Behavior Research Methods (2025) 57:154
https://doi.org/10.3758/s13428-025-02637-y

NOTES/COMMENTS/REPLY MANUSCRIPT

The keyboards are (still) all right in response time experiments

Pablo Gómez[1,2] · **Manuel Perea**[3,4] · **Ana Baciero**[5]

Accepted: 19 January 2025

Abstract
Response times (RTs) are a ubiquitous variable for assessing cognitive and motor processes. However, variability introduced by keyboards, especially in online experiments, has raised concerns among behavioral researchers. Here, we evaluate the impact of keyboard delays on RT measurements using linear mixed-effects models and grouped data *t*-tests through a series of simulations. The results showed that the impact of keyboard delays on statistical power is minimal in most cases. Keyboard-induced variability does not inflate type I error rates and has a negligible impact on power, except in rare scenarios of RT distribution shifts or in studies focused on individual differences with low signal-to-noise ratios. Thus, commercially available keyboards remain suitable for most RT experiments, including those conducted online.

Keywords Response time · Online experiments · Keyboards

Response times (RTs) are often used to measure various cognitive and motor processes, with longer RTs typically indicating greater processing difficulty (see Luce, 1991, for a review of early research). As a result, RTs serve as a critical dependent variable in numerous experimental paradigms involving human participants across diverse fields, including psychology, human factors, medicine, and education. In RT experiments, most data have been collected using mass-produced, commercially available keyboards, which might introduce variability to the RTs due to their different polling and scanning rates. This variability may be exacerbated now that RT experiments are increasingly conducted online (see Rodd, 2024, for a review), where researchers have limited control over the hardware used for data acquisition. These hardware issues raise concerns about potential measurement errors caused by keyboard delays. This has generated a somewhat prevalent but vague worry among researchers regarding their impact on the inferential analyses. The goal of this paper is to reduce the vagueness of such concerns and to reassure researchers that, in most cases, using keyboards is acceptable when the goal is to compare the differences between RTs in two or more experimental conditions (see Damian, 2010); evidently, if our goal were to obtain the true RT, the aforementioned noise would be undesirable.

We can illustrate the issue in question with a well-meaning statement from a review that has been edited for anonymity:

> *I have some doubts about the use of a keyboard to collect the participant responses. Keyboards introduce random errors in reaction time (RT) measurements. As shown by* Forster and Forster (2003)*, a 5-ms standard deviation of random error can significantly affect results when effects are only 10–20 ms. While the current data may not be systematically distorted, this raises concerns about replicability. The authors should acknowledge this issue or re-run an experiment using a mouse or a PIO-12 response box.*

Similarly, in journal articles advocating the construction or use of response boxes (e.g., Forster & Forster, 2003; Voss et al., 2007), the authors correctly point out that USB-based responses have a delay compared to real-time parallel port response acquisition methods; however, no discussion of the size or consequences of this problem is presented.

✉ Pablo Gómez
pgomez@skidmore.edu

[1] Skidmore College, 815 N Broadway, Saratoga Springs, NY 12866, USA

[2] California State University San Bernardino, Palm Desert Campus, Palm Desert, CA, USA

[3] Universitat de València, Valencia, Spain

[4] Universidad Nebrija, Madrid, Spain

[5] Universidad Complutense de Madrid, Madrid, Spain

Published online: 23 April 2025

In this paper, we hope to show that collecting data via keyboards is an appropriate practice in most cases; while this is not a new issue, it deserves renewed attention. Three decades ago, Ratcliff (1994) stated that the "variability introduced through the use of keyboards for response collection [was] a frequently mentioned, but ill-understood problem" (p. 95). Through a simple example in which the RT yields a standard deviation (SD) of 200 ms and the keyboard has an SD of 50 ms, Ratcliff pointed out that the measured SD is $\sqrt{200^2 + 50^2}$ (i.e., approximately 6 ms more). In other words, even if the SD added by the keyboard lag is one quarter of that from the latency (50/200 = ¼), the measured SD is only 3% larger than the real SD. Perhaps it is this counterintuitive proportion that gives rise to the above concern (see Damian, 2010, for a similar point).

To illustrate Ratcliff's (1994) statement, we present Fig. 1, which shows the measured SD in the data on the *y*-axis for different SDs in the RT (*x*-axis) and different keyboard delays (lines). The graph shows that in most situations, there is a barely noticeable increase in the *measured* SD. Critically, this was so even at unrealistically high keyboard delays and unusually low SDs in *real* response time experiments. Note that we chose the range of SD in the data using, as the low end, the latency of corrective saccades (as measured by eye trackers; Hollingworth et al., 2008), which is SD ≈ 30 ms, and as the high end, the latency in a tactile flanker task (Baciero et al., 2021), SD ≈ 380 ms.

Since the publication of Ratcliff's (1994) article, two significant developments have motivated us to revisit this issue. The first is the emergence of a wide range of hardware component qualities and the universal use of USB and Bluetooth keyboards instead of interrupt-based mechanisms like PS/2 keyboards. The prevalence of online RT experiments makes this issue even more timely. The second development is a shift in statistical practices, with the widespread adoption of linear mixed-effects models on raw RT data, where items and participants are considered random effects with a slope and an intercept. Indeed, in most fields of cognitive psychology, running *t*-tests or analyses of variance (ANOVAs) with data aggregated by subjects (or items) has become obsolete. While issues around averaging obscuring important features of data have been known for a long time (e.g., Estes, 1956; Clark, 1973), statistical practices have changed dramatically since the first decade of this century, notably spearheaded by Baayen et al. (2008). Linear mixed-effects models offer several advantages over aggregated *t*-tests, including the ability to handle data with more complex structures, account for both fixed and random effects, and provide more accurate

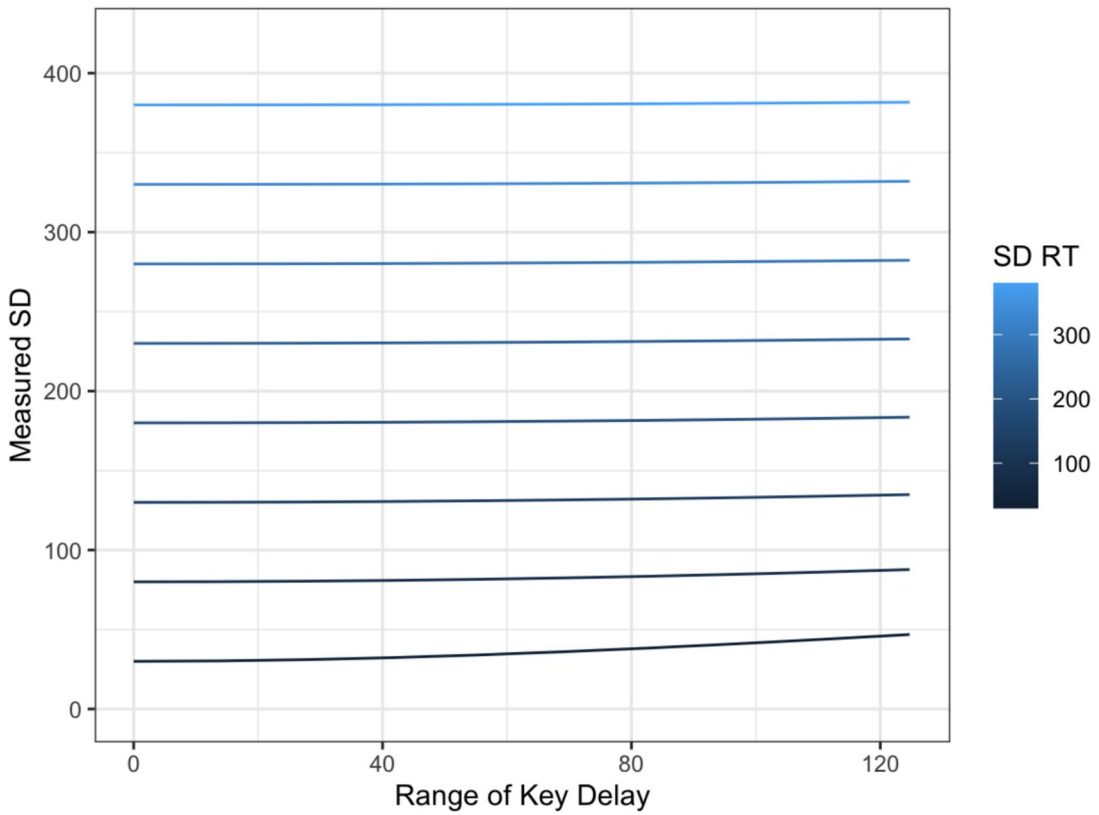

Fig. 1 Total measured standard deviation (SD, *y*-axis) assuming different delays ~U(0, range) in the keyboard (*x*-axis) at varying SD in the RT (colored lines)

estimations that might increase power. Relevant to this paper, it is possible that the averaging that occurs in the aggregated-by-participant *t*-test might attenuate the effect of equipment variability; however, this attenuation might not be present in linear mixed-effects models, as they use all individual observations in the analyses. For this reason, we compare both methods. Note that we use the term *aggregated* to describe analyses that first find the average per condition and per subject, and then perform a statistical test on those averages (i.e., a paired *t*-test); while there can be aggregations across items, for simplicity, here we only use aggregation across participants.

We now present some background on why USB and Bluetooth keyboards may induce extra variability in the obtained RTs. Then we will explore various scenarios to assess the potential impact of using these keyboards in RT experiments. To keep the analysis straightforward, and in line with Ratcliff (1993, 1994), we examined simulated experiments with only two conditions. Similarly, we assumed that the underlying RT distribution corresponds to the convolution of the Gaussian distribution (with parameters μ and σ responsible for the location of the RT distribution) and the exponential distribution (with parameter τ responsible for the skew), i.e., the ex-Gaussian distribution.

USB and Bluetooth keyboards

The delay in keyboards depends on two factors: polling rate and scan rate. The polling rate, a feature of USB connections, refers to how frequently the computer queries the USB device. The scan rate is the frequency of internal checks in the keyboard circuitry to detect key presses and releases. Both rates are measured in hertz (Hz).

High-end gaming keyboards often boast scan rates of 1000 Hz, meaning the computer receives updates from the keyboard every millisecond. Basic commercial keyboards typically have scan rates of 125 Hz, resulting in updates every 8 ms, and basic USB ports poll at 125 Hz as well.

It is easy to find gamers using a high-speed camera, a mechanical key presser, and a high-refresh-rate monitor to measure actual delays on keyboards (e.g., https://www.rtings.com/keyboard/tests/latency). This commentary assumes delays are uniformly distributed, with the effective polling rate approximately twice the reported mean delay. Some tests find the smallest delay to be 0.1 ms with a high-end gaming keyboard, while the largest delay is about 60 ms for a USB connection, so we assume that the measured RT is the sum of the real RT plus the scan and the pooling time, which have a uniform distribution:

$$measuredRT = RealRT + \epsilon_k \sim U\left(0, \frac{1000}{scanrate}\right) + \epsilon_m \sim U\left(0, \frac{1000}{poolingrate}\right)$$

As the extreme case—and using the findings from https://danluu.com/keyboard-latency/, we can assume that $measuredRT = RealRT + \epsilon \sim U(0, 66)$. For reference, the SD of a uniform distribution is defined as $= \sqrt{\frac{range^2}{12}}$, and when a distribution is a convolution of n different components, the total standard deviation is $\sigma_{Total} = \sqrt{\sigma_1^2 + \sigma_1^2 + \cdots + \sigma_n^2}$. This equation becomes particularly relevant for the analyses below, where we generate data using the ex-Gaussian distribution with two components of variance (one from the normal distribution, σ^2, and the other from the exponential distribution, τ^2). We then compare it with data generated with a third component of variance: the measurement error introduced by the equipment.

The simulated experiments

While the above formal analyses can be illuminating, to examine the potential cost induced by an increase in the variability at detecting a response in the keyboard, we generated simulated data from a hypothetical experiment featuring a within-participant factor with two levels (like Ratcliff, 1993); note, however, that the general implications should be the same for more complex designs.

The assumptions were the following:

- Subject was a random factor (20 or 40 participants; the random structure is explained below).
- Item was a random factor (20 or 40 items).
- There was one experimental effect factor (e.g., a manipulation like presentation time [short, long] of a given stimulus).
- The RTs follow the ex-Gaussian distribution[1] using the following parameters:

$$\mu = \mu_{Subject} + \mu_{item} + k \times \mu_{effect}$$
$$\sigma = 10$$
$$\tau = \tau_{Subject} + \tau_{item} + k \times \tau_{effect}$$

[1] Matzke and Wagenmakers (2009) provide an excellent overview of the problems in interpreting ex-Gaussian parameters as direct measurements of cognitive processes. However, in our case, we use it as a tool to generate data without a commitment to the theoretical meaning of such parameters (see Vadillo & Garaizar, 2016, for a similar argument). We use the ex-Gaussian parameters simply as a description of the shape of the RT distribution and the loci of effects.

In our simulations, there is variability across subjects and within subjects. The variability across subjects comes from the fact that each subject has their own μ_S and τ_S, each normally distributed: ($\mu_S \sim N(80,14)$;$\tau_S \sim N(80,14)$). The variability within each subject comes from the variability across items, as each item has its own μ_I and τ_I, which are also distributed normally ($\mu_I \sim N(80,14)$;$\tau_I \sim N(80,14)$), and also from the ex-Gaussian random number generator using the function `rexGAUS()` from the `gamlss.dist` package (Stasinopoulos & Rigby, 2023) in the R environment (R Core Team, 2023). In addition, k is the contrast code for the two conditions of the fixed factor (–.5 and .5).

We carried out two sets of simulations: the first one was pre-planned, and the second one was carried out to address questions that emerged from outcomes of the first set. For the first set of simulations, we explored three scenarios based on plausible loci of effects in the ex-Gaussian distribution:

1. Variability across subjects and items with a null effect of presentation time (i.e., the manipulated factor). In this case, $\mu_{effect} = 0$;$\tau_{effect} = 0$.
2. Variability across subjects and items with an effect of presentation time on the μ parameter of the RT distribution, which is distributed across subjects as $\mu_{effect} \sim N(5, 5)$. This produces a shift in the RT distribution for each participant, equal to the size of the effect (e.g., masked priming experiments show such pattern; see Gomez & Perea, 2020).
3. Variability across subjects and items with an effect of presentation time on the τ parameter of the RT distribution, which is distributed across subjects as $\tau_{effect} \sim N(5, 5)$. This produces a change in the tail (and the variability) of the RT distribution. When the τ parameter changes, effects are larger for the slower responses (e.g., the effect of word frequency in the lexical decision task yields such pattern; e.g., Ratcliff et al., 2004).

Because these simulations aimed to explore an extreme case of noise in computer keyboards, we used a relatively small standard deviation (SD) in the RT distributions, around 125 ms, and a small effect of the fixed factor (presentation time; for each subject ~N[5, 5]). While our focus was on the distribution of t values from the linear mixed-effects (LME) models, for comparison, we also report the distribution using aggregated by-subject t-tests. Finally, we examined the impact of keyboard-induced noise on correlational studies.

Results

We generated 1000 samples for each simulation, and we carried out the LME and the aggregated t-test analyses for each sample using both the raw RT and the –1000/RT transformation—this transformation is a common procedure to meet the normality assumption (see Balota et al., 2013; Lo & Andrews, 2015). We present the distribution of coefficients for the fixed factor (presentation time) in Fig. 2.[2] The outcome of the simulations is summarized in Tables 1, 2, 3 and 4, which include the proportion of t values above the standard critical value of 1.96 for both the LME and aggregated t-tests. For the LME, we implemented the following models:

dv ~ presentation + (1|item) + (1 + presentation|subject))
dv ~ presentation + (1|item) + (1|subject))

Both models produce remarkably similar findings, and the simpler model generates fewer than .05% singularities, while the more complex model yields up to 40% singularities in some of the simulations; hence, the results from the simpler model are presented in the tables. The code and results for the more complex model are available in the OSF site.

Null effect scenario First, the most reassuring aspect of these simulations is that when examining the null effect scenario, the distributions of coefficients overlap, and the proportion of scores above 1.96 is practically identical regardless of the number of trials, number of conditions, or data transformation: all are around the .05 level. In short, alpha is not inflated by the added variability introduced by keyboard use.

Effect only on the tail of the RT distribution (τ parameter) In the case of an effect on τ, which affects the tail of the distribution and its SD, some examination is in order. In the ex-Gaussian distribution, the exponential component of the RT will dominate the SD because $\sqrt{\sigma^2 + \tau^2}$ and in most empirical situations $\tau > \sigma$, which is reflected in our simulations. The results of our simulations show that the results using the actual RT versus the measured RT are practically identical for both LME and aggregated analyses, and no power is lost by using the measured RT instead of the real RT.

Effect only on the location of the RT distribution (μ parameter) In this scenario, there is a small but measurable loss of power in the measured versus the real RT. This loss is most evident when using the –1000/RT transformation in the 40-item, 40-participant simulation, and occurs for both the aggregated-by-subject and the LME analyses (both around a .05 cost). The loss in statistical power decreases for the simulations with fewer trials or fewer items. Thus, for higher

[2] Note that the intercepts show a difference between the measured and the real RT, which is obvious given that the measured RT includes the delay introduced by the uniform noise; the distribution of intercepts can be seen in the OSF repository, but they are trivial and irrelevant to the question at hand.

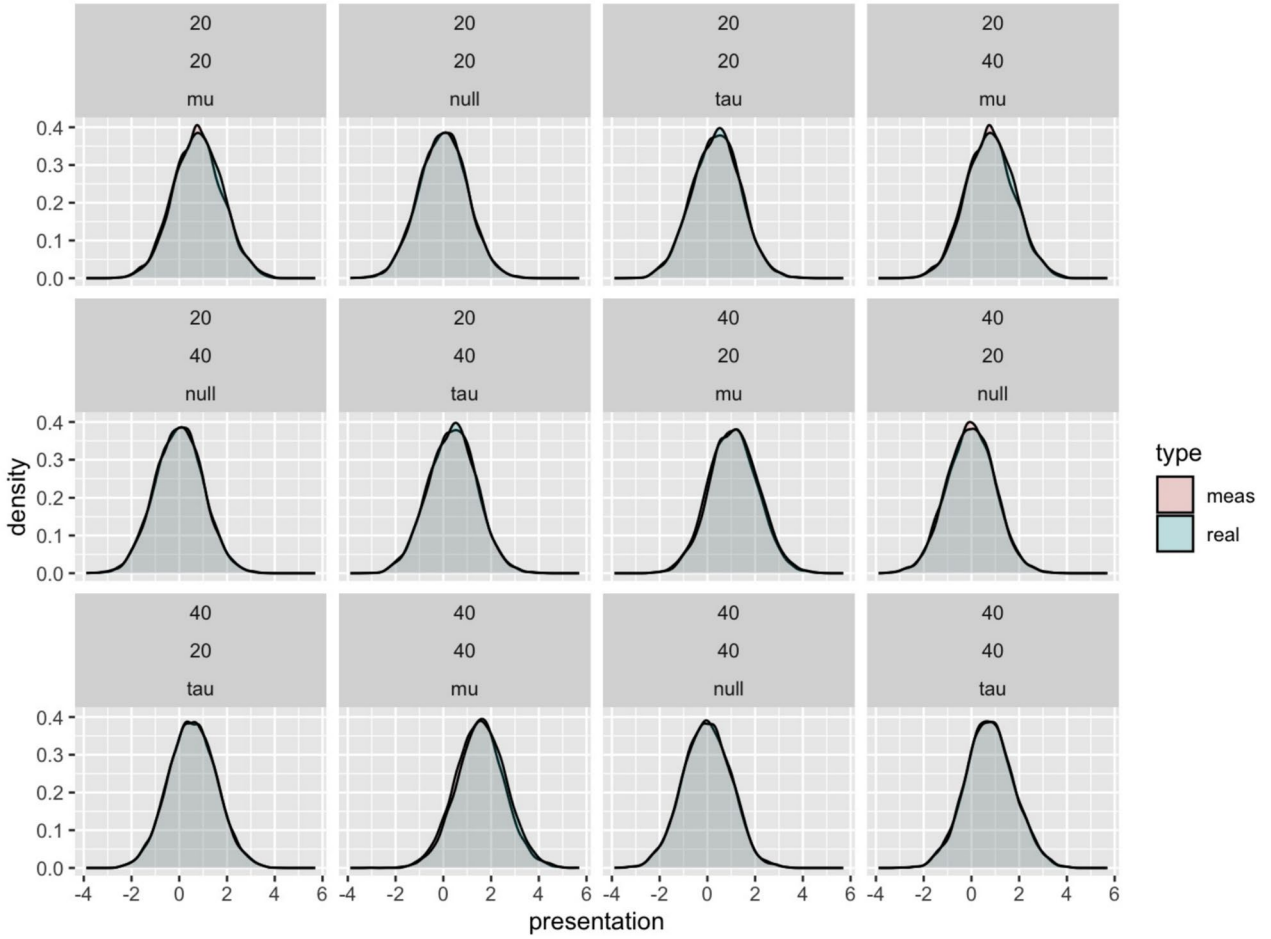

Fig. 2 Distributions of *t* scores for the fixed effect coefficient in the different simulations. Note that only the panel with 40 items and 40 participants, in which the effect is in μ, has a visible difference between the real and the measured –1000/RTs

numbers of items and participants, the cost of the keyboard delay is higher, even if power, overall, is obviously higher as well.

In short, the first set of simulations indicates that there is a loss in statistical power induced by equipment when the effect is located only in the μ parameter of the ex-Gaussian distribution. This led us to a second set of scenarios—each involving 1000 simulations—in which we explored the contours of the loss in statistical power, examining whether such loss is larger with larger effect sizes and whether it remains when there is an effect in τ as well.

1. The effect size on μ increased to ~$N(10, 5)$ across participants.
2. The effect affected both μ and τ, both ~$N(5, 5)$ across participants.
3. A larger variability in the real RTs, with the same effect as in (2), was achieved by increasing the σ parameter to 40 and the τ parameter to 100. The choice of these parameter values is consistent with realistic values in lexical decision tasks (see Matzke & Wagenmakers, 2009, for a variety values of ex-Gaussian parameters in different tasks and conditions).

The results are straightforward. When the RT is not transformed, neither the LME nor the aggregated *t*-tests show a sizable power advantage for the real RT over the measured RT (i.e., the differences are on the order of < .01). The highest cost in statistical power emerges in the case of a larger effect size on μ~ $N(10, 5)$, when using the –1000/RT transformation; in this case, the decrease in power produced by the keyboard delay is about .05. In the other simulations, with effects distributed in τ and μ, the cost in power of using the keyboard decreases to about .02 in the LME analyses with –1000/RT, and it is even smaller in the other analyses.

Naturally, researchers aim to maximize power and address the potential power loss due to keyboard noise. As stated above, this power loss depends on several experimental parameters. We chose to examine the scenario with the highest power loss and estimated how many additional trials

Table 1 Results from the simulations in the *null* scenario. The proportion of statistically significant simulations for each effect type and each data transformation is presented in the measured (*RT + U[0, 66]*) and the real (*RT*) columns

			Proportion of significant simulations		
Data transformation	No. participants	No. items	Measured RT	Real RT	Difference
–1000/RT					
LME	20	20	.055	.053	.002
LME	20	40	.050	.054	–.004
LME	40	20	.050	.044	.006
LME	40	40	.050	.047	.003
Aggregated *t*-test	20	20	.057	.050	.007
Aggregated *t*-test	20	40	.042	.043	–.001
Aggregated *t*-test	40	20	.060	.068	–.008
Aggregated *t*-test	40	40	.054	.056	.002
Raw RT					
LME	20	20	.057	.052	.005
LME	20	40	.045	.048	–.002
LME	40	20	.042	.038	.004
LME	40	40	.048	.049	–.001
Aggregated *t*-test	20	20	.056	.049	.002
Aggregated *t*-test	20	40	.047	.045	.002
Aggregated *t*-test	40	20	.049	.047	.002
Aggregated *t*-test	40	40	.065	.055	.010

Table 2 Results from the simulations in the τ scenario. The proportion of statistically significant simulations for each effect type and each data transformation is presented in the measured (*RT + U[0, 66]*) and the real (*RT*) columns

			Proportion of significant simulations		
Data transformation	No. Participants	No. items	Measured RT	Real RT	Difference
–1000/RT					
LME	20	20	.068	.073	–.005
LME	20	40	.090	.089	.001
LME	40	20	.070	.071	–.001
LME	40	40	.125	.124	.001
Aggregated *t*-test	20	20	.070	.072	–.002
Aggregated *t*-test	20	40	.075	.079	–.004
Aggregated *t*-test	40	20	.076	.075	–.001
Aggregated *t*-test	40	40	.113	.114	–.001
Raw RT					
LME	20	20	.074	.076	–.002
LME	20	40	.093	.095	–.002
LME	40	20	.086	.082	.004
LME	40	40	.146	.146	.000
Aggregated *t*-test	20	20	.067	.068	–.001
Aggregated *t*-test	20	40	.077	.078	–.001
Aggregated *t*-test	40	20	.094	.089	.005
Aggregated *t*-test	40	40	.148	.147	.001

per condition or extra participants would be necessary to counteract the impact of slow keyboards. The results are presented in Figs. 3 and 4, which depict the power curve under the assumptions of Simulation 1 from the second set described earlier: an effect size of $\mu \sim N(10, 5)$ across participants, using the –1000/RT transformation with grouped data.

Table 3 Results from the simulations in the μ scenario. The proportion of statistically significant simulations for each effect type and each data transformation is presented in the measured (*RT + U[0, 66]*) and the real (*RT*) columns

Data transformation	No. participants	No. items	Proportion of significant simulations		
			Measured RT	Real RT	Difference
−1000/RT					
LME	20	20	.136	.146	−.010
LME	20	40	.191	.215	−.025
LME	40	20	.190	.225	−.035
LME	40	40	.322	.367	−.046
Aggregated *t*-test	20	20	.110	.123	−.013
Aggregated *t*-test	20	40	.184	.206	−.022
Aggregated *t*-test	40	20	.175	.191	−.016
Aggregated *t*-test	40	40	.294	.332	−.038
Raw RT					
LME	20	20	.081	.082	−.001
LME	20	40	.101	.102	−.001
LME	40	20	.098	.106	−.008
LME	40	40	.148	.147	.001
Aggregated *t*-test	20	20	.074	.076	−.002
Aggregated *t*-test	20	40	.078	.080	−.002
Aggregated *t*-test	40	20	.096	.093	.003
Aggregated *t*-test	40	40	.105	.129	−.024

Table 4 Second set of simulations. The proportion of statistically significant simulations for each effect type and each data transformation is presented in the measured (*RT + U[0, 6]*) and the real (*RT*) columns

Effect	Measured	Real	Difference
−1000/RT LME			
mu 10	.856	.899	-.043
mu + tau 10	.859	.877	-.018
mu + tau 10 [larger SD in RT]	.856	.879	-.023
Grouped RT. Aggregated T-Test on −1000/RT			
mu 10	.826	.876	-.050
mu + tau 10	.701	.703	-.002
mu + tau 10 [larger SD in RT]	.710	.714	-.004
Raw RT LME			
mu 10	.427	.430	-.003
mu + tau 10	.724	.730	-.006
mu + tau 10 [larger SD in RT]	.747	.753	-.006
Grouped RT. Aggregated T-Test on Raw RT			
mu 10	.399	.409	-.010
mu + tau10	.701	.703	-.002
mu + tau 10 [larger SD in RT]	.704	.707	-.003

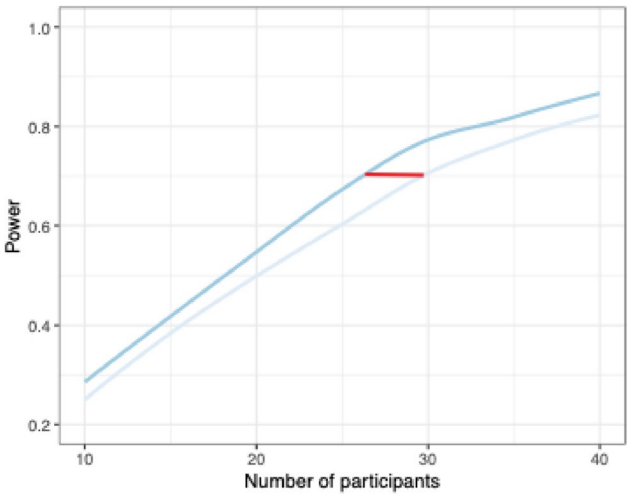

Fig. 3 Power curve illustrating the number of additional participants needed to compensate for keyboard delay, based on Simulation 1 with an effect size of μ ~*N*(10, 5). The darker line represents power with no keyboard delay, and the lighter line represents power with keyboard delay. The arrow indicates the horizontal offset at the 0.7 power level, corresponding to four additional participants

The figures are straightforward to interpret: the darker lines represent power without keyboard delay, while the lighter lines show power with keyboard delay. To estimate the number of additional participants or items needed to achieve a certain power level, one can measure the horizontal offset between the lines. For example, the horizontal offset at the 0.7 power level, highlighted with arrows in both figures, indicates that four additional participants are needed, as shown in Fig. 3. Similarly, in Fig. 4, the arrow

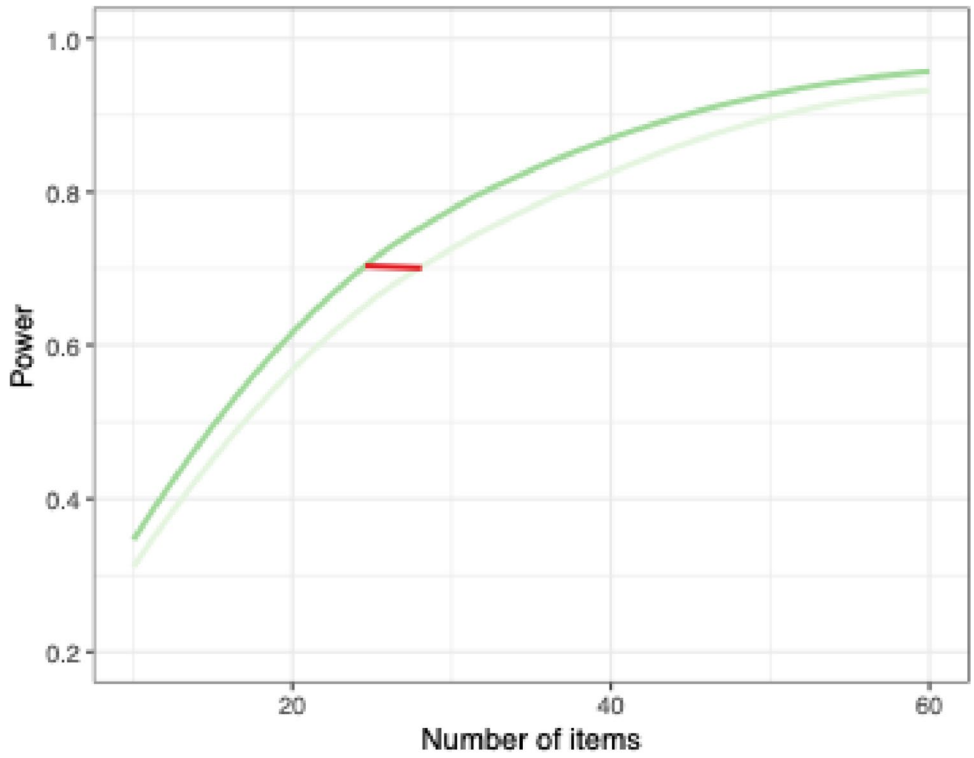

Fig. 4 Power curve illustrating the number of additional items needed to compensate for keyboard delay, based on Simulation 1 with an effect size of μ ~$N(10, 5)$. The darker line represents power with no keyboard delay, and the lighter line represents power with keyboard delay. The arrow indicates the horizontal offset at the 0.7 power level, corresponding to two additional items

suggests that two extra items compensate for the keyboard noise. Notably, the offset between the lines never exceeds five units. These results suggest that adding approximately five items or five participants per condition can effectively compensate for the power loss caused by slow keyboard latency.

The case of individual differences In the simulations described above, we focused on the effects of keyboard noise in experiments that examined group-level effects. To investigate this issue in the context of individual differences research, we conducted a third simulation**.** This time, we introduced even larger response-induced noise, distributed as ~ uniform(0,100), and used the following premise: there is a cognitive task for which the RT has a .35 correlation with IQ $\sim N\left(100, 15^2\right)$. In this case, it is useful to think of each participant *j* having a true mean μ_j, which is distributed $\sim N\left(\mathrm{M}, \eta^2\right)$ where *M* is the overall mean and η^2 is the variance across participants.

The estimation of *M* is aided by increasing the number of participants or the number of trials. However, to locate μ_j (the mean for individual participants), adding additional participants does not help; only adding trials matters. Consequently, we carried out simulations to explore the effects of adding trials.

When estimating the correlation between a cognitive task's RT and IQ, or even between two cognitive tasks, the critical component is locating μ_j, and to do so we use the measured average latency, which is determined by the across-trial variability and the number of trials: $SE_{\mu j} = \frac{\sigma}{\sqrt{n}}$, where σ is the across-trial variability and *n* is the number of trials. The precision of locating μ_j is very important if there is small variability across participants. However, if there is large variability across participants, one can tolerate less precise location of μ_j. In other words, the ratio of across participant variability and across-trial variability is very relevant. For this reason, we explored those two forms of variability in our simulations (for a more complete treatment of the issues around variability between and within participants, see Hedge et al., 2018; Rouder et al., 2023).

The simulations are summarized in Fig. 5, where variability across participants (η) is displayed from left to right, and variability across trials is displayed from top to bottom. Importantly, when the variability across participants is small, the added variability across trials caused by the keyboard delay attenuates the correlations in a sizable manner; when the variability across participants is small, any increase in the cross-trial variability incurs a cost. The good news is that the attenuation of the correlations diminishes

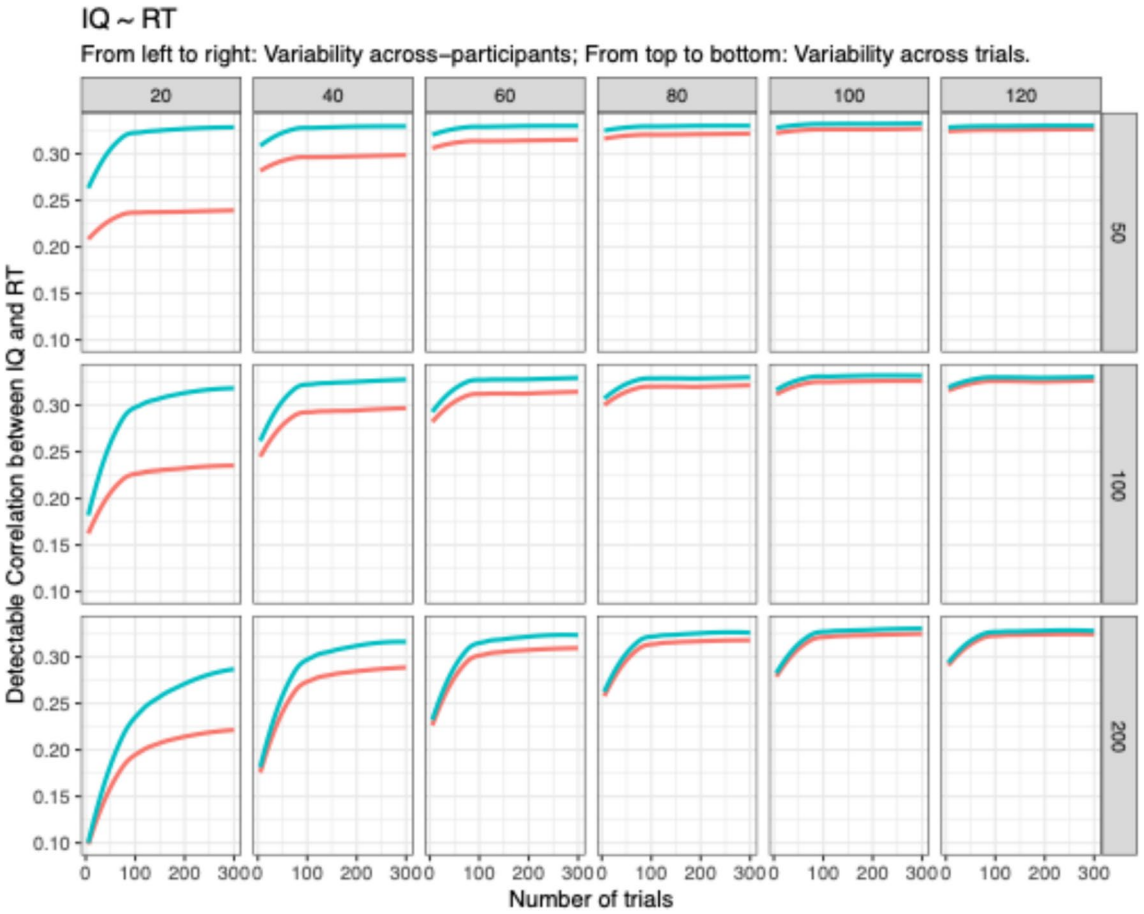

Fig. 5 Simulation results exploring the effects of noise on individual differences research, focusing on the correlation between response time (RT) and IQ (μ ~*N*(mean = 100, SD = 15)). The figure displays different levels of variability across participants (*SD* of the RT across participants) from left to right and variability across trials from top to bottom. The added variability due to keyboard delay attenuates correlations, especially when across-participant variability is small. However, this attenuation diminishes when the SD across participants is > 80. The figure underscores the importance of trial variability in estimating individual participant means (μ_j) and the impact of across-participant variability on correlation attenuation

when the variability across participants exceeds $\eta > 80$ (which is not uncommon in tasks like lexical decision; see, for example, Mella et al., 2015; Zoccolotti, et al., 2018). Importantly, when the individual variation that is submitted to the correlation comes from subtractive conditions (e.g., incongruent minus congruent trials), the effect will be small, and η will also be small.

Discussion

Increasingly, RT experiments are being conducted online, raising a recurring concern about whether the (uncontrolled) added noise from participants' keyboards has a sizable detrimental effect on statistical power. In previous papers, Ratcliff (1994) and Damian (2010) argued that the potential cost in experiments with aggregated data is minimal. However, a remaining issue is whether the scenario would differ in the now prevalent analyses via linear mixed-effects (LME) models, which rely on individual RT data rather than data aggregated by subjects or items.

The findings from the presented simulations are clear and suggest that using mass-produced keyboards in RT experiments generally does not have a detrimental effect in statistical power, as shown by inferential data analyses with both linear mixed-effects models and aggregated by-subject *t*-tests. However, some caveats must be addressed. In the LME analyses, the distributions of coefficients were similar whether RTs were measured with or without added noise. Nevertheless, noticeable but small differences emerged in specific situations, which are discussed below.

First, the detrimental consequences of keyboard delays can be sizeable for scenarios with very little variability in RTs, as shown in the bottom-most line in Fig. 1. Perhaps

most importantly, if the effects of the manipulation are expressed as a shift (only) in the RT distributions (i.e., an effect on μ), there is a small but measurable cost in power of the LME models and in the aggregated *t*-test when using the –1000/RT transformation. To contextualize the minor loss of power when using mass-produced keyboards, our simulations suggest that in the worst-case scenario with a 0.05 reduction in power, adding just five participants or increasing the number of trials by approximately 5–10 would restore the original power level. These adjustments represent a feasible solution for researchers concerned about hardware variability in their RT experiments.

In addition, for correlational or individual variability research, researchers should consider the signal-to-noise ratio (variability across participants relative to variability across trials); lower ratios indicate greater sensitivity to keyboard-induced noise. This can be achieved relatively simply by computing the SD across participants and across trials and comparing those numbers to Fig. 5. Conversely, researchers can use methods employing hierarchical trial-level models suggested by Rouder et al. (2023) to address this issue.

While the scenarios that yield this loss in statistical power are plausible, they occur in our simulations designed to maximize the chances of obtaining a difference between the real and the measured RT—specifically, we used the simulations with the longest possible pooling and scan times, and relatively small effect sizes in the RT. We must bear in mind that in most cases, the power loss would be smaller than the ones in the extreme scenario reported here. Furthermore, manipulations that only shift the RT distribution (i.e., an effect on μ, which corresponds to a shift in the RT distributions) are highly infrequent (e.g., masked priming is one of the few exceptions; e.g., see Gomez & Perea, 2020). Critically, once we simulated the more common scenario of an effect affecting both the location and tail of the distributions (i.e., μ and τ parameters), the cost in statistical power was reduced by half. Second, our findings apply to behavioral experiments. In other scenarios, such as response-locked event-related potential (ERP) analyses, it is advisable to use an extremely precise measure for RTs via response boxes in the laboratory because the averaging of brainwaves is likely to be sensitive to the jitter introduced by equipment; in fact, most electroencephalography (EEG)/ERP equipment includes high-quality response boxes.

Given the increasing prevalence of online experiments in psychology and other fields (Rodd, 2024), revisiting the issue of equipment-based noise in RT experiments is particularly timely. In the last few decades, statistical practices have shifted significantly moving from analyses of aggregated RT data to those based on individual RT data, which may be potentially more sensitive to measurement noise. Related to the present work, Vadillo and Garaizar (2016) analyzed the ability to recover ex-Gaussian parameters from functional fits and parameters of the diffusion model when noise was introduced by imprecise measuring of RTs. They found that the recovery of generating parameters was robust to this added noise. In this work, we have quantified the reduction in power because of added noise, in both linear mixed-effects models and aggregated *t*-test analyses, and while one could focus on the situations where the loss in power is sizable (when the effect involves a shift in the RT distributions or when it occurs in studies of individual differences with low signal-to-noise ratios), in most situations, the effect of mass-produced equipment is only modestly detrimental.

To conclude, while there is some gain for precise measurement of RTs, access to such hardware is not universally feasible in online experiments (see Pronk et al, 2020, for an in-depth discussion of the different versions of online experiments and their implications). Importantly, commercially available keyboards remain suitable for most RT studies, provided researchers implement simple adjustments, such as increasing the number of trials or participants, to mitigate potential variability. On the OSF, we have included a script that can easily be modified to estimate, under different assumptions, plausible losses in power due to equipment variability. This practical approach allows for robust and reliable RT data collection, even in the era of widespread online experimentation.

Author contributions All authors contributed to the study's conception and design. The first draft of the manuscript was written by P.G. and M.P. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The research reported in this article has been partially supported by Grants PID2020-116740 GB-I00 (funded by the MCIN/AEI/1013039/501100011033) and PID2023-152078NB-100 from the Spanish Ministry of Science, Innovation, and Universities to Manuel Perea, Grant CIAICO/2021/172 from the Department of Innovation, Universities, Science and Digital Society of the Valencian Government to Manuel Perea, and NSF grant SMA-2127135 to Pablo Gomez.

Data availability The stimuli, data, scripts, and outputs are available at https://osf.io/4z8rn/.

Code availability The stimuli, data, scripts, and outputs are available at https://osf.io/4z8rn/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Not applicable.

No human data was collected, nor used.

Consent to participate Not applicable.

No human data was collected, nor used.

Consent for publication All authors consent publication.

References

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*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baciero, A., Uribe, I., & Gomez, P. (2021). The tactile Eriksen flanker effect: A time course analysis. *Attention, Perception, & Psychophysics, 83*, 1424–1434. https://doi.org/10.3758/s13414-020-02172-2

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: The influence of trial history and data transformations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1563–1571. https://doi.org/10.1037/a003218

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

Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards? *Behavior Research Methods, 42*, 205–211. https://doi.org/10.3758/BRM.42.1.205

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*, 134–140. https://doi.org/10.1037/h0045156

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*, 116–124. https://doi.org/10.3758/BF03195503

Gomez, P., & Perea, M. (2020). Masked identity priming reflects an encoding advantage in developing readers. *Journal of Experimental Child Psychology, 199*, 104911. https://doi.org/10.1016/j.jecp.2020.104911

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hollingworth, A., Richard, A. M., & Luck, S. J. (2008). Understanding the function of visual short-term memory: Transsaccadic memory, object correspondence, and gaze correction. *Journal of Experimental Psychology: General, 137*, 163–181. https://doi.org/10.1037/0096-3445.137.1.163

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Matzke, D., & Wagenmakers, E. J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review, 16*, 798–817. https://doi.org/10.3758/PBR.16.5.798

Mella, N., Fagot, D., & de Ribaupierre, A. (2015). Dispersion in cognitive functioning: Age differences over the lifespan. *Journal of Clinical and Experimental Neuropsychology, 38*, 111–126. https://doi.org/10.1080/13803395.2015.1089979

Pronk, T., Wiers, R. W., Molenkamp, B., & Murre, J. (2020). Mental chronometry in the pocket? Timing accuracy of web applications on touchscreen and keyboard devices. *Behavioral Research Methods, 52*, 1371–1382. https://doi.org/10.3758/s13428-019-01321-2

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Ratcliff, R. (1994). Using computers in empirical and theoretical work in cognitive psychology. *Behavior Research Methods, Instruments, & Computers, 26*, 94–106. https://doi.org/10.3758/BF03204600

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review, 111*(1), 159–182. https://doi.org/10.1037/0033-295x.111.1.159

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language, 134*, 104472. https://doi.org/10.1016/j.jml.2023.104472

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review, 30*(6), 2049–2066. https://doi.org/10.3758/s13423-023-02293-3

Stasinopoulos, M., & Rigby, R. (2023). *gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape*. R package version 6.1–3. https://github.com/gamlss-dev/gamlss.dist. Accessed 13 March 2025.

Vadillo, M. A., & Garaizar, P. (2016). *The effect of noise-induced variance on parameter recovery from reaction times*. BMC Bioinformatics, 17(1). https://doi.org/10.1186/s12859-016-0993-x

Voss, A., Leonhart, R., & Stahl, C. (2007). How to make your own response boxes: A step-by-step guide for the construction of reliable and inexpensive parallel-port response pads from computer mice. *Behavior Research Methods, 39*, 797–801. https://doi.org/10.3758/BF03192971

Zoccolotti, P., De Luca, M., Di Filippo, G., Marinelli, C. V., & Spinelli, D. (2018). Reading and lexical-decision tasks generate different patterns of individual variability as a function of condition difficulty. *Psychonomic Bulletin & Review, 25*, 1161–1169. https://doi.org/10.3758/s13423-017-1335-3

Open practices statement The code for the simulations and the analyses is available at https://osf.io/4z8rn/; https://doi.org/10.17605/OSF.IO/4Z8RN.