Does variability in human performance outweigh imprecision in response devices such as computer keyboards?

MARKUS F. DAMIAN

University of Bristol, Bristol, England

In chronometric research, it is well-known that many response devices used with personal computers, especially computer keyboards, suffer from measurement inaccuracies due to infrequent polling. In this article, it is investigated whether it is worth being concerned by this added error, given that human performance inherently exhibits a considerable degree of variability. By comparing inaccuracy due to response device imprecision with the level of variability already present in human performance, it is shown that the former is very unlikely to have a negative impact on statistical outcomes. In other words, millisecond resolution of response devices, although desirable, is not a general requirement.

As scientists pursuing a quantitative approach, we take pride in the rigor of our experimental designs and the accuracy of our measurements. Because cognitive psychologists interested in the time course of mental processes typically look at events on the order of milliseconds, it would appear self-evident that the tools with which we collect our data likewise must have a similar (i.e., millisecond) accuracy. This—typically, implicit—assumption is occasionally made explicit; for instance, in one of the standard texts on human response times, Luce (1986) stated the following: “Without precision of measurement, one cannot possibly answer some of the detailed questions raised by theory . . . it is absolutely essential that [techniques of measurement] be of sufficient quality to get the millisecond accuracy found in most of the data discussed” (p. 3). However, an even more fundamental question is not typically asked—namely, whether such measurement precision is really required, given the inherent variability of human performance.¹

The advent of the personal computer has put an accessible and inexpensive tool for the design and implementation of chronometric research within easy reach. A number of off-the-shelf software packages for this purpose are now readily available (e.g., DMDX, SuperLab Pro, E-Prime, Psychlab, and others), and most of them pride themselves on their ability to measure responses with millisecond resolution. But of course, measurements collected by computers are only as accurate as the input devices that are used, and here many of them fall drastically short of achieving millisecond precision. For experimental tasks requiring manual responses, the most obvious choice of response device is the computer keyboard, and using keyboards to collect response latencies remains in widespread use in

published work. However, computer keyboards were not designed for millisecond accuracy, mainly because their typical use—text and numerical inputting—does not require such a degree of precision. One of the main sources of loss of precision is the fact that their status is only relatively infrequently sampled by the CPU. Most modern keyboards are connected to computers via the USB port (e.g., Anderson, 1997), and unless specifically tuned (typically, for gaming purposes), this entails that the keyboard is polled by the computer with a resolution of 125 Hz, or every 8 msec. Unless special measures are taken to increase the polling rate beyond its default setting, this limitation also extends to other response devices that are connected via USB, such as joysticks, game pads, and so forth. The more traditional way of connecting keyboards to computers via PS/2 connectors implies typical polling rates of 60 Hz, or every 17 msec (e.g., Jordan & Churchill, 1990). There are additional sources of error inherent in standard computer keyboards that could potentially increase variability: The keyboard microprocessor may scan successive keys at a certain rate in an order that typically varies from product to product and, hence, is unpredictable; the transmission time between keyboard and computer may take some time; and a motherboard keyboard chip may require additional processing time. The operating system also typically has its own polling rate at which it interacts with peripherals. Furthermore, there is no industry standard to which such devices must conform. J. C. Forster (2007) provided a systematic assessment of several response devices for chronometric research by benchmarking their accuracy against an automatically generated electronic switch closure and concluded that “the keyboard should never be used as an input device.”

M. F. Damian, m.damian@bristol.ac.uk

Intuitively, a worse-than-millisecond resolution of a response device seems highly unsatisfactory, given that the aim is to measure the duration of mental processes on the scale of milliseconds. And indeed, the author has repeatedly received reviews of manuscripts submitted to scientific journals in which expert referees have criticized the use of computer keyboards as inappropriate due to their inherent imprecision. To be sure, there are response devices available that provide the desired millisecond accuracy—most notably, via the use of tailor-made response boxes connected to dedicated digital I/O cards or to the serial port. The central issue assessed below is whether, from a pragmatically driven point of view, this is worth the effort, given the typical variability already present in human responses. To answer this question, it is worth noting that to overcome the variability of human performance, standard practice is to accumulate a substantial number of observations per condition, which are then averaged to arrive at a valid estimate. Most psychologists have developed “rules of thumb” about how many data points are necessary in order to interpret results. For typical chronometric experiments, it is common to average across several hundred observations, and only then can the means provide an adequate impression. The question is whether, against this background of noise, it is necessary to require millisecond resolution from response devices.

In what follows, I will focus on the effects of infrequent polling of the computer keyboard. Infrequent polling of the response device entails that the measured latencies “lag” behind the “true” latencies by a certain amount. Using a device with limited resolution R , such as a keyboard, adds a certain amount of error variance to each “true” response time—that is, a value between 0 and R msec that is uniformly distributed within that range. This “lag” has a mean of $R/2$, and the corresponding constant value will be added to conditional means, but this is of little concern, because it applies to all experimental conditions alike. More interesting is the variability of the added error. The standard deviation (SD) of a uniform distribution ranging from x to y is given by

$$SD = \sqrt{\frac{(y - x)^2}{12}}. \quad (1)$$

For a resolution of 17 msec, such as that found in computer keyboards connected via PS/2, the SD of the error associated with individual responses is 4.91. The important question is whether such a degree of additional variability could conceivably make a difference when chronometric experiments are conducted.

The answer crucially depends on the magnitude of the variability introduced by response device imprecision, relative to the variability inherently present in human performance. One of the difficulties in developing an intuition for questions of this type lies in the fact that standard practice in cognitive psychology is to average, for each participant, the response times within a condition and then enter the resulting conditional means into an inferential statistical procedure, such as an ANOVA. As a consequence, the SD s reported in Results sections overwhelm-

ingly report variability among participants’ conditional means, rather than variability of raw data. Raw data are oftentimes not inspected, and this approach is encouraged by auxiliary software tools (e.g., the ANALYZE utility provided with DMDX; K. I. Forster & Forster, 2003) that allow for an automatic transformation of raw data into conditional means.

In what is outlined below, the general idea is as follows. It is assumed that a conditional mean provides an imperfect estimate of a “true” underlying mean. The first step is to characterize the variables that render this estimate less than perfect. The second step is to assess the additional impact that error variability due to response measurement inaccuracy can have, relative to that already present in the data. Inferential statistical tests are typically performed on conditional means, and with regard to the additional variability introduced by response device inaccuracies, one could argue that, if the computed means deviate only very slightly, or not at all, from the “true” estimates, the results of the inferential procedures will necessarily be very similar or identical. In such a case, there is no reason to be alarmed by the added degree of error.

To this aim, the perturbation of estimates due to various error sources will be conceptualized as the *mean absolute deviation* (MAD) from a “true” mean—that is, the average degree to which an estimate tends to deviate from the underlying mean (Howell, 2008; Kenney & Keeping, 1962). The choice of MAD as a measure of dispersion, over the ones more conventionally used in psychology (such as standard deviations) is motivated by the fact that it transparently relates to the main objective: Given a certain level of noise and a particular number of data points, how many milliseconds (if any) are our measured estimates, on average, bound to be different from the “true” mean? The MAD of a set $\{x_1, x_2, \dots, x_n\}$ is calculated as

$$MAD = \frac{1}{n} \sum_{i=1}^n [x_i - m(X)]. \quad (2)$$

In what follows below, however, we are not interested in the MAD among a set of raw data but, rather, in the average deviation of multiple computed means from their “true” mean.

Estimating the Variability of Human Performance

In judging whether it is worth being alarmed by the inaccuracies of measures taken by digital devices, the first step consists of forming an impression of the noise level within raw response times. To this aim, 5 participants were asked to carry out a number of standard, widely used, experimental reaction time tasks. The simple aim was to provide an impression of the level of noise in tasks of this type. The tasks were a simple reaction time task (e.g., Donders, 1868/1969), a choice reaction time task (e.g., Donders, 1868/1969), a Stroop task with manual responses (e.g., Keele, 1972), an Eriksen flanker task (e.g., Eriksen & Schultz, 1979; Sanders & Lamers, 2002), and a lexical decision task (e.g., Meyer & Schvaneveldt, 1971).

Each task consisted of 64 experimental trials and 8 practice trials. Each of the 5 participants carried out all five tasks, and the order of tasks was counterbalanced. All the stimuli were centrally presented in Arial 14-point font on a computer monitor, on a light gray background. All the stimuli were presented in black, except in the Stroop task (see below). A new random sequence of trials was generated for each participant. The specific manipulations were as follows.

Simple reaction task. The letter X was presented 1,500, 2,000, 2,500, or 3,000 msec after trial onset, in the center of the screen. The participants pressed a response key as soon as the target was detected.

Choice reaction task. A fixation dot was presented for 500 msec, followed by either the letter X or O. The participants pressed the right key if X was the target and the left key if O was the target.

Manual Stroop task. A fixation dot was presented for 500 msec, followed by a word in either red or green ink. The participants pressed the right key if the target was presented in green and the left key if it was presented in red. Words consisted of the four color words *green*, *red*, *blue*, and *yellow*, which were presented eight times each.

Flanker task. A fixation dot was presented for 500 msec, followed by a target in which a target letter was flanked by distractor letters, such as “K K K H K K K.” The participants pressed the right key if the central target letter was either H or K and the left key if the target letter was either S or C. Each of the eight possible combinations of target and flanker was presented four times.

Lexical decision task. A fixation dot was presented for 500 msec, followed by either a word or a pseudoword. Pseudowords were chosen to be orthographically and phonologically plausible. The participants pressed the right key if the target was a word and the left key if the target was a pseudoword.

The results are shown in Table 1. The observed *SDs* of raw latencies are in the range of 30–160 msec, with substantial variability dependent on participant and task.

What Variables Constrain the Accuracy of a Conditional Mean Estimate?

The answer to the question of what variables constrain the accuracy of a conditional mean estimate is straightforward. When response latencies are averaged to estimate a conditional mean, the two variables affecting

the accuracy of a mean estimate are (1) the variability of raw responses and (2) the number of data points per condition. Large variability of individual responses, in association with a low number of measurements per condition, is most prone to delivering a mean estimate that differs from the “true” mean. To systematically explore the effects of these two variables, a number of simulations were run in which response times were generated from a Gaussian distribution with a mean of 500 and an *SD* range typical of human performance, based on the results from the experiments reported above (20–140 msec). It is, of course, well-known that human response times are often not normally distributed but, typically, have a rather long right tail and, hence, are frequently modeled with an ex-Gaussian distribution consisting of a convolution of a Gaussian and an exponential function (e.g., McGill, 1963; Ratcliff, 1979; see Lacouture & Cousineau, 2008, for an overview and tutorial). However, for the matter investigated below, additional simulations that varied the contribution of the exponential portion indicated that it is irrelevant whether data are generated from Gaussian or ex-Gaussian distributions, which result in numerically identical MADs.

Individual response times were averaged across numbers of data points per condition, ranging from 5 to 50; this should provide a range of data points per cell that approximates the one typically found in empirical studies. For each combination of *SD* and number of data points per cell, 100,000 simulations were run, and the MAD of the estimated mean from the “true” mean of 500 msec was computed.² The results are shown in Figure 1.

Not surprisingly, the most accurate measurement is obtained with many data points per cell and low variability of the task, and the most imprecise estimate is obtained with only a few observations per cell and high variability. In the most favorable case (50 trials per condition and an *SD* of only 20 msec), an observed mean can be expected to be within ~2 msec from the “true” conditional mean. On the other extreme, with only five data points per cell and high variability (*SD* = 140 msec), the estimates deviate, on average, approximately 50 msec from the “true” conditional mean.

Having available an impression of the degree of variability in human performance and of the effects of having a less-than-optimal number of observations per cell, we can now progress to the central question of whether added

Table 1
Results: Mean Latencies (in Milliseconds, With Standard Deviations)
and Percentages of Errors (PEs) From Five Experimental Tasks, Separately for Each of 5 Participants

Participant	Task														
	Simple Reaction			Choice Reaction			Manual Stroop			Flanker			Lexical Decision		
	<i>M</i>	<i>SD</i>		<i>M</i>	<i>SD</i>	PE	<i>M</i>	<i>SD</i>	PE	<i>M</i>	<i>SD</i>	PE	<i>M</i>	<i>SD</i>	PE
1	324.8	51.2		445.3	36.4	0.0	518.0	59.6	6.3	526.7	49.2	1.6	517.8	44.4	3.1
2	280.5	41.8		398.6	85.4	0.0	387.4	64.8	1.6	492.9	125.0	3.1	574.5	126.1	6.3
3	324.0	38.2		456.2	72.1	0.0	465.5	68.0	0.0	547.5	78.7	1.6	696.3	113.5	4.7
4	323.4	65.4		410.3	86.0	1.6	564.4	163.4	1.6	590.9	124.8	1.6	760.3	131.4	3.1
5	379.4	103.1		442.3	46.3	0.0	612.8	56.3	3.1	541.8	44.2	0.0	558.8	44.2	6.3
<i>Average</i>	326.4	59.9		430.5	65.2	0.3	509.6	82.4	2.5	540.0	84.4	1.6	621.5	91.9	4.7

Figure 1. Mean absolute deviation from “true” mean (in milliseconds), across 100,000 simulations, when latencies with particular standard deviations (SDs) are averaged across a variable number of trials per condition.

variability due to response input device inaccuracies can adversely affect results.

To What Extent Does Measurement Error Associated With Response Devices Affect Conditional Means?

The variability added by an imprecise measurement device such as a computer keyboard adds a further source of variability to already noisy data that is unrelated to other sources of errors. In this section, the amount of error added by this imprecision will be assessed. As was outlined in the introduction, the logic is as follows. Consider two scenarios, one in which responses were collected with an instrument with true millisecond precision, and another one in which the same responses were measured with a device of coarser resolution. It appeals to common sense that, if averaging across several data points to arrive at a conditional mean (as is standard practice) resulted in an outcome that was identical (or very similar) under both scenarios, the measurement error added by the input device could be safely neglected. The question then is how large the additional measurement error caused by an imprecise input device would have to be in order to cause a conceivable distortion of a conditional mean, and on the other hand, what other factors would affect the outcome.

Because measurement error due to a response input device is unrelated and, therefore, statistically additive to other sources of variability, these can be safely ignored; the possible deviation from a “true” estimate added by the device imprecision is statistically independent of the possible deviation due to other sources. The one remaining factor is the number of data points per cell. To assess the relationship between a particular response device resolution and the number of data points per cell, I

conducted simulations in which response latencies were generated with an SD of 50 msec (to reiterate, other SDs render identical results because the two sources of error variability are unrelated), and to each data point was then added a random measurement error ranging from 0 to R msec, R being the resolution in milliseconds, ranging from 18 down to 3 msec. I assessed a range of data points per cell from 5 to 50. For each step, 100,000 simulations were run, and the deviation from the unperturbed mean was then averaged. The results are shown in Figure 2. Note that, as was stated above, the response inaccuracy due to “lag” of the response device polling, when averaged, adds a constant value with a mean of $R/2$ to the estimates. Because, as long as the same response device is used for all the experimental conditions, this lag is added to all the conditional means, it is of little concern. For the sake of exposition, this constant has been subtracted in order to arrive at the means displayed in the figure.

Comparison of Figures 1 and 2 clearly reveals that the amount of deviation from the true mean due to response measurement inaccuracies is very minor; in fact, combining the two figures would render the effects of response inaccuracies invisible to the naked eye altogether. Even in the most unfavorable case of employing a response device with a rather poor 18-msec resolution (similar to a PS/2 keyboard) and only five data points per cell, the average absolute deviation from the “true” mean caused by response device imprecision is less than 2 msec. By comparison, with such few data points per cell, the average absolute deviation due to inherent human response variability ranges between 7 and 50 msec, dependent on the SD of the specific task (see Figure 1). For practically

Figure 2. Mean absolute deviation from “true” mean (in milliseconds), across 100,000 simulations, as a result of added variability due to infrequent response device polling, averaged across a variable number of trials per condition.

all other cases, the additional variability is more in the range of 1 msec, and in fact, in many cases, the average deviation, when rounded to the nearest integer, is 0 msec. In other words, even with poor resolution and relatively few trials per cell, the measured mean converges on the “true” mean within 1 msec or less. Given that conditional means from a sizable number of participants will eventually be entered into an inferential statistical procedure, the outcome of that analysis is likely to be extremely similar—and, in many cases, identical. And of course, the descriptive latency means reported in a scientific article represent average values from multiple participants, making it extremely unlikely that the already very small deviations per participant due to response device imprecision have a tangible effect on the reported values.

Conclusions

It goes without saying that, having a choice, scientists should prefer more accurate measurement tools over less accurate ones; establishing precision in measurement is generally advantageous, and loss of precision should be avoided if possible. Furthermore, it is generally desirable to quantify the degree of error inherent in a particular method. However, it is worth keeping in mind that a given measurement method needs only to offer sufficient, rather than optimal, precision for a particular task at hand. In this article, I pursued a pragmatically motivated issue: Given the degree of variability inherent in human performance, is it worth being concerned about whether the devices used to collect manual response latencies have suboptimal (i.e., less-than-millisecond) accuracy? The answer is that the degree of imprecision by inaccurate response devices added to already substantially noisy human performance data is vanishingly small and very unlikely to make a difference in the statistical outcomes. Specifically with regard to computer keyboards as input devices, there may be other valid reasons to prefer alternative response devices, relating to, for example, ergonomics; however, measurement accuracy does not warrant serious concern. Hence, the concerns expressed by J. C. Forster (2007) and others against the use of keyboards in cognitive studies seem unwarranted.

It is worth noting that, apart from the error variability introduced by infrequent polling of the keyboard by the computer, there are additional potential sources of error variability inherent to keyboards (due to the keyboard microprocessor, transmission time between the keyboard and computer, the motherboard keyboard chip, etc.). But as for the error variability associated with infrequent polling of the keyboard, the argument is that the loss of statistical power is largely negligible in the context of the inherent variability observed in human performance.³

The measurement inaccuracies associated with standard computer keyboards are well-known but are considered largely irrelevant in typical usage situations, such as data or text entry. However, it is certainly the case that for the purpose of running behavioral experiments, commercial solutions are available that either minimize the error variability (for instance, by using keyboards specially designed and optimized for gaming, or by avoiding key-

boards as response devices altogether and using dedicated response boxes instead) or at least allow a precise assessment of the introduced variability (e.g., the Black Box Toolkit described in Plant, Hammond, & Turner, 2004). Given these possibilities, it is worth asking whether there are experimental situations in which taking this approach would be imperative, and not just principally desirable.

As was outlined above, an imprecise response device has two effects: Relative to the “true” latency, it (1) increases error variability and (2) introduces a “lag.” To this point, the added error variability of imprecise response devices has been exclusively characterized in terms of potentially introducing a Type II error—that is, a false negative. It is in principle conceivable that an experimenter will fail to replicate a significant result, which was originally obtained with a precise response device, with a less precise device. But if the argument advanced in this article holds, the additional error introduced by the imprecise response device is bound to be small relative to human variability, and only in experiments that lack statistical power in the first place will the added variability potentially have an effect. On the other hand, if statistical power is sufficient, the added error will likely be negligible. Similarly, it is conceivable that the added error variability could result in a Type I error—that is, in a false positive—but again, this will happen only in studies that lack statistical power to begin with.

The second statistical consequence associated with imprecise computer keyboards arises from the “lag,” which manifests itself as a constant value added to computed means. Typically, this lag can be ignored, because it merely adds a constant value to all the experimental conditions alike but does not affect relative differences. However, a false positive or negative result could conceivably be produced in some restricted circumstances, such as when different response devices are used for experimental conditions (e.g., when a participant generates responses with two hands, each on a different response device afflicted with a different degree of “lag,” or when latencies are compared between two groups of participants, each group tested with a different response device). However, these are rather contrived situations, and in more typical experimental tasks, it is again difficult to see how a constant value added to latency means by an imprecise response device could affect the outcome of a study.

It should be noted that the inferences drawn from the present study are not particularly novel. Ulrich and Giray (1989) estimated the effects of poor resolution of the internal clock of a computer on the likelihood of rejecting the null hypothesis in response time measurements and concluded that the effect was negligible (note that this specific problem is no longer applicable to modern CPUs). Similarly, Lincoln and Lane (1980) investigated the error associated with the use of CRT displays. As with the limited resolution of most response input devices, an image presented on such a monitor is created by an electron beam sweeping across successive horizontal lines of the display with a particular resolution (typically, between 60 and 100 Hz). The computer’s internal clock, on

the other hand, operates at a much higher speed, implying that, without sophisticated programming, the display will appear on the screen a certain amount of time later than it should. Note that in recent experimental packages such as DMDX and E-Prime, considerable programming efforts have been invested to synchronize the clock with the raster, in order to eliminate this type of error. By analyzing this degree of variability relative to the one already present in response time data, Lincoln and Lane arrived at a conclusion similar to the one derived from the work presented here: Inaccuracies of this magnitude are negligible when compared with those found in human performance and are extremely unlikely to decrease the power of an experiment to reject the null hypothesis. This conclusion needs to be qualified only for studies that rely on relatively subtle variations in display durations. Because CRT monitors display events only in discrete increments (e.g., 14.3, 28.6, 42.9 msec, etc.), for a screen with a refresh rate of 70 Hz, studies that crucially depend on an exact variation of display duration, such as those using subliminal priming, are extremely sensitive to exact timing: The difference between, say, a 29- and a 43-msec display duration of a masked target can oftentimes have dramatic perceptual consequences (see Kouider & Dupoux, 2004, for an example), and for such purposes, an exact synchronization of the internal clock with the raster is essential. Apart from such restricted circumstances, display errors of the type investigated by Lincoln and Lane are unlikely to be of much concern.

On a more general level, one could, of course, argue that various sources of additional and undesired error variability (due, e.g., to response device imprecision, screen refresh delay, internal clock resolution, and other sources) may add up to a degree of error that eventually poses a problem. In addition, in situations in which several computers need to be synchronized (e.g., in neuroimaging studies), the error associated with each setup will add to a total error that might be considerable, or several error sources may interact in complex ways. Hence, it is advisable to minimize error at all levels. It is difficult to disagree with this suggestion on a general level; nevertheless, it remains the case that in the context of variability of human performance in standard behavioral tasks, errors associated with the measurement imprecision of input devices such as computer keyboards are bound to be very small. Chronometric research normally requires averaging across a large number of trials (typically, several hundred) to arrive at a valid estimate for a conditional mean, and for this reason, most psychologists intuitively know that the amount of additional error introduced by response device imprecision is unlikely to matter (to reiterate, the "lag" arising from imprecision is irrelevant, except under very particular circumstances, because it will affect all conditions alike). This intuitive knowledge accounts for the continued use of computer keyboards as response devices in published research on human cognition, and according to the argument above, there is nothing wrong with that practice.

AUTHOR NOTE

The author thanks Jeff Bowers, Simon Farrell, and Richard Plant for very helpful suggestions and discussion. Correspondence concerning this article should be addressed to M. F. Damian, Department of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, England (e-mail: m.damian@bristol.ac.uk).

Note—Accepted by the previous editorial team, when John H. Krantz was Editor.

REFERENCES

- ANDERSON, D. (1997). *Universal serial bus system architecture*. Reading, MA: Addison-Wesley.
- DONDERS, F. C. (1969). On the speed of mental processes (W. G. Koster, Trans.). *Acta Psychologica*, **30**, 412-431. (Original work published 1868)
- ERIKSEN, C. W., & SCHULTZ, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception & Psychophysics*, **25**, 249-263.
- FORSTER, J. C. (2007). *DMDX updates page*. Retrieved December 4, 2008, from www.u.arizona.edu/~jforster/dmdx.htm.
- FORSTER, K. I., & FORSTER, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, **35**, 116-124.
- HOWELL, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thomson/Wadsworth.
- JORDAN, L. E., & CHURCHILL, B. (1990). *Communications and networking for the IBM PC and compatibles* (3rd ed.). New York: Prentice Hall.
- KEELE, S. W. (1972). Attention demands of memory retrieval. *Journal of Experimental Psychology*, **93**, 245-248.
- KENNEY, J. F., & KEEPING, E. S. (1962). Mean absolute deviation. In *Mathematics of statistics: Part 1* (3rd ed., pp. 76-77). Princeton, NJ: Van Nostrand.
- KOUIDER, S., & DUPOUX, E. (2004). Partial awareness creates the "illusion" of subliminal semantic priming. *Psychological Science*, **15**, 75-81.
- LACOUTURE, Y., & COUSINEAU, D. (2008). How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. *Tutorials in Quantitative Methods for Psychology*, **4**, 35-45.
- LINCOLN, C. E., & LANE, D. M. (1980). Reaction time measurement errors resulting from the use of CRT displays. *Behavior Research Methods & Instrumentation*, **12**, 55-57.
- LUCE, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford: Oxford University Press.
- MCGILL, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 309-360). New York: Wiley.
- MEYER, D. E., & SCHVANEVELDT, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, **90**, 227-234.
- PLANT, R. R., HAMMOND, N., & TURNER, G. (2004). Self-validating presentation and response timing in cognitive paradigms: How and why? *Behavior Research Methods, Instruments, & Computers*, **36**, 291-303.
- RATCLIFF, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, **86**, 446-461.
- R DEVELOPMENT CORE TEAM (2009). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at www.R-project.org.
- SANDERS, A. F., & LAMERS, J. M. (2002). The Eriksen flanker effect revisited. *Acta Psychologica*, **109**, 41-56.
- TAYLOR, J. R. (1997). *An introduction to error analysis: The study of uncertainties in physical measurements* (2nd ed.). Sausalito, CA: University Science Books.
- ULRICH, R., & GIRAY, M. (1989). Time resolution of clocks: Effects on reaction time measurement: Good news for bad clocks. *British Journal of Mathematical & Statistical Psychology*, **42**, 1-12.

NOTES

1. Throughout this article, the terms *accuracy* and *precision* are used largely interchangeably. Note that in science and engineering, *measurement precision* is conventionally defined as the degree to which further measurements show the same or similar results (repeatability/reproducibility), and the relevant statistical index is error variability. *Accuracy* is typically defined as the degree of closeness of a measurement to its actual value, and it is indexed by bias (a nonrandom or directed effect; e.g., Taylor, 1997). Below, it is argued that the difference between a computer keyboard and a customized response device optimized for millisecond resolution lies in the degree of error variability, or measurement precision; by contrast, both devices have comparable accuracy because (ignoring the “lag,” which will apply to all experimental conditions alike) measurements taken from both devices will rapidly converge on a sufficiently similar mean.

2. Simulations were run in the statistical package R (R Development Core Team, 2009). The routines that generated the output reported in Figures 1 and 2 are available from the author upon request.

3. The argument is that the slight loss of power due to response device imprecision will be easily overcome by including a small number (relative to the one already required) of additional trials, in order for a true difference to reach statistical significance. This does not, of course, imply that researchers should aim at the bare minimum of trials necessary to reach a desired result; instead, formal or informal a priori power analyses should be used to determine the number of necessary observations for a given study.

(Manuscript received December 5, 2008;
revision accepted for publication October 20, 2009.)