126.4.349

Table S1
Type-1 Error Rates As a Function of Random Slope Variance

	Random slope SD = 0				Random slope SD = 0.15				Random slope SD = 0.3			
	K = 10	*K* = 30	*K* = 50	*K* = 70	*K* = 10	*K* = 30	*K* = 50	*K* = 70	*K* = 10	*K* = 30	*K* = 50	*K* = 70
G	.226	.221	.235	.226	.218	.204	.194	.199	.196	.165	.155	.141
G_k	.234	.235	.247	.238	.229	.211	.201	.201	.205	.175	.151	.146
G_s	.226	.222	.23	.227	.222	.20	.192	.199	.194	.164	.153	.141
G_w	..236	.235	.248	.238	.229	.213	.200	.201	.204	.175	.152	.146
*G**	.191	.218	.230	.226	.178	.204	.191	.198	.164	.161	.149	.142
r_{pb}	.240	.237	.245	.236	.230	.214	.202	.202	.204	.174	.154	.148
r_b	.235	.2	.240	.236	.226	.210	.195	.200	.198	.173	.154	.144
r_{pc}	.228	.231	.235	.233	.221	.208	.192	.199	.197	.166	.153	.144
Logistic	.179	.218	.229	.231	.170	.201	.190	.195	.153	.166	.148	.141
d_a	.231	.235	.236	.228	.224	.208	.190	.196	.199	.166	.151	.142
A_z	.233	.236	.237	.229	.224	.209	.191	.195	.200	.165	.149	.142
D	.230	.224	.239	.233	.225	.206	.189	.199	.196	.166	.152	.141
Mixed model *z*	.050	.045	.045	.049	.054	.054	.056	.054	.059	.054	.061	.062
Mixed model LRT	.048	.045	.045	.049	.052	.052	.054	.051	.056	.052	.059	.060

Note. *K* = number of participants. *G* = Goodman & Kruskal's gamma correlation. G_k = corrected gamma correlation proposed by Kim (1971). G_s = corrected gamma correlation proposed by Somers (1968). G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by

Equation 5. *D* = Hart's difference score. Mixed model *z* test = *z* value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model.

Figure Captions

Fig. S1. Type 1 error rates as a function of number of participants and number of items in Simulation 1, when random item effect is absent. G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. D = Hart's difference score. Mixed model z test = z value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value (α = .05) is highlighted by the dotted line.

Fig. S2. Adjusted statistical power (see text for further explanation) as a function of number of participants and number of items in Simulation 1, when random item effect is present (true slope = 0.2). G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. D = Hart's difference score. Mixed model z test = z value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model.

Fig. S3. Type 1 error rates as a function of number of participants and number of items in Simulation 5, when random item effect is present but random item effect X group interaction is absent. G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. D = Hart's difference score. Mixed model z test = z value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value (α = .05) is highlighted by the dotted line.

Fig. S4. Type 1 error rates as a function of number of participants and number of items in Simulation 6, when random item effect is present but random item effect X group interaction is absent. G_w = corrected gamma correlation proposed by Wilson (1974). r_{pb} = point-biserial correlation. r_b = biserial correlation. r_{pc} = polychoric correlation. *Logistic* = logistic regression coefficient. d_a = a signal detection measure with unequal variance computed by Equation 4. A_z = a signal detection measure with unequal variance computed by Equation 5. D = Hart's difference score. Mixed model z test = z value more than 1.96 with mixed-effects model. Mixed model LRT = Log-likelihood ratio test with mixed-effects model. The predetermined alpha value (α = .05) is highlighted by the dotted line.

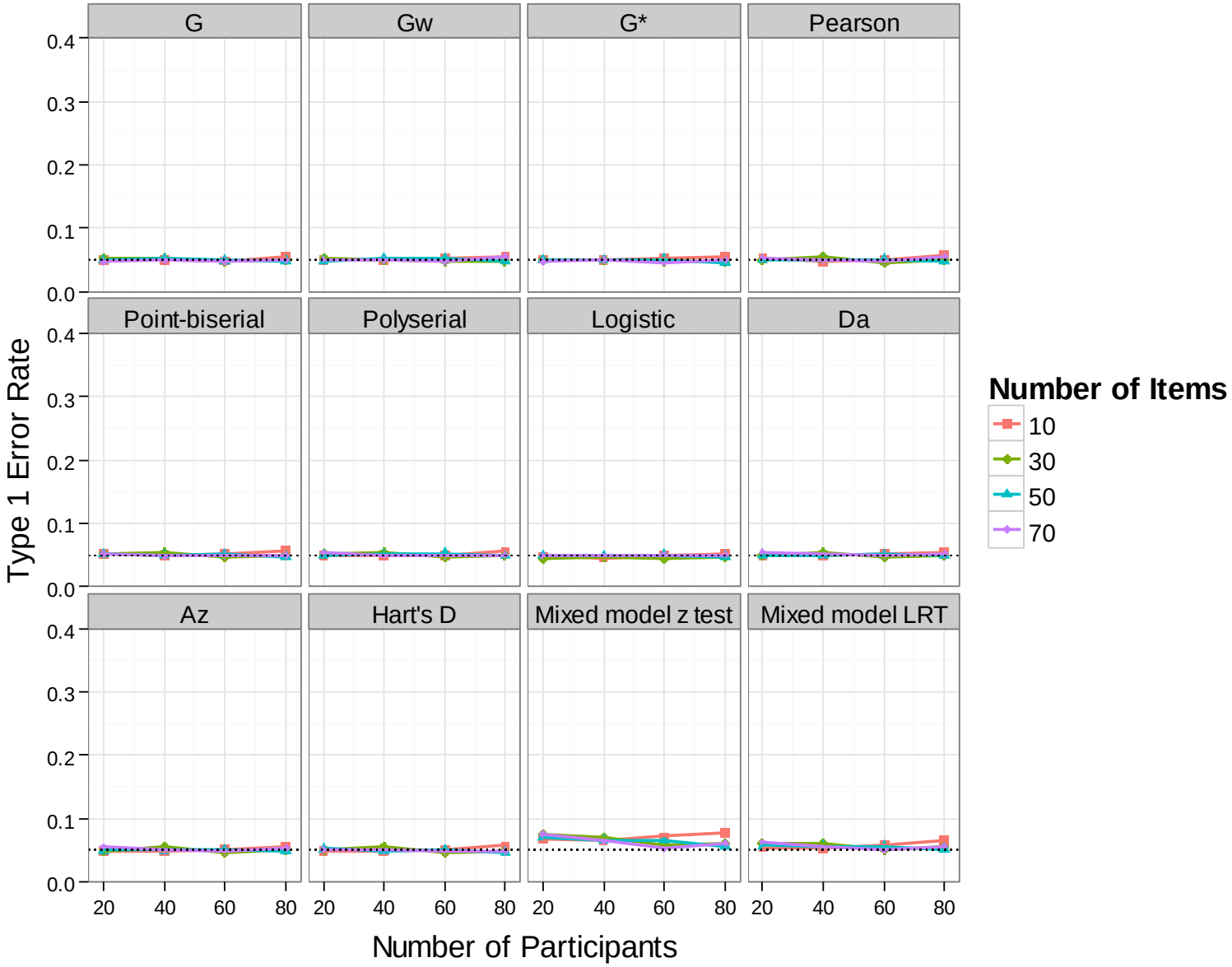
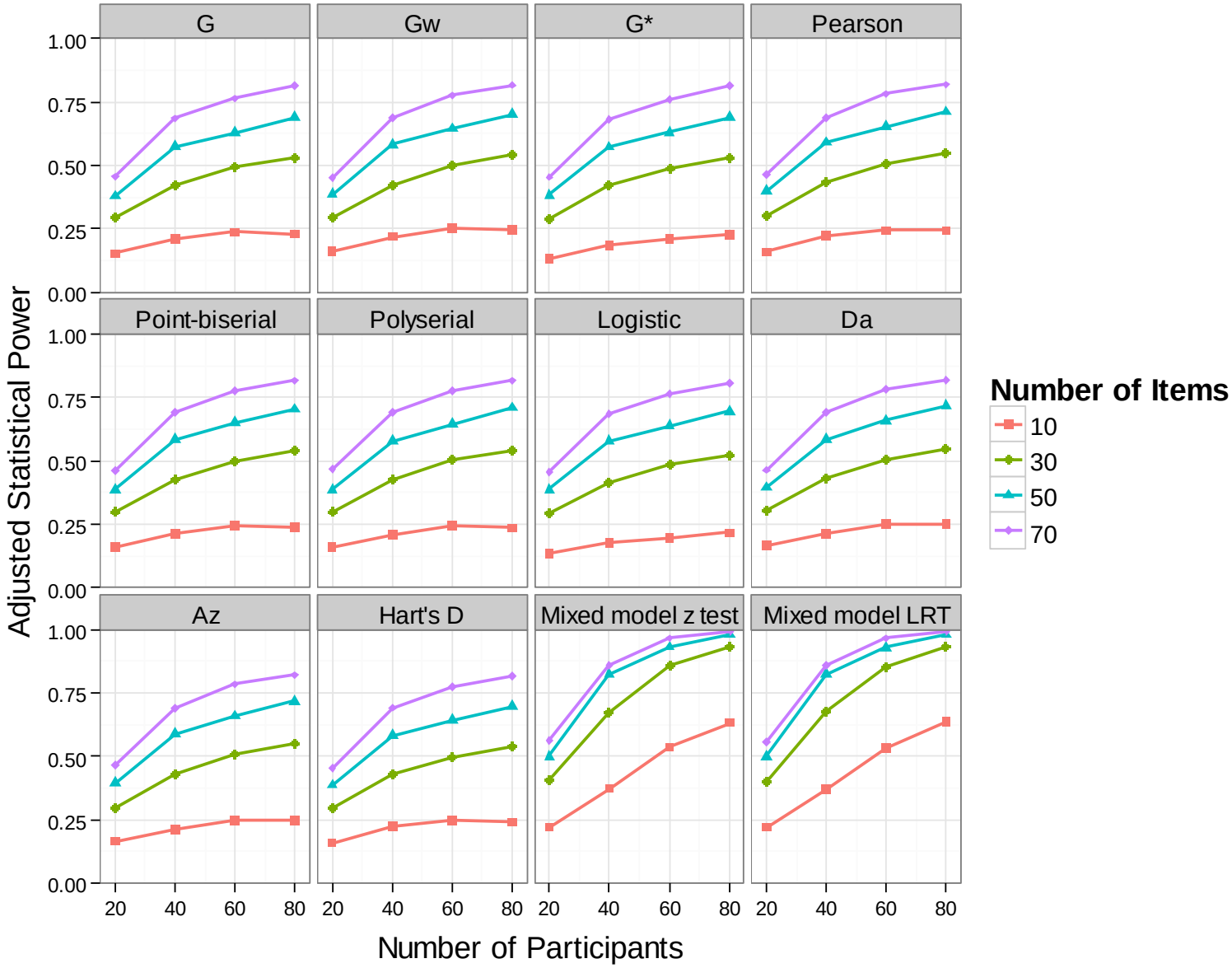

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
20
40
60
80
Number of Participants
Number of Items
10
30
50
70

Figure S1

Figure S2


G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Adjusted Statistical Power
0.00
0.25
0.50
0.75
1.00
20
40
60
80
Number of Participants
Number of Items
10
30
50
70

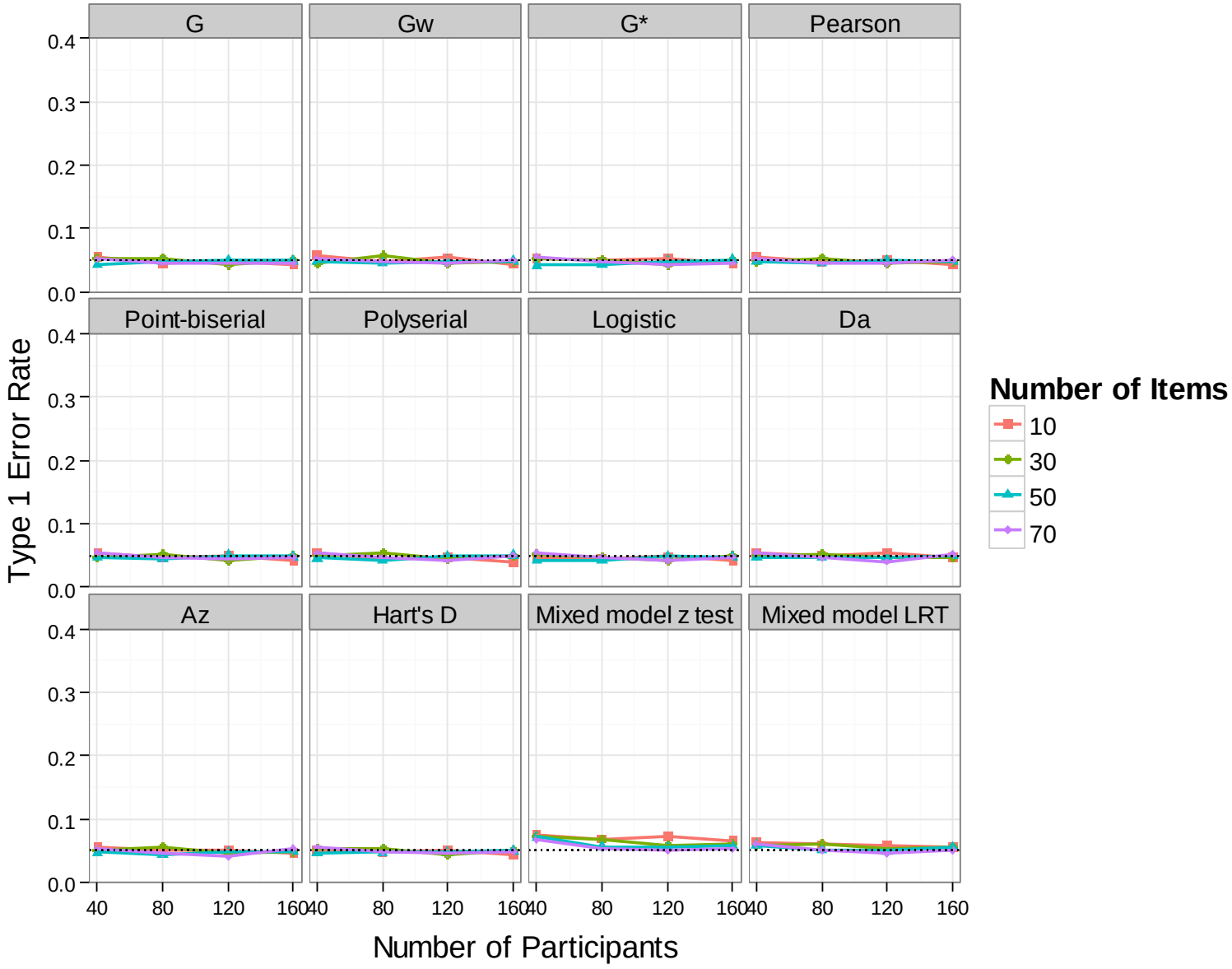

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
40
80
120
160
Number of Participants
Number of Items
10
30
50
70

Figure S3

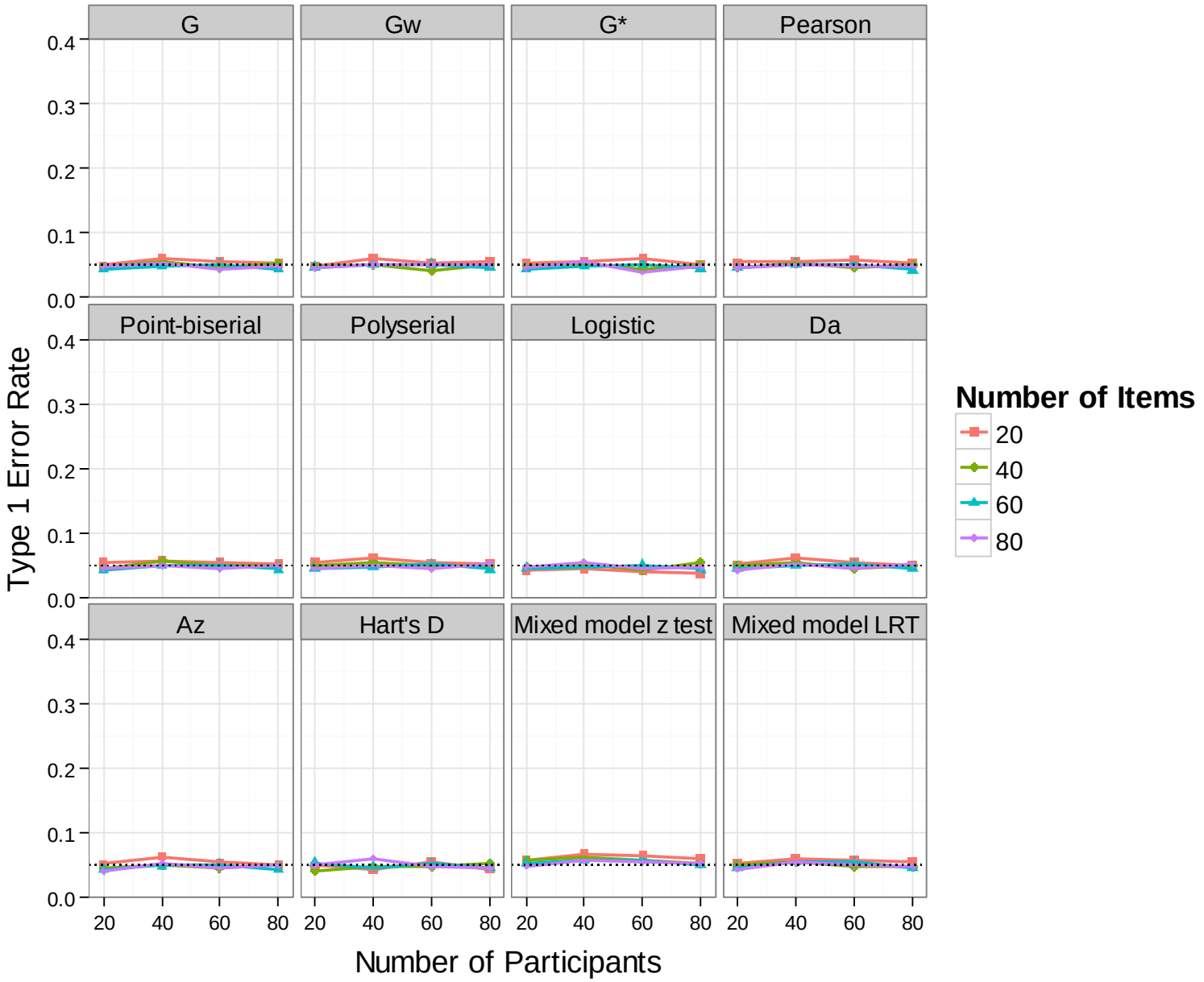

G
Gw
G*
Pearson
Point-biserial
Polyserial
Logistic
Da
Az
Hart's D
Mixed model z test
Mixed model LRT
Type 1 Error Rate
0.0
0.1
0.2
0.3
0.4
20
40
60
80
Number of Participants
Number of Items
20
40
60
80

Figure S4