WORKING
PAPER

Graph Comprehension

An Experiment in Displaying Data as Bar Charts, Pie Charts and Tables with and without the Gratuitous 3rd Dimension

MATTHIAS SCHONLAU
ELLEN PETERS

WR-618

September 2008

This paper series made possible by the NIA funded RAND Center for the Study of Aging (P30AG012815) and the NICHD funded RAND Population Research Center (R24HD050906).

This product is part of the RAND Labor and Population working paper series. RAND working papers are intended to share researchers' latest findings and to solicit informal peer review. They have been approved for circulation by RAND Labor and Population but have not been formally edited or peer reviewed. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.
RAND® is a registered trademark.

LABOR AND POPULATION

Graph Comprehension: An experiment in displaying data as bar charts, pie charts and tables with and without the gratuitous 3rd dimension

Matthias Schonlau[1], Ellen Peters[2]

[1]RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213; email: matt@rand.org

[2]Decision Research and University of Oregon;
Decision Research, 1201 Oak Street, Suite 200, Eugene, Oregon 97401; email: empeters@decisionresearch.org

Abstract

We investigated whether the type of data display (bar chart, pie chart, or table) or adding a gratuitous third dimension (shading to give the illusion of depth) affects the accuracy of answers of questions about the data. We conducted a randomized experiment with 897 members of the American Life Panel, a nationally representative US web survey panel. We found that displaying data in a table lead to more accurate answers than the choice of bar charts or pie charts. Adding a gratuitous third dimension had no effect on the accuracy of the answers for the bar chart and a small but significant negative effect for the pie chart. Viewing the graph/table for less than 8 seconds resulted in less accurate answers. Older age was associated with increased average viewing time (1.2 seconds per 10 years increase in age) but did not affect the accuracy of the answers. Greater numeracy was associated with more accurate answers.

Introduction

Literacy is usually operationalized into prose, document, and quantitative literacy (Kirsch, 1993). Document literacy is "the knowledge and skills required to locate and use information contained in various formats, including […] tables, and graphics" (Murray et al. 1997, p17). Graph comprehension is part of document literacy. Graph comprehension can be an important component in making intelligent, reasoned decisions including decisions about individuals' health. Data displays that require more cognitive effort will tend to decrease comprehension and use of information, particularly for individuals who are less numerate, a group that disproportionately includes the elderly (Peters, Dieckmann, et al., 2007; Peters, Hibbard, et al., 2007).

Pie charts and bar charts (also called pie graphs and bar graphs, respectively) are two commonly used graphs which both display categories of a single variable. The statistical-graphics community dislikes pie charts because it is difficult to compare relative size of slices of the pie chart. According to Steven's (1957) power law, individuals misjudge relative frequencies in a pie chart when the frequency is proportional to the area. Instead, Steven finds that displaying the area to the power of 0.7 allows individuals to make more accurate comparisons. We are not aware of any commercial software package that has implemented this law.

Cleveland and McGill (1985) have developed a taxonomy that implies a strict hierarchy of graphs for the purpose of comparison. The easiest task by accuracy is judging position along a common scale. This is true for bar graphs where the bars are aligned along the horizontal axis. Somewhat less easy it is the judging of lengths as can be found, for example, in stacked (or divided) bar charts. In a stacked bar chart, only the lowest segment within each bar is aligned on the horizontal axis, the stacked segments are not aligned with one another. Even less accurate are angle comparisons as is the case for pie charts. A similar study (Simkin and Hastie, JASA, 1987) in part confirmed this taxonomy. They found that the common scale (simple bar chart) was most accurate for the comparison judgment. The length judgment (stacked bar chart) was somewhat accurate; the comparison of angles (pie chart) was least accurate.

Other researchers have since reached different conclusions. Carswell (1992) found little difference in accuracy for position, length, or angle judgments. For judging

proportions-of-the-whole (rather than comparisons), Simkin and Hastie (1987) found pie charts to be as accurate as bar charts. Stacked bar charts were less accurate. Bar graphs were found to have consistently yielded lower scores on gist and verbatim comprehension than other graphs, including pie charts and modified pie charts (Fagerlin et al. 2007).

Despite the lower comprehension, Fagerlin et al. (2007) found that bar graphs are preferred over other graphs, including pie charts and modified pie charts. Some people perceive themselves as having difficulty interpreting pie graphs because they are unable to determine the exact proportion being represented; however, pie charts are useful for judging relative proportions (Hollands et al 2002) which may be important to comprehending the gist of the message, in particular.

Whereas some people prefer bar graphs, others may prefer line graphs or pie charts. Therefore, one option might be to present an array of graphic formats and ask people which they prefer (commonly available computer programs readily convert data into different graphic formats).

Graphs versus Tables

Very little research compares graphs to tables. Sanfey and Hastie (1998) found that respondents made more accurate judgments based on a narrative than respondents who were given the same information in bar graphs or data tables. Speaking on tables only Ehrenberg (1981) recommends table should include only two significant digits to facilitate mental arithmetic, rows should be ordered sensibly and a brief summary might help comprehension.

The gratuitous third dimension

The gratuitous third dimension refers to rendering two-dimensional graphs in three dimensions (i.e. with depth perception) even though the third dimension contains no information about the data. Such 3-d graphs are common in business presentations. For example, the default chart in PowerPoint is a 3-d side-by-side barchart.

The statistics-graphics community considers adding a depth cue using a third dimension to be a bad idea. One study found that the gratuitous third dimension was associated with lower accuracy (Carswell, Frankenberger, and Bernhard, 1991). This

study used line, bar, and pie charts and created versions of each, both with and without depth shading. Relative to 2-d bar charts 3-d bar charts are associated with slower decision times (Fischer, 2000). Fischer embedded two-or three dimensional bars with two or three dimensional frames and asked which of two bars was greater either before or after displaying the graph. The accuracy of reading numbers of a 3-d bar chart is lower than in a 2-d bar chart (Zacks et al, 1998).

Finally, some people simply prefer 3-d displays over 2-d displays (Levy et al. 1996; Fagerlin et al. 2007) and to the extent that that encourages them to read this may well influence the decision of how graphs should be displayed.

Method

We are interested in the following questions:

- Can information about the data be more easily recalled in a pie chart, a bar chart or a table?
- Does the gratuitous third dimension in pie charts and bar charts affect recall?
- Does better recall correlate with longer graph viewing time?
- Does recall decrease with age?
- How does numeracy relate to graph comprehension?

We conducted an experiment with respondents of the American Life Panel (ALP), a web survey panel (http://rand.org/labor/roybalfd/american_life.html). Respondents in the panel either use their own computer to log on to the Internet or use their own television, telephone, and Web TV (http://www.webtv.com/pc/) to access the Internet. The technology allows respondents without Internet access from their own computer to participate in the panel and furthermore use the Web TV for browsing the Internet or using email. About once a month, respondents receive an email with a request to visit the ALP URL and complete questionnaires that are, typically, no more than 30 minutes in length. Respondents are paid an incentive of about $20 per thirty minutes of interviewing (and proportionately less if an interview is shorter). The respondents in the ALP are recruited from among individuals age 18 and older who are respondents to the Monthly Survey (MS) of the University of Michigan's Survey Research Center (SRC). The

American Life Panel includes respondents from all US states. While roughly nationally representative it has few Hispanic members.

Each respondent was randomized to one of five experimental arms. Specifically, each respondent was shown the following text. "In a survey respondents were asked about their own health. The results are summarized in the [graph/table] below. Please look at the results and then go to the next page, which contains questions about this [graph/table]." The five graphs/table corresponding to the five experimental arms are shown in Figure 1. The four graphs consisted of two pie charts and to bar charts each shown as a two-dimensional graph and as a three-dimensional graph. The third dimension in the three-dimensional graphs did not convey any information. All five graphs/table displayed the same five percentages for self-assessed health. All graphs were constructed in Excel using default settings. No attempt was made to rotate the pie charts in a particular way.

On the next page the respondents were asked the following three questions (answer choices in parentheses): 1) Which was the largest answer category? (excellent, very good, good, fair, poor) 2) Which was the smallest answer category? (excellent, very good, good, fair, poor) 3) Roughly what percentage of respondents chose the largest category? [0..100].

For the largest and smallest category we report the percentage of correct answers in each arm and test whether they are significantly different from one another. For the third question we analyze the difference between the percentage reported and the true percentage. We also analyze how the accuracy changes as a function of viewing time and age.

Numeracy was measured using a scale developed by Lipkus et al), 2001). The 11 item scale contains various questions about percentages and fractions.

Results

A total of 897 panel respondents participated in the experiment. The individual experimental arms contained between 165 and 182 respondents each.

<u>2-D pie chart, 2-D bar chart, table</u>

Table 1 gives results for the two-dimensional pie chart, the bar chart and the table. Specifically, the percentages of respondents with the correct answer for the largest and smallest categories and the average absolute difference of the estimated minus actual percentage for the largest category are shown. Because the largest and the second largest category were very close (34.9 vs. 34.7), one of the rows analyzes the question counting either of these two answer choices as correct. In all four comparisons the experimental arm "Table" does best. In particular for the largest category (counting the second largest category as incorrect) the experimental arm "Table" achieves 60.4% versus 41.9% and 41.0% respectively. The 2-D pie chart and the 2-D bar charts do about equally well. Significance will be discussed below in the context of age and viewing time.

The gratuitous 3rd dimension

Table 3 gives the results comparing the bar charts and pie charts with and without the gratuitous third dimension. With one exception the gratuitous third dimension seems to only have a small effect. The 3-D pie chart does significantly worse than the 2-D pie chart (t-test, p=0.03), though the difference is small (2.4% reduction; from 11.3% to 8.9%).

The exception is a large difference between the three-dimensional and the two-dimensional pie chart for the percentage of correct answers for the largest category (counting the second largest category as incorrect). The percentage of correct answers is much higher for the three-dimensional pie chart. We attribute this difference to the orientation of the three-dimensional pie chart which we believe distorts the relative difference between the largest and the second largest category. We cannot verify this hypothesis because our design included only one orientation of the pie charts.

Viewing Time and Age

Respondents spent between 2.7 and 275 seconds looking at the graphs before proceeding to the next page. The median was 13.8 seconds. Time spent did not vary significantly by experimental arm (ANOVA, p=0.92). Older respondents looked at the graph longer. On average, each additional ten years in age resulted in an additional 1.2

seconds of viewing time ($p<0.001$) (Increase on full data is 1.5 seconds per ten years of age. After removal of some large outliers [time>60sec] increase was reduced to 1.2 seconds).

Figure 2 shows a scatter plot of the absolute difference between estimated and actual percentage of the largest category over viewing time. A lowess smoother is added to guide interpretation. A lowess smoother is useful for identifying trends. The algorithm takes a weighted average of earlier and later observations to smooth out noise. The lowess smoother suggests that as long as the respondent looked at the graph for at least 8 seconds time was not related to the accuracy of the answer. Below about 8 seconds, the accuracy of the answer significantly decreased (chi squared for smallest category, $p=0.016$). There were 58 respondents who looked at the graph eight seconds or less versus 828 who looked at it longer. However, these 58 respondents on average were also much younger (average 36.8 vs. 52.3 years old).

To investigate whether the eight second viewing time is associated with worse accuracy for all questions and even after adjusting for age we conducted three logistic regression analyses of whether or not a respondent answered correctly for the largest category (with and without counting the 2nd largest category as correct), and the smallest category. We also conducted a regular Gaussian regression analysis of the absolute difference of estimated minus actual percentage for the largest category (Table 3). For the largest category, viewing time greater than eight seconds significantly increases the odds of a correct answer between 2.5 and 3.5-fold (exp (0.91) and exp (1.26)). There was also an increase for the smallest category; however, after adjusting for age it was no longer significant. Viewing time longer than eight seconds also improved the percentage estimate for the largest category reducing the absolute difference by an average of 4.2%.

The referent category for the five experimental arms in Table 3 is “Table”. For the absolute difference for the largest category all pie and bar charts are significantly different from “Table”. For percentage correct for the smallest category none of the arms is significantly different from “Table”. For percentage correct for the largest category the result is more complex. If one only counts the one category as correct, then all other experimental arms (except the 3-D pie chart as previously noted) are significantly worse

than the table. If one counts the second largest category as correct, then only the pie charts are significantly worse than the table.

"Cheating" works

After seeing the questions some respondents used the "back" button in the web browser to look at the graphs a second time. The "back" button was supposed to be disabled but was indeed functioning. Respondents did not know that the "back" button was supposed to be disabled. A total of 8 respondents (<1%) went back. Their efforts were rewarded. 100% of the back button users identified one the largest two categories vs. 96% of those who didn't. 88% of the back button users identified the smallest category versus 73% of those you didn't. Most convincingly, the average absolute difference in remembering the largest category was 1.1% for those who used the back button versus 9.1% for those who did not. None of the differences were significant due to the small number of back button users.

Numeracy

Respondents in this sample were fairly numerate. The average score was 8.54 (out of 11) and the quartiles were as follows: 7 (1^{st} quartile), 9 (median), and 11 (3^{rd} quartile). Higher numeracy was significantly correlated with correctly identifying the largest category (with 2 correct answer choices) ($p<0.001$) , the smallest category ($p<0.001$) and with the absolute percent difference for the largest category ($p<0.001$) even after adjusting for the variables in Table 3. Numeracy was not significant in predicting the largest category by itself. Based on adjusted regressions increasing the numeracy from the first quartile (7) to the third quartile (11) increased the probability of a correct answer for the largest category (with 2 correct answer choices) from 94% to 97%, for the smallest category from 69% to 79%, and reduced the absolute difference between the estimated and true percentage of the largest category from 12.1% to 6.7%. There were no significant interactions between numeracy and age or numeracy and experimental arm for any of the questions.

Conclusion

Displaying the data in a table achieved equally or more accurate responses compared to displaying the data in a pie chart or in a bar chart. The table was particularly successful in distinguishing between the two largest categories which were nearly identically. Adding a gratuitous third dimension to the bar chart did not affect the accuracy of the answers. Adding a gratuitous third dimension to the pie chart lead to a small (2.4% out of 11.3%) but significant decrease in the average absolute difference of estimated minus actual percentage. The three dimensional pie chart achieved surprisingly accurate responses in distinguishing between the two largest categories. However, we attribute this to the particular rotation of the 3-D pie chart which appeared to amplify the difference between the two largest categories. Identifying the smallest category seemed to work equally well for all charts-both two and three dimensional - and the table. We found that each additional 10 years in age increased the viewing time of the graphs by 1.2 seconds. Longer viewing time did increase the accuracy of the answer as long as respondents had looked at the graphs at least for about eight seconds. A short viewing time (<8 seconds) was associated with a significant decrease in accuracy for all questions. After adjusting for short viewing time, age did not affect accuracy except for one question. In this question, the smallest category, old-age was associated with a very slight increase in the odds to answer the question correctly. The increase was so small that it is of no practical consequence.

Greater numeracy was significantly related to better recall about the graph. Numeracy (both objective and subjective scales) has also been found to relate to the understanding of survival graphs (Zigmund-Fisher et al. 2007). Because numeracy measures and graph comprehension measure two separate domains of literacy – quantitative and document literacy – these findings imply some overlap between these domains.

Our experiment is limited in several ways. While we randomized the type of the display (bar chart, pie chart, table) and whether or not the graph had a gratuitous third dimension, we used only one set of percentages for the five categories. It is unclear whether the findings generalize to substantially different percentages. It is also unclear whether the results generalize to (much) larger number of categories. The fact that the

table yielded the most accurate responses is consistent with the idea that for few categories tables may be the best representation. Respondents might have had an opinion on the topic we were asking about, self-assessed health. This opinion might have helped respondents to guess the answer. However, this would have been true across all experimental conditions. Our experiment was unusual in that the five categories were ordered.

It is an open question whether different graphs are better for different users. It may be useful to give users more than one display format to let them choose which one to look at. Visual displays may be especially helpful to those who are least numerate. Context may matter a lot. Research in graph comprehension and document literacy in general has the potential to be invaluable if it succeeds in improving the ability of some respondents to better understand financial information (such as terms of credit card or mortgage loans) or information about health care and health care related choices.

Acknowledgement

Support for this research comes from grant 1R01AG020717from the National Institute of Aging of the U.S. National Institutes to RAND (Arie Kapteyn, P.I.). Additional support comes from the National Science Foundation (#0517770) (Peters, P.I.)

References

Carswell, C.M., Frankenberger, S., and Bernhard, D. (1991). Graphing in depth: Perspectives on the use of three-dimensional graphs to represent lower-dimensional data, Behaviour & Information Technology, 10(6), 459-474.

Cleveland, W. S., McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. Science, 229 (4716), 828-833.

Cleveland, W. S., McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387), 531-554.

Ehrenberg, A.S. C. (1981). The Problem of Numeracy. The American Statistician, 35(2), 67-71.

Fagerlin, A., Zikmund-Fisher, B., Ubel, P., Smith, D. (2007). Measuring numeracy and the impact of numeracy on medical decision making. Presented at the Annual Meeting of the Society of Behavioral Medicine. Washington, DC.

Fagerlin, A., Zikmund-Fisher, B.J., Ubel, P.A., et al. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. Medical Decision Making, 27 (5), 672-680.

Fischer, M.H. (2000). Do irrelevant depth cues affect the comprehension of bar graphs? Applied Cognitive Psychology, 14(2), 151–162.

Friel, S.N., Curcio, F.R., Bright, G.W. (2001). Making sense of graphs: critical factors influencing comprehension and instructional implications. Journal for Research in Mathematics Education, 32(2), 124-158.

Hollands, J.G., Tanaka, T., Dyre, B.P. (2002). Understanding bias in proportion production. Journal of Experimental Psychology: Human Perception and Performance, 28(3), 563-574.

Kirsch I, Jungeblut A, Jenkins L, Kolstad A (1993). Adult literacy in America: A first look at the findings of the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Levy, E., Zacks, J., Tversky, B., Schiano, D. (1996). Gratuitous graphics? Putting preferences in perspective. Conference on Human factors in computing systems. Vancouver, BC, 1996, 42-49.

Lipkus, I.M., Samsa, G., &Rimer, B.K. (2001). General performance on a numeracy scale among highly educated samples. Medical Decision Making, 21, 37–44.

Murray, T. S., Kirsch, I. S., Jenkins, L. B. (1997). Adult literacy in OECD countries: A technical report for the first international adult literacy survey. Washington, DC: National Center for Education Statistics, U.S. Government Printing Office.

National Center for Education Statistics (1992). National Adult Literacy Survey (NALS).

National Center for Education Statistics (2003). National Assessment of Adult Literacy (NAAL).

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J.H., Mertz, C.K. (2007). Less is more in presenting quality information to consumers. Medical Care Research & Review, 64(2), 169-190.

Peters, E., Hibbard, J.H., Slovic, P., Dieckmann, N.F. (2007). Numeracy skill and the communication, comprehension, and use of risk and benefit information. Health Affairs, 26(3), 741-748.

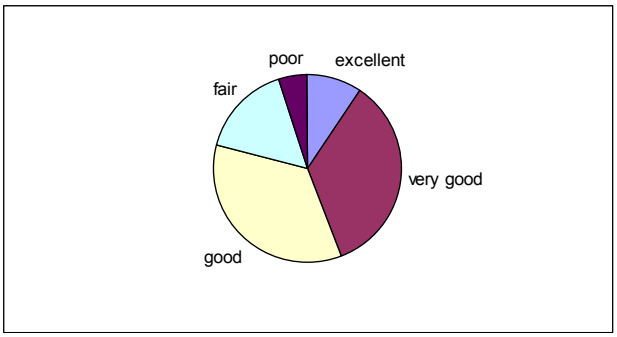
Sanfey, A., Hastie, R. (1998). Does evidence presentation format affect judgment? Psychological Science, 9(2), 99–103.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review* 64(3):153–181.

Zacks, J., Levy, E., Tversky, B., Schiano, D. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. Journal of Experimental Psychology Applied, 4(2), 119-138.

Zikmund-Fisher B. J., Smith D.M , Ubel P. A. and Fagerlin A. (2007). Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations Med Decis Making, 27, 663.

	Percent
excellent	9.4
very good	34.7
good	34.9
fair	16.2
poor	4.9



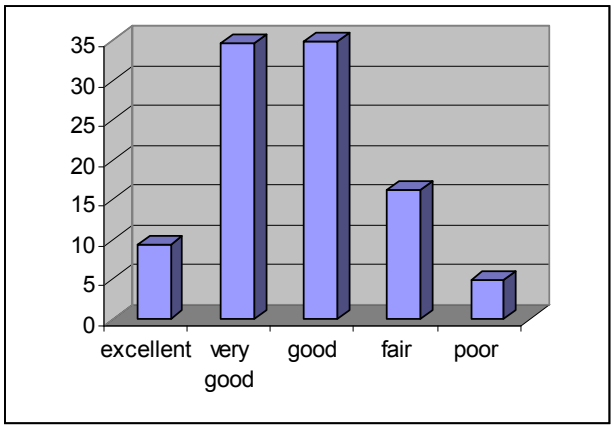


Figure 1: The five graphs/table corresponding to the five experimental arms

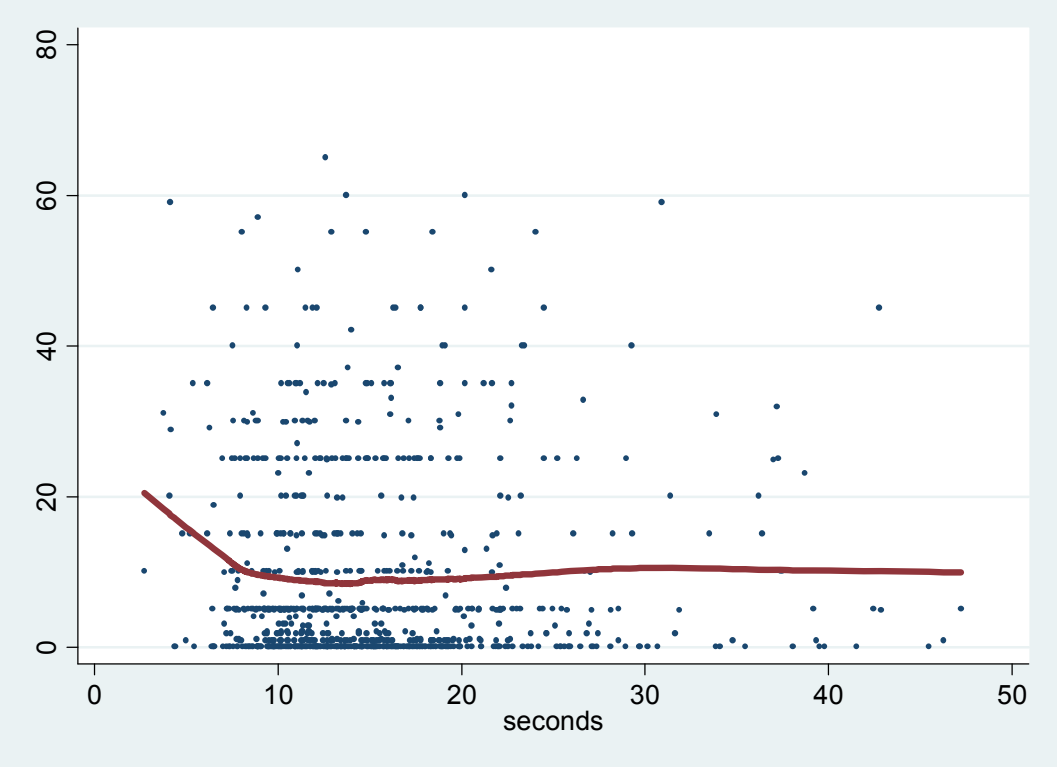


Figure 2: Absolute difference between estimated and correct percentage over time spent looking at graph/table. The line is a lowess smoother.

	2d pie chart	2d bar chart	Table
largest category (% correct)	41.9	41.0	60.4
largest category (% correct incl 2nd largest categ)	91.1	94.9	97.7
smallest category (% correct)	72.1	75.8	74.2
absolute difference (estimate - actual)	8.9	9.5	5.8

Table 1: Percentage of respondents with the correct answer (absolute difference of the estimated versus actual percentage of the largest category) for the two-dimensional pie chart, bar chart and table.

	3d pie chart	Diff 3D - 2D pie chart	3d bar chart	Diff 3D - 2D bar chart
largest category (% correct)	59.2	17.3	40.5	-0.5
largest category (% correct incl 2nd largest categ)	93.6	2.5	95.7	0.8
smallest category (% correct)	72.2	0.1	75.6	-0.2
Average absolute difference to actual value	11.3	2.4	10.6	1.1

Table 2: Percentage of respondents with the correct answer (absolute difference of the estimated versus actual percentage of the largest category) for the three-dimensional and bar charts. Gray shading indicates significance at 5%.

		Largest category log odds ratio		Largest category (2 correct answers) log odds ratio		smallest category log odds ratio		absolute difference for largest category Coefficient	
age		-0.01		0.01		0.03	**	0.01	
	p-value	0.19		0.42		0.00		0.70	
Indicator		0.91	**	1.26	**	0.31		-4.16	*
time >8 sec	p-value	0.00		0.00		0.30		0.02	
Table		0.00		0.00		0.00		0.00	
	p-value	NA		NA		NA		NA	
3D bar		-0.78	**	-0.73		0.04		4.90	**
	p-value	0.00		0.25		0.87		0.00	
2D bar		-0.76	**	-0.96		0.00		3.92	**
	p-value	0.00		0.12		0.99		0.01	
3D pie		-0.03		-1.23	*	-0.18		5.62	**
	p-value	0.87		0.04		0.47		0.00	
2D pie		-0.73	**	-1.52	**	-0.13		3.09	*
	p-value	0.00		0.01		0.60		0.03	
constant		-0.13		2.27	**	-0.45		9.01	**
	p-value	0.73		0.00		0.23		0.00	

* p<0.05, ** p<0.01

Table 3: Coefficients for four three logistic regressions of whether or not a respondent answered correctly for the largest category (with and without counting the 2nd largest category as correct) and the smallest category, and a Gaussian regression of the absolute difference on covariates.