DOES PSYCHOLOGY REQUIRE MORE (REFINED) DATA?

Denis Cousineau
School of psychology, University of Ottawa, Ottawa, Canada
Denis.Cousineau@UOttawa.ca

Abstract

Experimental psychology is a discipline that collects data to advance knowledge, foster theories and falsify wrong leads. It is a very fertile scientific field, collecting possibly billions of new data every year. Researchers often struggle hard to come up with innovative designs but this richness in methodological sophistication is not mirrored in the behavioral measures collected. Too often, research articles report preferences (or percent correct if there is a best choice) and mean response times. How many experiments, as clever as they might be, will it take to unravel the mysteries of the mind when there are only two variables observed? Here, we will argue that a lot of information is wasted in psychology. Even if it all starts with preferences and response times, averaging is only the beginning. We will discuss the importance of variances and skew in response times. We will also argue in favor of coefficients of changes. These are just examples; we wish to encourage researcher to use the same amount of imagination on measurement as on experimental design to stop the big datawaste.

Psychology and psychophysics are not easy fields of research. Researchers in those fields aim at finding the rules of perception and decision making. Yet, a percept is not a clearly defined phenomenon. Decision-making, presumably based on evidence gathering, is also elusive. Learning and training effects are likewise hard to define: do they imply local or global changes. Neurons implement these abilities. Yet, we cannot define what the *carrier of information* is across cells: is it the spikes, the synchronous firing, the strenght of activation or the time of activation? We are not even sure that neureons alone are responsible for our cognition, some suggesting that astrocytes may play an important role in learning.

Within such a loosely defined framework, it is important to support a multiplicity of approach, both from a methodological perspective and from a quantitative perspective. Yet, whereas we devote much energy to devise wise experimental manipulation, we cannot say the same regarding quantitative analyses. A lot of research papers only examine percent correct and mean response times. How many experiments will it take to uncover the rules of the mind when we use only two measures? It is time that we delve deeper into the data.

Considering that experimental psychologists collect in the order of billions of data per year, it represents a huge investment. We can approximate the cost of one datum in psychophysics and cognitive psychology at about 1 cent (counting the apparatus and the wages), which is affordable, but in total, it represents tens of millions of dollars spent gathering information.

Here, we will argue that we must invest the same ingenuity deriving meaningful measures. We do so by highlighting a few examples.

Beyond the mean

An intriguing paper was published a decade ago "What can a million trials tell us about visual search" (Wolfe, 1998). With so many data, the expectations were high that we would finally know whether visual search for a target proceeds in serial or in parallel. Sadly, all the response times (RTs) were aggregated using the mean so that in effect, there was much fewer data than announced. And the conclusion turned out to be a disappointment: both serial and parallel processing were possible considering the range of means observed. The only true conclusion of that study must in reality be that the means are not informative enough regarding this issue. More data do not represent more information when irrelevant results are kept.

One reason is that the standard Serial Self-Terminating Search model (SSTS), one of the two class of model tested, is based on two unknown parameters, but only one prediction was examined, a prediction based on the mean (see below). Hence, the data under determined the model.

The prediction tested is the ratio of increase in means for target absent trials as display size increases (*D*) relative to the increase in means for target present trials. Let $\mathbf{RT} = \sum_{i=1}^{d} \mathbf{T}_1 + \mathbf{T}_0$ where *d* is the number of locations visited (variable from trials to trials, but depending on the total number of locations *D*), $\mathbf{T}_1$ is the time to scan and decide target presence or not at one location, and $\mathbf{T}_0$ is the residual time (motor and perceptual processing). Both $\mathbf{T}_0$ and $\mathbf{T}_1$ are unknown parameters. Hence, $E(\mathbf{RT}|D) = E(\mathbf{d})E(\mathbf{T}_1) + \mathbf{T}_0$. The change in mean from a high display size condition relative to a low display size condition is

$$\Delta E(\mathbf{RT}|D \text{ vs. } 1) = E(\mathbf{RT}|D) - E(\mathbf{RT}|1) = (E(\mathbf{d}) - 1) \times E(\mathbf{T}_1) \tag{1}$$

Contrasting target present (self-terminating processing for which $E(\mathbf{d}) = (D+1)/2$) and target absent conditions (exhaustive processing for which $E(\mathbf{d}) = D$), we get

$$\frac{\Delta E(RT|D \text{ vs. } 1, \text{ target present})}{\Delta E(RT|D \text{ vs. } 1, \text{ target present})} = \frac{(D-1)E(\mathbf{T}_1)}{(\frac{D+1}{2} - 1)E(\mathbf{T}_1)} = 2 \tag{2}$$

This is the famous 2 to 1 ratio of target absent to target present slopes. It is a pointwise prediction regarding the ratio of changes in means. More importantly, this measure based on mean response times is independent of the two unknown parameters $\mathbf{T}_0$ and $\mathbf{T}_1$.

As seen from this example, means can be assembled to create *second-order* measures. Assuming a specific model, here the SSTS, this measure, the ratio of changes in mean, is predicted to be a constant and this is an easy-to-test prediction. It is based on changes because it is the best solution to remove residual processes' influences.

Other descriptive statistics can be used instead of the mean. For example, the change in variance for target absent trials is predicted to be a linear function of *D* according to the SSTS model.

Likewise, skew is expected to tend towards zero as display size is increased. This is predicted under SSTS by the Central Limit Theorem: As the number of location increases, distribution of RT should tend toward a normal distribution.

Finally, minimum RT can be used to make predictions. Indeed, for many parallel models, mean and variance can increase, but the best RT should be constant. Hence, the first percentile (if the number of observations permits) should be constant across conditions under such a model.

These last sources of information are rarely, if ever, mentioned in the Result sections of published research papers. Further, this information cannot be deduced from inspection of the mean results (or inspection of the means and variances if both aspects of the data are provided). Hence, a potentially important source of information is annihilated in the publication process.

Coefficient of change in variations

Means and variances can also be mixed in what are called Coefficient of change in variation relative to change in means (noted in short ΔCV).

Within SSTS, we get a nice prediction is we suppose that the variance of $\mathbf{T}_1$ is small:

$$\begin{aligned}
\Delta CV(\mathbf{RT}) &= \frac{\sqrt{\Delta Var(\mathbf{RT}|D \text{ vs. } 1, \text{ target present})}}{\Delta E(\mathbf{RT}|D \text{ vs. } 1, \text{ target present})} \\
&= \frac{\sqrt{Var(\mathbf{d})E^2(\mathbf{T}_1) + Var(\mathbf{T}_0) - Var(1)E^2(\mathbf{T}_1) - Var(\mathbf{T}_0)}}{E(\mathbf{d})E(\mathbf{T}_1) - E(1)E(\mathbf{T}_1)} \\
&\approx d\frac{\sqrt{\frac{D^2}{12}E^2(\mathbf{T}_1)}}{\frac{D+1}{2}E(\mathbf{T}_1)} = \frac{2}{\sqrt{12}}\frac{D}{D+1} \approx 0.577
\end{aligned} \tag{3}$$

because $Var(\mathbf{d}) = \frac{(D-1)(D+1)}{12} \approx D^2/12$ and for D large, $D/(D+1)$ tends toward 1. This prediction is again a pointwise prediction independent of the parameters $\mathbf{T}_0$ and $\mathbf{T}_1$. By using subtraction, it removes the residual times and by using a ratio, it removes the time of a single scan-and-decide process.

In a different task, this second-order measure can also make a useful prediction. In the redundant target detection task (Miller, 1982), RT is the time to locate one target attribute, but more than one can be presented. Detecting a single target attribute requires reaching a threshold *k* so that an asymptotic counting model predicts that $\mathbf{RT} = \min_{k:r\times N}(\mathbf{T}_1) + \mathbf{T}_0$ where *r* is the number of redundant attributes, *N* is the total number of channels activated by a target attribute (*N* considered large) and $\mathbf{T}_1$ is the time to register single evidence. Hence,

$$E(\mathbf{RT}|r) = \frac{E(\mathbf{T}_1|1)}{\sqrt[q]{r}} + E(\mathbf{T}_0) \quad \text{and} \quad Var(\mathbf{RT}|r) = \frac{Var(\mathbf{T}_1|1)}{\sqrt[q]{r^2}} + Var(\mathbf{T}_0) \tag{4}$$

so that, assuming again that the variance of the residual processes is small,

$$\Delta CV'(\mathbf{RT}) = \frac{SD(\mathbf{RT}|r) - SD(\mathbf{RT}|1)}{E(\mathbf{RT}|r) - E(\mathbf{RT}|1)} = \frac{SD(\mathbf{T}_1|1)}{E(\mathbf{T}_1|1)} = constant \tag{5}$$

In the redundant target detection task, we have predictions on skew, on minima and on the coefficient of change. However, there is not a single interesting prediction regarding the mean RTs. Mean response times are sometimes overrated.

Distributions

Going beyond the mean is just a first step. Variance (or standard deviations) can be used to make interesting predictions. More importantly, they can be combined into coefficients of change that are convenient to make predictions independent of the parameters. Finally, skew and minima are also important measures to test many parallel and serial models.

To go further, we can also make predictions on whole RT distributions. The Miller inequality (1982) is one such prediction in the redundant target detection task. It turns RT distributions in some baseline conditions into a bound use to test RT distributions in

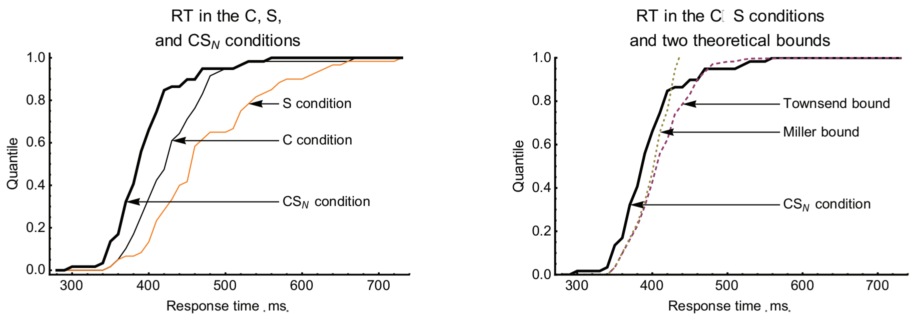

Figure 1. Example data from one participant in Cousineau, Engmann and Mestari (submitted) showing RT distributions for the two baseline conditions (only one target attribute is present, S or C) as well as the RT distribution in the critical CS condition). On the right, we see two boundaries built from the baseline conditions, the Miller bound and the Townsend bound.

conditions with high level of redundancy. This bound is useful to test some models of parallel processing.

Figure 1 shows example data (taken from Cousineau, Engmann and Mestari, submitted) with the Miller bound as well as the Townsend bound (Engmann, 2009). The observed cumulative distribution shows faster RTs than predicted, an indication supporting pooling of activation across the attribute detectors.

Likewise, the Δ plots are nice tools to make simple predictions. This plot shows the difference between the quantiles of two experimental conditions as a function of the mean between the quantiles.

Figure 2 shows three examples of Δ plots in the redundant target detection task. It is based on the triply redundant stimuli vs. the baseline conditions in which only one attribute is presented. See Schwarz and Miller (2012) for more details on this plot.

It is possible to show that in the visual search task, SSTS makes predictions regarding the target absent trials. It predicts that the slope of the Δ plot will be 1. This prediction is based on the technical assumptions that the distribution of processing times is part of the location-scale family of distribution. This assumption is taken from granted by researchers doing vincentizations. However, it is not clearly demonstrated.

For the redundant target detection task, the predicted relation is linear as well, but the slope depends on a free parameter:

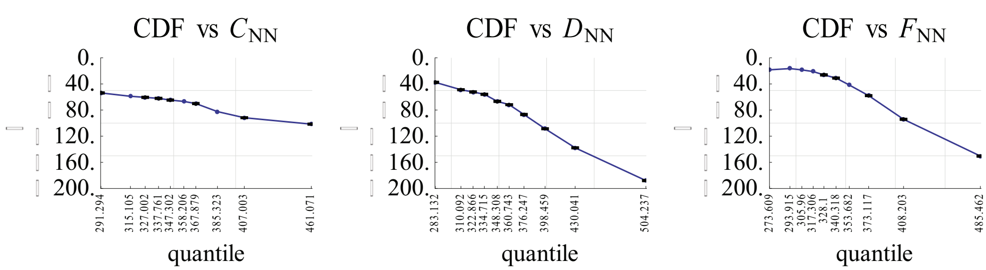

Figure 2. Examples of Δ plots based on the group results of Cousineau, Engmann, & Mestari (in press). The plot shows 9 points, one for each of the quantiles 0.05, 0.15, 0.25, …, 0.95. For each point, we see the difference between the slowest condition and the fastest condition on that quantile as a function of the mean of both quantiles.

$$\frac{\Delta Q(p)}{\Delta E(Q(p))} = 2\frac{\sqrt[\gamma]{r}-1}{\sqrt[\gamma]{r}+1} \quad (6)$$

in which *r* is the amount of redundancy and γ is the skew in the target detection RTs when there is only one attribute. As an illustration, for the γ offering the best fit to the RTs in the single attributes conditions ($\hat{\gamma} = 1.76$), we find a predicted slope of 0.605 when contrasting baseline conditions with the triple redundant condition and a predicted slope of 0.388 when contrasting baseline conditions with the double redundant conditions. The empirical results are 0.680 and 0.489 respectively.

Again, these are nice predictions that are obtained not just using one or two means, but using a whole range of quantiles. Hence, these are predictions that applies to the RTs, whether they are slow (e.g. the first quantile or below), medium (e.g., about the median) or fast (e.g., above the third quantile). Whereas alternate models could predict similar means, it is difficult to image strikingly different models predicting the same quantile dispersion.

Conclusion

In mechanics, physicists were faced with a similar problem: the two measures they could access were the position (*x*) of a mobile at certain times (*t*). It is exactly analogous to what we have in behavioral psychology where we have only access to response choice and response time. However, early physicists overcame this limitation by deriving measures of a higher order, such as speed and acceleration. These concepts turned out to be fertile and nowadays, no one could imagine mechanics without them.

Psychologist must do the same and find the equivalent of speed and acceleration regarding perception and decision making. When they are found, it will be possible to extract the maximum of information out of the raw data that are decision times and decision choices.

Here we have examined possible candidate measures. Among others, the coefficient of changes may be fruitful as it removes the residual times using a subtraction. At least, it turned out that for many models, it is possible to derive the theoretical value of the coefficient of change. Hence, they synthetize in a compact form some signature of the data that would go unnoticed if only the mean results were reported.

The overall message is simple: stop the data waste. Extract as much information from your results as you can by going beyond the mean results. The definitive measures may not be known at this time, but we should be encouraged to look for useful measures in any possible ways. Quantitative investigation of your results should take as much energy and ingenuity than experimental designs. As a final recommendation, always find a way to make your raw data available to the research community (web sites are less expensive than data), so that future propositions could be benchmarked on existing, relied upon, data sets.

Acknowledgements

This research was supported by grants from the CRSNG and the University d'Ottawa. I would like to thank the members of the Quibb joint laboratory for discussions and commentaries.

References

Cousineau, D. & Larochelle, S. (1997). *PASTIS: A Program for Curve and Distribution Analyses*. Behavior Research Methods, Instruments, & Computers, 29, 542-548.

Cousineau, D., Engmann, S., & Mestari, Z. (2012). *Evidence for parallel coactivation with double and triple redundant targets*. (submitted for publication), submitted, 1-59.

Engmann, S. (2009). Redundancy gain: manifestations, causes, and predictions. Université de Montréal: unpublished doctoral dissertation.

Miller, J. (1982). *Divided attention: Evidence for coactivation with redundant signals*. Cognitive Psychology, 14, 247-279.

Schwarz, W., Miller, J. (2012). *Response time models of delta plots with negative-going slopes*. Psychonomic Bulletin & Review, 19, 555-574.

Wolfe, J. M. (1998). *What can 1 million trials tell us about visual search?*. Psychological Science, 9, 33-39.