Perception & Psychophysics
2008, 70 (7), 1325-1326
doi: 10.3758/PP.70.7.1325

Better than average? When can we say that subsampling of items is better than statistical summary representations?

Dan Ariely
Duke University, Durham, North Carolina

Myczek and Simons (2008) have described a computational model that subsamples a few items from a set with high accuracy, showing that this approach can do as well as, or better than, a model that captures statistical representations of the set. Although this is an intriguing existence proof, some caution should be taken before we consider their approach as a model for human behavior. In particular, I propose that such simulation-based research should be based on a more expanded range of phenomena and that it should include more accurate representations of errors in judgments.

In a very interesting article, Myczek and Simons (2008) proposed that when individuals view a set of similar objects—a set of circles, trees, cars, books, or bananas—they do not extract an average property of the entire set (as I and others had proposed previously). Instead, they proposed that human observers extract these properties only for a subset of the items in the set and make their judgments on the basis of this subset. To support this claim, Myczek and Simons compared the results of human observer experiments with simulated performances and, by doing so, provided a starting point for an interesting discussion. Because I like the article and the general approach, I would rather not pick holes at particular experiments and specifications of the simulations and will, instead, focus on the ways that I think research using such simulations should be developed. In particular, I would like to propose that when engaging in simulation-based research, we should try to expand the range of phenomena accounted for and that we should worry about the error term that we use in these types of simulations.

What Data Should Simulation-Based Research Account For?

One of the worries about simulation-based research is overfitting, wherein a very specific model is designed for a narrow situation and its fit is specific not to the process, but to the setup. For example, in my article (2001), I included data not only on the mean judgment accuracy, but also on the ability to tell whether a particular size of circle was in the set of circles just presented or not. I called this task the *membership task*, and I personally found it particularly interesting that whereas the ability to tell whether an item was a member of a set was very close to being random, the ability to tell the mean of the set as a whole was relatively very good. In fact, the accuracy of the membership judgment was basically the same regardless of whether the number of items (members) in the set was 4, 8, 12, or 16. In my mind, it was this contrast between knowing very little about the individual items and being rather accurate about the mean of them all as a set that made this puzzle so curious. This difference, and not the performance on the mean task, was the basis for the *representation by statistical properties* hypothesis.

The process proposed by Myczek and Simons (2008) has very clear predictions for the membership task. If people subsample a number of the items and have a perfectly accurate representation of these, their accuracy should be proportional to the number of items they can select from the set. For example, if human observers can select two items and the task has four items, they should be correct in 75% of the cases (accurately identifying two of the members and randomly guessing for the other two). The fact that the results from my experiments show that observers are very close to 50% in their accuracy (see Figure 3 in Ariely, 2001) suggests that the simulation approach taken by Myczek and Simons cannot account for this aspect of the data. Of course, if we relax their assumption of accurately representing the individual objects that are sampled, the accuracy prediction of their model would be lower, but it is unlikely to reach 50%. Thus, even for the domain of size representations, I think it is premature to consider their model as a general mechanism.

In addition to accounting for slightly different tasks, such as membership tasks, it would be highly desirable if models such as that of Myczek and Simons (2008) would make predictions about other aspects of the task that they are modeling. In particular, researchers who have worked

D. Ariely, dandan@duke.edu

in the domain of statistical representation (mostly Chong & Treisman, 2003, 2005a, 2005b) have used vastly different exposure times (ranging from 50 to 1,000 msec), with relatively similar results. It seems to me that the Myczek and Simons approach should predict very clear patterns of performance on the basis of such variations—with shorter exposure times leading to a smaller number of items being sampled, and longer exposure times leading to a larger number of items being sampled—and higher accuracy. Although it is clearly possible to speculate within the Myczek and Simons approach about why exposure times have relatively small effects on the accuracy of statistical representations, it is also possible that their model could be expanded in some way to account for such findings.

Finally, I think that a general objective of simulation-based research is to take into account other, more distant tasks. For example, it would be nice if the same mechanism as that proposed by Myczek and Simons (2008) could also account for the quick and effortless extraction of averages in motion perception of multiple objects and for the ability to perform highly accurate representations of the mean orientation of multiple objects.

There is, of course, a question of what any individual research project should aim to accomplish, and you may wonder whether I am not demanding too much from the Myczek and Simons (2008) approach. It is possible, for example, to look at the Myczek and Simons approach as simply a *proof of concept*—showing that this type of system could produce these results. However, to actually make progress with such approaches, I suspect that a proof of concept is not enough and that the burden of proof for such simulations should be higher than the burden of proof we usually place on empirical research. But of course, this is my subjective bias. In my mind, successful simulation-based research should be applied to related data that it was not modeled directly after (such as the membership data), and more important, it should be used to generate new hypotheses that can be tested (e.g., exposure time). Without such efforts, there is a risk of making very specific models of particular paradigms, and not of the psychological process.

The Role of Error in Human Judgments and in Simulations

The second general point concerns the ways in which we should think about and implement errors in simulations. The simulations of Myczek and Simons (2008) have a very particular process, with its own built-in error (let's call this *process error*) but no judgment error. In this sense, they are *ideal observer simulations*. In contrast, human observers have their own decision-making process, which might include some built-in error (process error) *plus* some error and fluctuations in judgments over time (let's call this *judgment error*). This means that the performances of Myczek and Simons's simulations are limited by process error, whereas human performance is limited by process error *and* by judgment error.

Now the question is, How can we compare these two types of performances (simulation based and human based), and what can we say about this comparison? If the two processes yield the same performance, and if we all agree that human observers have an additional component of judgment error, would that not mean that the human observers are necessarily better than the simulation (presumably by the same magnitude as their judgment error)? In fact, I worry that the Myczek and Simons (2008) simulations simply replaced one type of error (judgment error) with another (process error) and that this modeling choice makes it very hard to directly compare these two types of performances. In essence, it seems to me that any simulation-based model that is used as a direct comparison with human performance must include a judgment error component. How to estimate this judgment error is, of course, not an easy task, but in the case of the task at hand (judging the average of multiple objects), I suspect that using a very small set (maybe one or two items) would be a good start to get an estimate of judgment error. With this estimate at hand, adding it as a constant to a simulation of the main process of interest (subsampling, in this case) could be much more informative.

In summary, I find the simulation approach proposed by Myczek and Simons (2008) interesting and provocative, and I am hoping that using this type of approach will be helpful in increasing our understanding of how individuals examine sets of similar objects. At the same time, I hope that future models of this nature will attempt to expand the scope of data that they aim to account for and, at the same time, that they will aim to integrate judgment error as an integral component of the model.

AUTHOR NOTE

Correspondence concerning this article should be addressed to D. Ariely, Duke University, 1 Towerview Road, Box 90120, Durham, NC 27708 (e-mail: dandan@duke.edu).

REFERENCES

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, **12**, 157-162.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, **43**, 393-404.

Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, **67**, 1-13.

Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, **45**, 891-900.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, **70**, 772-788.

(Manuscript received October 23, 2007;
revision accepted for publication July 9, 2008.)