Schonlau M, Peters E. (2012).Comprehension of Graphs and Tables Depends on the Task: Empirical Evidence from two web-based studies. *Statistics, Politics and Policy*. 3(2); Article 5.

Comprehension of Graphs and Tables Depend on the Task: Empirical Evidence from Two Web-Based Studies

Matthias Schonlau[1] and Ellen Peters[2]

[1]University of Waterloo

[2]Ohio State University

Matthias Schonlau is a Professor, Department of Statistics and Actuarial Sciences, University of Waterloo, 200 University Ave West, Building M3, Waterloo, Ontario, Canada, N2L 3G1 , (email: schonlau@uwaterloo.ca);

Ellen Peters is an Associate Professor, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH 43210 (email:peters.498@osu.edu).

Acknowledgment

The work was conducted while the first author was employed at RAND and the German Institute for Economic Research. This research was supported from grant 1R01AG020717 from the National Institute of Aging (USA) to RAND (Kapteyn, P.I.) with additional support from the National Science Foundation (SES-0517770, -1047757, and -0820197) (Peters, P.I.). We thank two anonymous reviewers for their thoughtful reviews which improved the paper.

Comprehension of graphs and tables depend on the task: Empirical evidence from two web-based studies

Abstract

Graphs and tables are an effective means of communication. However, relatively little experimental work exists examining differences between various formats in how well people understand provided information. We conducted two web-based experiments with a large, diverse sample to explore the effects of display format on respondents' comprehension. We found that comprehension depended on task. Graphs were better for estimating differences; however, tables were better when estimating equality and sums. We found 3D display formats reduced comprehension of pie charts but not of bar charts. Although pie charts never assisted comprehension, they often did not significantly impair comprehension either. Comprehension based on a 3-way table was as good as that for clustered bar charts but was worse for divided bar charts. Information can be conveyed graphically even with 3-way tables, but the choice of display format needs to be sensitive to the task at hand.

Key words: Clustered Bar chart, Divided Bar Chart, Pie Chart, 3D Display, Table

1. Introduction

Mass media play a central role as an intermediary between public policy and private citizens (Koch-Baumgarten and Voltmer 2010). Effective communication of information is essential to reach both the public and busy policy makers. Graphs such as pie charts and bar charts often emerge as a method of choice for conveying quantitative policy information. For example, on May 6, 2008 the New York Times ran a story "E.R.'s are busy, but fewer patients are uninsured" that displayed a set of two pie charts about the usual source of medical care of patients in emergency rooms (ERs) in 2003 and 1996. The graph in the New York Times was based on a graph in a research paper (Weber, et al. 2008, Figure 3). However, the New York Times made some changes: The original graph was a divided bar chart with four bars corresponding to one year each. The New York Times created a pie chart for the first and the last year leaving out two intermediate years. Was the use of pie charts a good choice in this case?

Most professional data analysts dislike the use of pie charts on grounds that individuals assess angles less accurately than length, but others support their use ("The general antipathy statisticians have toward pie charts is misinformed."(Wilkinson 2001) and "The traditional prejudice against the pie chart is misguided" (Spence and Lewandowsky 1991)). The controversy may be due, in part, to an insufficient focus on the focal task. The answer to "Which graph type or table is most easily understood?" may vary by the task to be performed.

Controversy is not restricted to pie charts. Three-dimensional (3D) graphical representations of two-dimensional (2D) data (e.g., 3D pie and bar charts) have evoked strong opinions among some scientists: "The icon of the garish is the 3-D pie chart" (Wilkinson 1994). The substantial resistance to 3D graphs is based both on minimalist design principles (Tufte 1983) and on the argument that the depth cues they provide make the graphs difficult to comprehend (Haemer 1951).

Conducting experiments with graphical displays is important to verify the veracity of such claims (Cleveland and McGill 1987). Experimentalists have confirmed the gratuitous 3rd dimension leads to less accurate magnitude judgments (Siegrist 1996, Zacks, Levy, Tversky and Schiano 1998, Stewart, Cipolla and Best 2009) and slower decision times (Siegrist 1996, Fischer 2000). However, these effects tend to be small and not particularly robust. In addition, participants also express a preference for 3D displays (Levy, Zacks, Tversky and Schiano 1996) and therefore may spend more time with 3D displays. Whether or not to employ 3D display may be a tradeoff between accuracy and visual interest (Shah, Freedman and Vekiri 2005).

These communication issues are important because statistical graphs are widely used, using about 6.6% of the total print space in scientific publications

alone (Cleveland 1984). The use of graphs in natural-science publications is greater than that in social–science publications (Best, Smith and Stubbs 2001). About 1.6% of printed graphs are 3-dimensional (Zacks, Levy, Tversky and Schiano 2001) though the use of such graphs at scientific meetings may be much larger, because some people like them (Stewart, et al. 2009) and software such as Microsoft Excel and PowerPoint makes them easy to create. Pie charts are more frequently used in the popular press than in scientific publications, but even in the popular press, bar charts outnumber pie charts (Spence 2005).

Studies have demonstrated that comprehension of some types of graphs and tables varies across tasks.

- First, for comparing the relative size of two categories or cells, judgment is most accurate along a common scale (simple bar chart), has intermediate accuracy when assessing length (divided or stacked bar charts), and is least accurate when assessing angles (pie charts) (Cleveland and McGill 1984, 1985, Simkin and Hastie 1987, Heer and Bostock 2010).
- However, for estimating the absolute size of proportions, bar charts yield results similar to pie charts (Simkin, et al. 1987, Spence 1990). Comprehension based on divided bar charts is worse (Simkin, et al. 1987). Tables may have a slight edge in accuracy but also take longer to read (Spence 1990). Tables are more accurate and faster than line charts (Meyer, Shamo and Gopher 1999). In general, tables are thought to be better than graphs for point reading and recall (Vessey 1991). When information providers do not provide axes and scales on bar charts, estimating proportions becomes easier with pie charts than with bar charts (Hollands and Spence 1998).
- Third, for estimating differences, a bar chart was more accurate than a pie chart in a population of cancer patients but not in a population of students (Feldman-Stewart, Kocovski, McConnell, Brundage and Mackillop 2000).
- Fourth, for judging which of two categories is bigger, tables, pie charts and bar charts tend to yield similar accuracy scores (Spence, et al. 1991, Meyer, Shinar and Leiser 1997) though another study found that pie charts performed worse than bar charts (Feldman-Stewart, et al. 2000). Providing a scale for this task improves accuracy, and providing the numerical value in addition reduced the time needed to complete the task (Feldman-Stewart, Brundage and Zotov 2007). When restricting time, tables perform worse (Spence, et al. 1991).
- Fifth, for judging which of two sums of proportions are greater, pie charts were more accurate than bar charts (Spence, et al. 1991) because neighboring cells are visually easy to sum.
- Sixth, for identifying a maximum of a row tables were more accurate than bar charts or line graphs (Meyer, et al. 1997).

- Finally, graphs tend to be more accurate than tables when identifying trends and patterns and for data interpolation (Vessey 1991, Meyer, et al. 1997).

A more specialized literature has emerged for the communication of health risks. A review of graphs such as pictographs, risk ladders and survival curves was conducted in the context of health risk communication (Ancker, Senathirajah, Kukafka and Starren 2006). The authors concluded that the best design depends on whether the purpose is to produce quantitative judgments or to promote behavior change. Pictographs display an array of icons (e.g. 10 rows of 10 little stylized people each) some of which are shaded (e.g. 10 of the 100 stylized people) to represent percent risk (e.g. 10%). Research has shown that pictographs work well to communicate risks, especially for populations with lower numeracy (Fagerlin, Wang and Ubel 2005, Hawley, et al. 2008, Garcia-Retamero and Galesic 2009). Another study comparing icon arrays filled with ovals, single bars (vertical or horizontal), and pie charts (with a single slice) found that a vertical bar chart yielded the greatest comprehension (Feldman-Stewart, et al. 2007).

The literature thus far has not evaluated how graphs and tables perform for other common tasks such as identifying which category increased most from one graph to another, or performing simple numerical operations ("Compared to year 1, is the category approximately twice as large in year 2?"). Also, research thus far has concentrated on one or two well thought out tasks usually conducted with college-student subjects rather than considering overall performance on multiple tasks in a more diverse population. With the notable exception of (Cleveland, et al. 1984), published studies also have focused on the visualization of one variable at a time. Clustered bar charts which display two variables have therefore not been explicitly evaluated. Even comparisons of a variable in two different years require the visualization of two variables and have not been studied empirically. It is unclear to what extent findings generalize for the display of multiple variables and more complex questions.

We conducted two experiments embedded in a web survey to explore whether and to what extent the gratuitous third dimension and display type affected respondents' graph comprehension. The experiments involved the visualization of 2 and 3 variables, respectively. Our experiments were conducted with a large, diverse sample of subjects and we considered a greater range of tasks or comprehension questions than have previous studies. Sections 2 and 3 describe the two experiments. Section 4 concludes with discussion.

2. **Study 1**

2.1 Experimental Design

We conducted a randomized experiment in a web survey of 2,254 respondents in the American Life Panel (www.rand.org/labor/roybalfd/american_life.html).

Respondents in the panel either use their own computer to log on to the Internet or a Web TV (http://www.webtv.com/pc/), which allows them to access the Internet, using their television and a telephone line. 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 MS is the leading consumer sentiments survey that incorporates the long-standing Survey of Consumer Attitudes (SCA) and produces, among others, the widely used Index of Consumer Expectations. The American life panel is a convenient and fast platform for such research but maintaining a national panel comes at a price: Currently, the ALP costs $2 per person per survey minute (e.g. a 30-minute survey with 2,254 respondents costs $ 135,240; a large portion of which goes to respondents in the form of incentives).[1]

Data were shown concerning where people usually receive health care in five usual-care categories for two different years. The data stem from the previously mentioned article in the May 6, 2008 issue of the New York Times; however, we did not use any of the text in the article. The New York times article was based on a research paper (Weber, et al. 2008). Respondents were given the following explanatory text: "Patients who went to the emergency room were asked about their usual source of medical care in 1996 and 2003. The graphs below give percentages in various categories of usual source of medical care. Please look at the graphs below and use them to answer the following questions." Accurate comprehension of such graphs might be important, for example, for understanding the effects of health policy shifts in the intervening years.

Each respondent was shown one of 5 different displays: a 2-way table (Figure 1), a 2D bar chart for each year (Figure 2), a 2D pie chart for each year (Figure 3), and two arms with 3D versions of the graphs (Figure 4 and Figure 5). The experimental arm with the 3D pie chart consisted of 4 versions, specifically four 90 degree rotations of the 3D pie chart (one additional rotation is shown in Figure 6). We created each graph in Microsoft Excel 2007 because Excel is a widely available package for creating such graphs. Except for font size (explained below), we used the default setting in Excel. Further, each of the 8 displays (5 types+ 3 additional rotations) consisted of two versions with labels in 9pt and 12pt font sizes of type Calibri (default for excel graphs). (We controlled the font size of the display including labels. We did not control font size of the

[1] Some experiments are now being done using Amazon's Mechanical Turk. Such experiments are much cheaper, though little is known about the participants and the control over presentation format may be more limited. Heer, J., and Bostock, M. (2010), "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design," in *CHI*, Atlanta, Georgia: ACM, pp. 203-212.

questions). We found that font size had no effect on comprehension; therefore we will not consider font size further.

Source of medical care	Year 1996	Year 2003
HMO	4	3
Emergency room	8	7
None	10	10
Health centers	26	22
Doctor's office	52	59

Figure 1 : Study 1: Display type is a table.

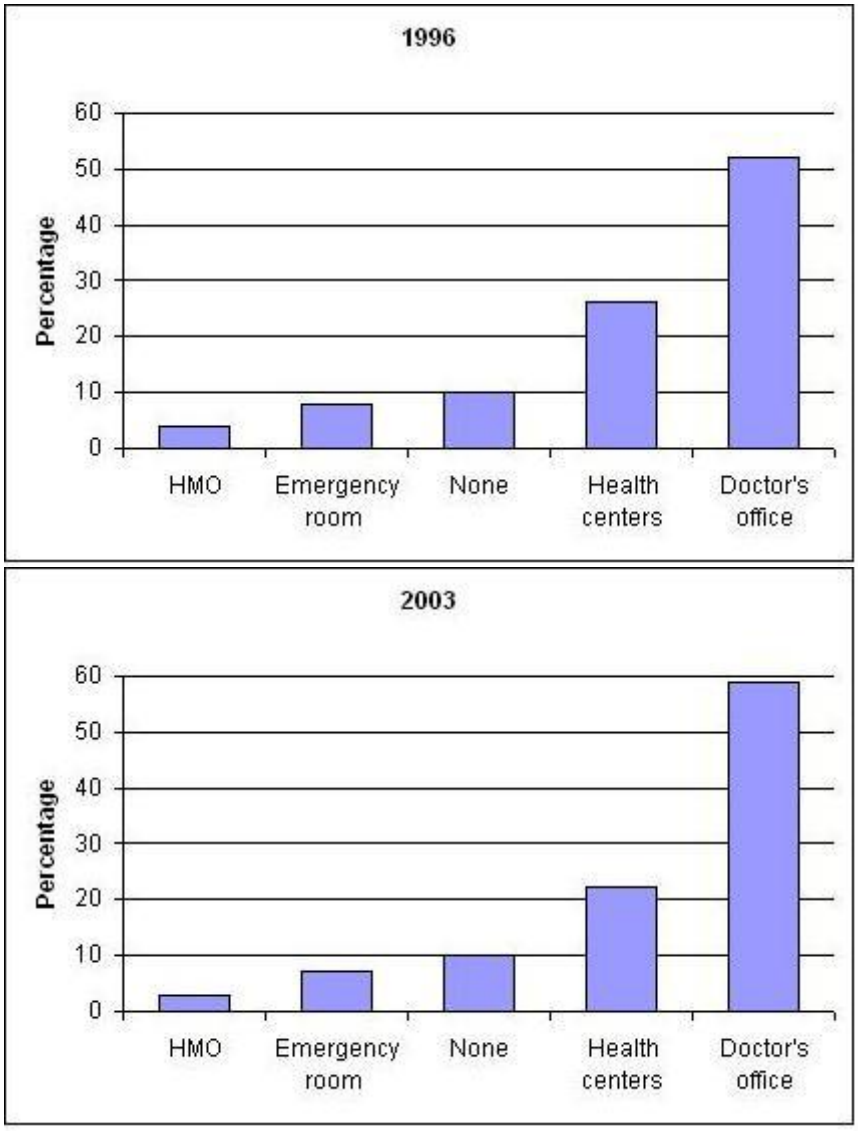

Figure 2 : Study 1: Display type 2D bar charts

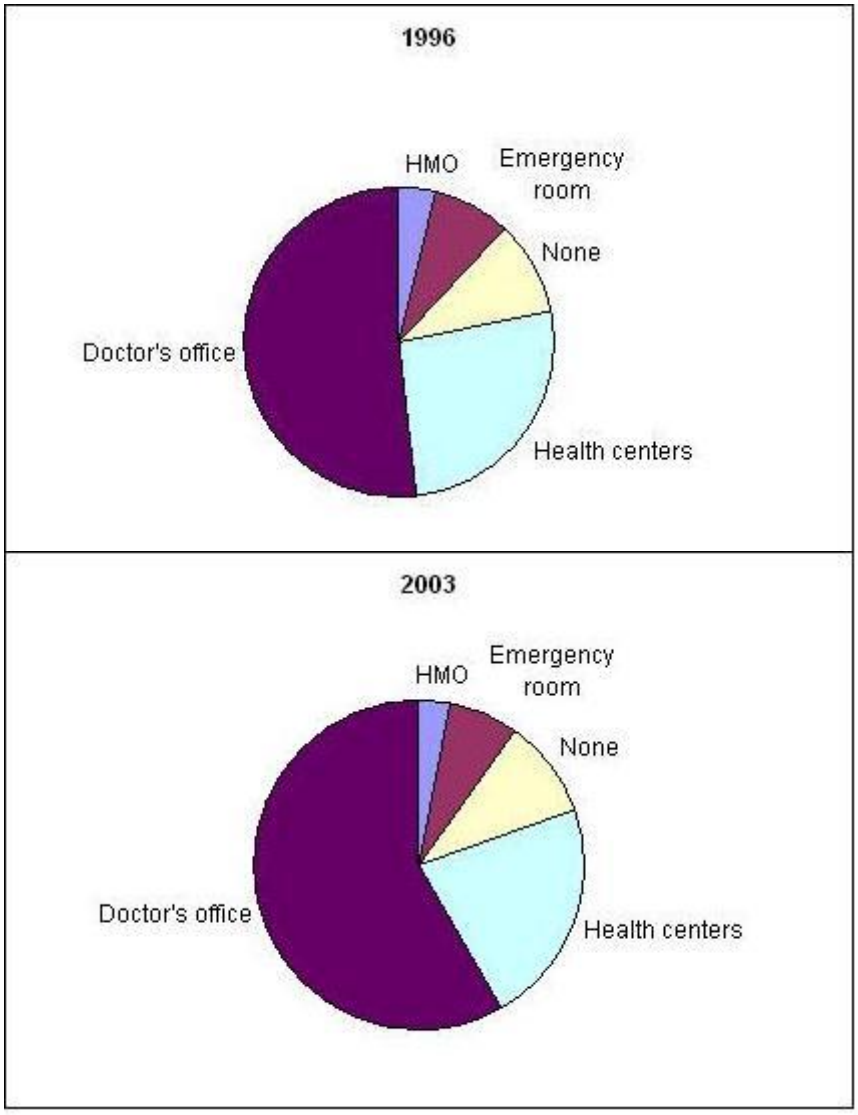

Figure 3 : Study 1: Display type 2D piecharts

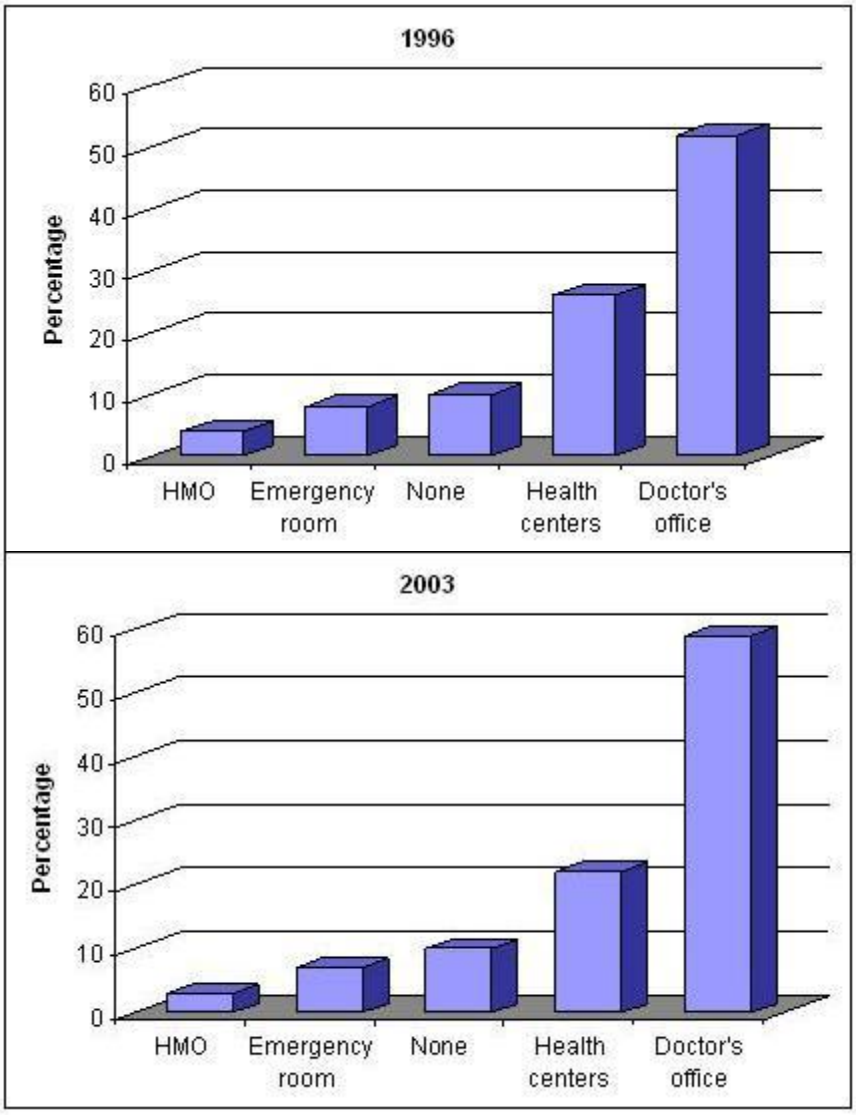

Figure 4: Study 1: Display type 3D bar charts

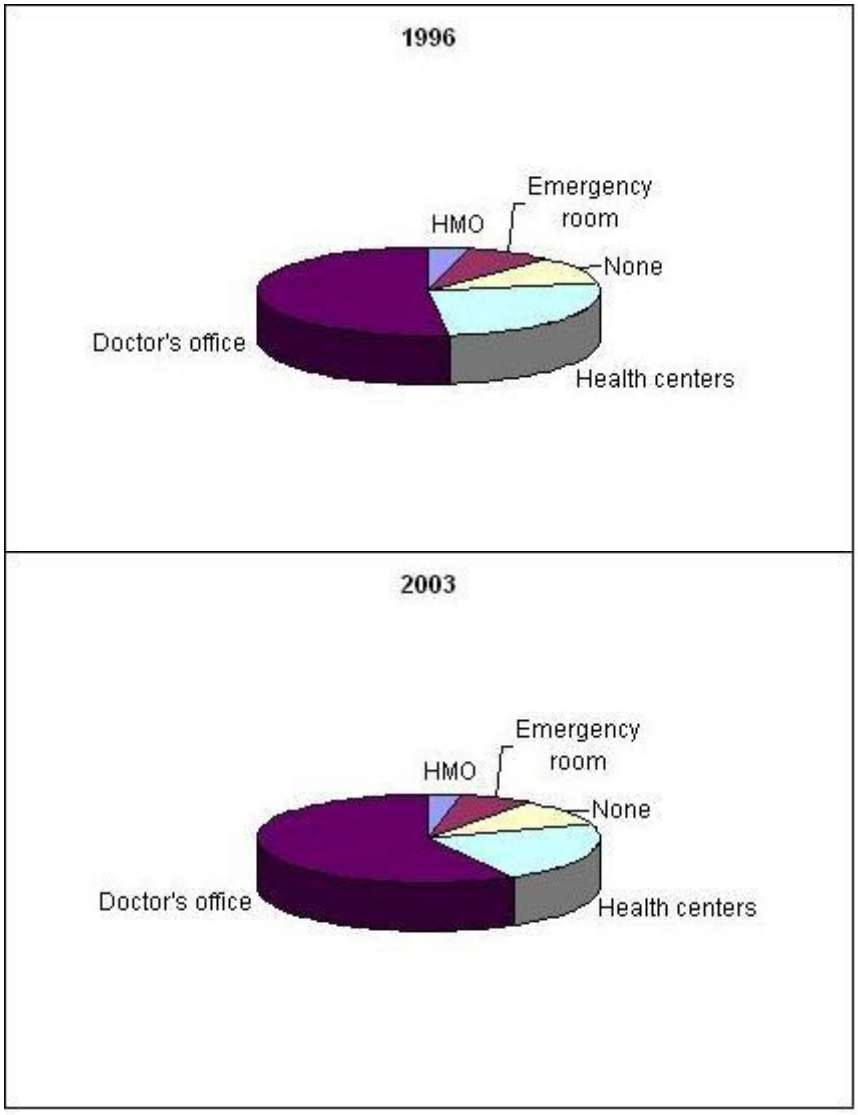

Figure 5: Study 1: Display type 3D pie charts, rotation: large slice left

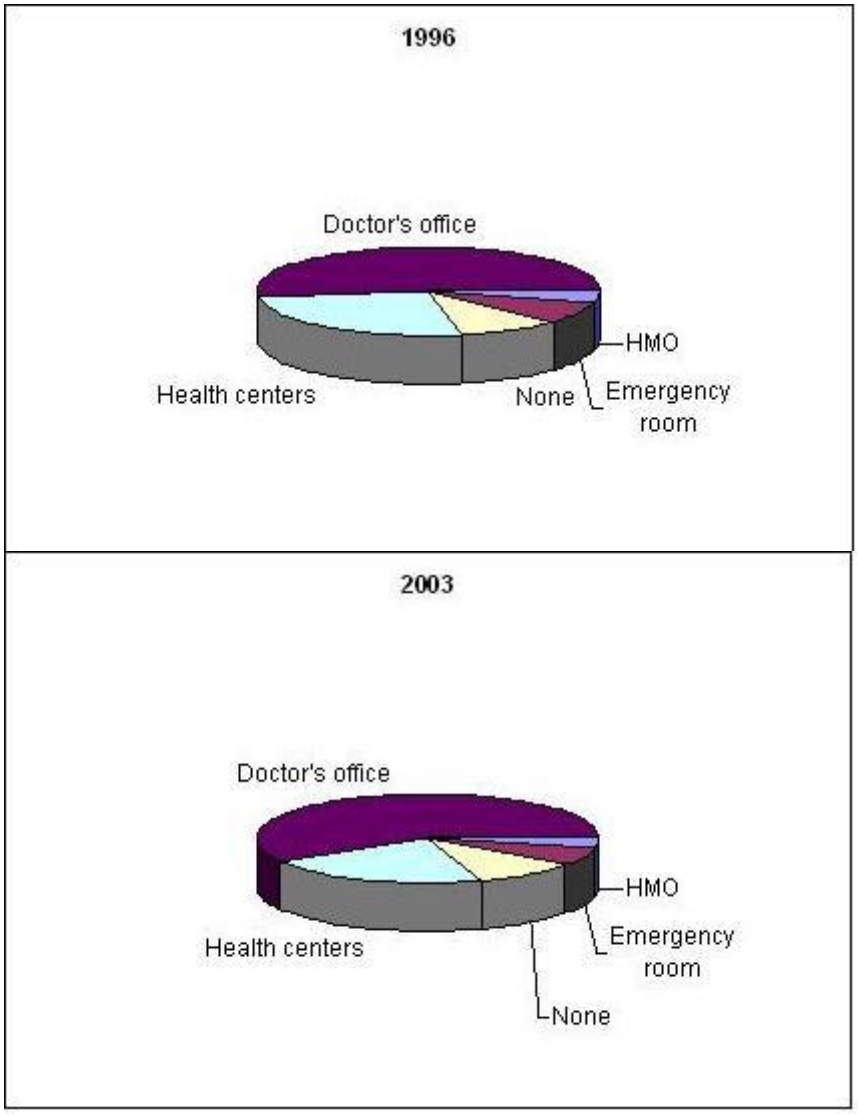

Figure 6: Study 1: Display type 3D pie charts, rotation: large slice back

Because each respondent saw a random display, some displays were seen less often by chance. Each random display was seen at least 266 times in either font size. Displays were shown in color, and the size of the display could not be adjusted. All respondents were asked the same 7 comprehension questions (hereafter called tasks; Table 1) in the same order. The comprehension questions varied and included percentage estimates of one category, estimated percentage differences between categories, within graph (or table) comparisons, and between graph comparisons. The questions were chosen to reflect a variety of different question types rather than to test understanding of an important message of any particular graph. Overall comprehension scores were computed as the sum of correct responses in each experiment separately. Missing answers were counted as incorrect answers.

Se-quence	**Question Text**	**Percen-tage answered correctly**
Q4	In 1996, what was the approximate difference in percentage between patients who answered "None" vs "Health centers". (integer)	52%
Q3	In 1996, the percentage of patients in the emergency room with usual care in a doctor's office was approximately twice as large as the one with usual care in health centers. (True/False)	66%
Q2	In 1996, the number of patients in the emergency room whose usual care was "doctor's office" was larger than all other categories combined. (True/False)	71%
Q7	Which category of usual care has the largest change between 1996 and 2003? (1 HMO, 2 Emergency room, 3 None, 4 Health centers, 5 Doctor's office)	77%
Q1	What is the percentage of patients who had no usual medical care in 1996? (integer)	79%
Q5	The number of emergency room visits in 1996 and 2003 was approximately the same. (True/False)	82%
Q6	The percentage of patients with usual care in a doctor's office increased from 1996 to 2003. (True/False)	91%

Table 1: Study 1: Questions and percentage of correct answers.

Two of the seven comprehension questions required estimating percentages. Estimated percentages based on graphs were counted as correct if they were within 6 percentage points of the correct percentage and incorrect otherwise. Communication of approximate percentages can be more relevant than communicating exact percentages. More generally, gist understanding (a general impression of the data) is often thought to be more important for judgment and decision making than verbatim (specific numerical) knowledge (Hawley, et al. 2008, Reyna 2008). For example, the correct answer in Question Q4 was 16% and we scored estimates from 8% through 20% as correct. (We chose 6 rather than 5 percentage points because respondents' answers accumulate around multiples of 5 or 10. Because the correct verbatim answer is 16%, we did not want to score 20% as correct and 10% as incorrect for approximate comprehension. We analyze the sensitivity of this decision below). For tables, respondents did not need to estimate percentages as the tables contained the correct percentages. Therefore, for tables only, the exact percentage was scored as correct.

Rather than dichotomizing answers to comprehension questions into correct and incorrect, we also considered analyses using continuous variables. This turned out not to be helpful because regression assumptions were far from being met even when allowing for transformations of the response. Answers to most comprehension questions naturally categorized into correct and incorrect; therefore, we decided to dichotomize the 2 continuous questions in study 1 and one such question in study 2 also.

We first compared average overall scores for each experimental arm. For each task, we then used logistic regression to regress comprehension on indicator variables for graph type, and whether the display was 3 dimensional. Differences between any two conditions were given by the corresponding indicator variable (relative to the default) or by contrasts between any two indicator variables. Comparison graph types were table, pie chart, and bar chart in the first experiment and table, divided bar chart, and clustered bar chart in the second experiment. In both experiments, we also explored interactions between 3D format and graph type.

2.2 Participants

Of 2725 invited panel members, 2257 participated in the survey and answered at least one comprehension question. This corresponds to a response rate of 81.8%. Of these, 2197 respondents (96.5%) completed all comprehension questions, and 80 respondents (3.5%) answered some questions but did not answer others.

Respondents represented a diverse cross section of the US population, though some socio-economic strata were overrepresented. The sample contained a large proportion of females (57.7%), and it was older (average age 50.3 years, 1st quartile 39 years, median 52 years, 3rd quartile 61 years), more white (88.0% non-

Hispanic white, 5.8% non-Hispanic African American, 3.8% Hispanic, 2.4% other non-Hispanic), more educated (2.8% no high school degree, 15.6% high school, 37.7% some college, 24.8% 4 year bachelor degree, 19.1% more than a 4yr bachelor degree), and had a higher income (median family income bracket $50,000-$50,999) than one would expect under a strictly proportional representation. The sample contained respondents from 49 states (there were no respondents from Hawaii) as well as Washington D.C. Over-representation of groups mostly reflects panel composition rather than selective non-response.

2.3 Overall comprehension

The mean and median comprehension scores for Study 1 were 5.2 (sd=1.4) and 5, respectively. Scores covered the full range from 0 to 7. Overall comprehension scores by display format for study 1 are shown in Figure 7. They suggest, overall,

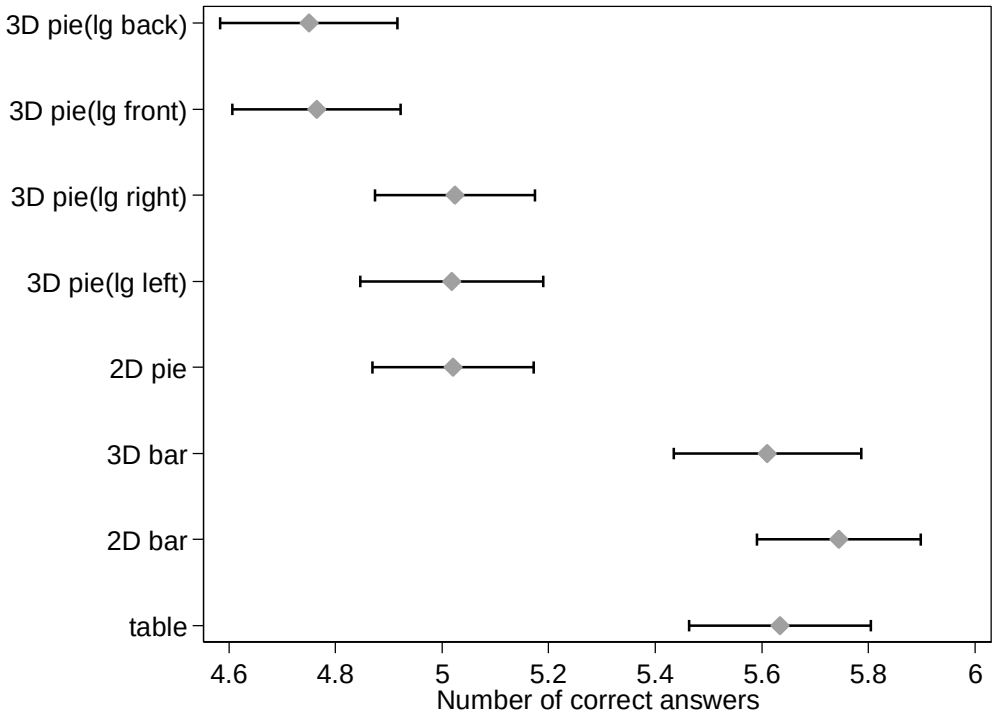

Figure 7: Study 1: Average number of correct answers (out of 7) by graph format. Four different 90 degree rotations of the 3D pie charts are labeled for the position of the largest slice (large slice front, left, back, and right). 95% confidence intervals are shown.

that bar charts and tables led to substantially better comprehension than pie charts. Comprehension varied with the four rotations of the 3D pie charts. The gratuitous third dimension did not affect comprehension of the bar charts, and had a modest effect for the pie charts. This is now discussed in greater detail for individual tasks.

2.4 Comprehension of individual tasks

When conducting multiple regressions with the same set of independent covariates, it is sometimes easier to spot patterns when displaying the results in graphical form (Gelman, Pasarica and Dodhia 2002). Figure 8 shows the coefficients (log odds ratios) of logistic regressions of correct responses based on the display format for each of the 7 questions or tasks in Study 1. The tasks are sorted by question difficulty (% correct answers). If the confidence interval does not touch the horizontal line (coefficient=0) the coefficient is significantly different from zero ($p<0.05$).

The tasks involved estimating percentages (Q1) or differences of percentages (Q4), locating the category with the maximum difference (Q7), and making comparisons between two categories (Q5, Q7, Q8) or one category and a constant (Q2). The comparison could be within the same graph (e.g. comparing two slices of a pie chart) or across the two graphs (e.g. comparing the same slice across two different pie charts representing the two different years). All comparisons required dichotomous yes/no answers.

For questions that required estimating a number (Q1 and Q4) rather than a yes/no decision, the log odds ratio of the pie chart was significantly lower than that of the table or the bar chart. For judging approximate equality of two categories (Q5), the pie chart also had a significantly lower coefficient than the table; however, this coefficient was not very large and was canceled out for 3D pie displays. No significant differences in comprehension emerged between pie charts and tables or barcharts and tables for a range of comparison tasks (Q2, Q3, and Q6).

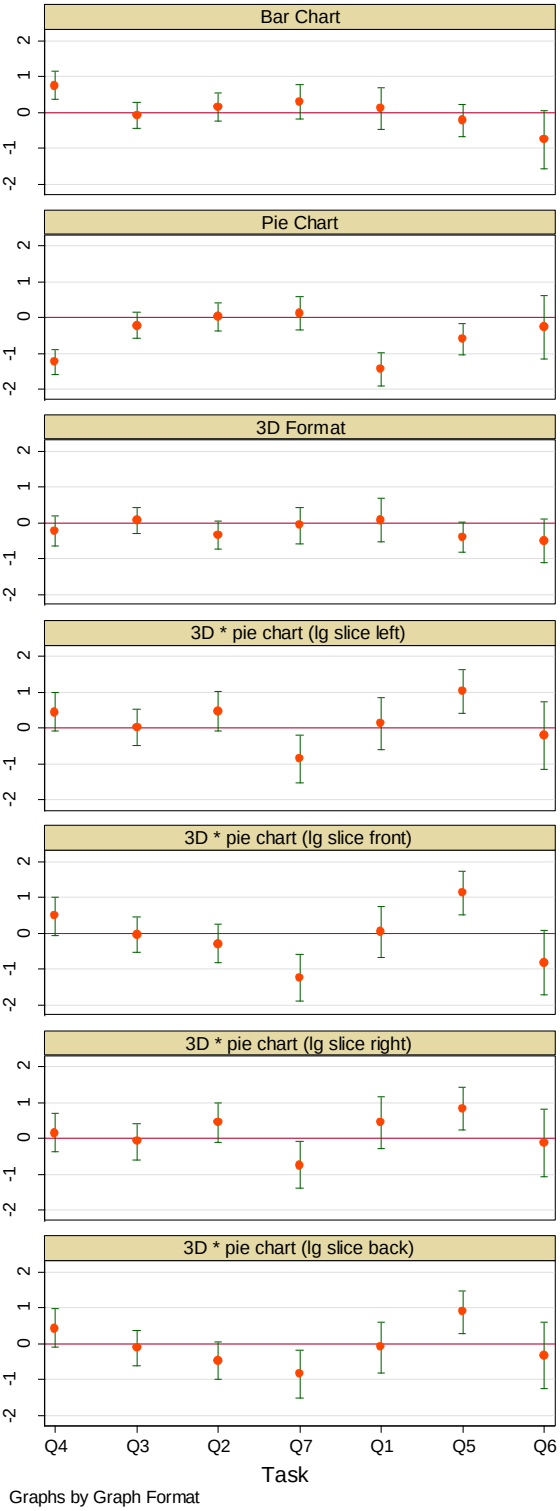

Figure 8: Study 1: Coefficients and 95% confidence intervals from 7 logistic regressions of correct responses on display format for all 7 comprehension questions (tasks). Tasks are sorted by question difficulty (% incorrect answers) starting with the most difficult question. The reference category is the table.

Because the table did not appear in 3D format, the main effect of "3D format" and the corresponding interactions in Figure 8 refer to the bar charts and pie charts only. 3D format was used for bar charts and for the 4 different rotations of the pie chart. The interaction "3D format*bar chart" was chosen as the reference category for the five different 3D displays. There was only one main effect for pie charts; there were no rotation-specific main effects of "pie chart" because only the 3D pie charts were rotated, not the 2D pie chart.

3D format was not a significant predictor of comprehension despite resistance in the literature to its use, but the interactions between 3D format and the rotations of the pie charts were sometimes significant (Q5, Q7). Interaction coefficients were positive for Q5; however, the combined main effects of "3D format" plus "pie chart" cancelled out the interaction effect.

For the most difficult question, Q4, we found that approximate comprehension was significantly better based on bar charts as compared to tables. This question required estimating a percentage. The correct answer was 16% and because we allowed estimates for approximate comprehension to differ by 6%, we scored answers of 10% as correct. We found that this result was sensitive to the threshold of 6%. For thresholds of 6% and larger, we found that approximate comprehension was significantly better for bar charts as compared to tables. For thresholds of 4% and 5% the difference was not significant. For thresholds of 3% or lower, the relationship reverses and comprehension based on bar charts was significantly worse as compared to tables. (Comprehension based on pie charts remained significantly worse than that for tables regardless of the threshold used.) This result confirms that, for exact or nearly exact comprehension, tables outperformed bar charts, whereas for approximate comprehension the reverse can happen.

2.5 Study 1: Discussion

For estimation of absolute proportions (Q1), we found that approximate comprehension based on bar charts and tables was better than that of pie charts. This is consistent with Cleveland's theory (Cleveland, et al. 1984, 1985) stating that comparison involving angles is more difficult than comparison along scales. It is also consistent with experimental work with the use of scales (Hollands, et al. 1998) but inconsistent with another study (Simkin, et al. 1987) that found that comprehension based on bar charts and pie charts was approximately equal.

For judging which of two categories is greater (Q6), we found similar approximate comprehension for tables, bar charts and pie charts. This is consistent with one study displaying only one variable (Spence, et al. 1991). Another study suggested comprehension based on pie charts is worse (Feldman-Stewart, et al. 2000) than that based on either bar charts or displaying numbers; however, this study included comparing some relatively small differences and was based on a single variable with two categories.

Moreover, for judging whether two categories are equal (Q5), we also found that comprehension based on pie charts was worse. Question Q6 involved a comparison of two categories representing 52% and 59%, respectively. Therefore, for judging which of two slices are greater, we believe that approximate comprehension is similar using a table, bar chart or pie chart. Comprehension based on the pie chart becomes gradually worse when the comparison involves categories of nearly equal size.

For estimating differences, we found that approximate comprehension was best for bar charts, intermediate for tables, and worst for pie charts. However, the slight advantage of bar charts over tables reversed when an exact answer was required. For estimating differences, one study (Feldman-Stewart, et al. 2000) was inconclusive: a bar chart was more accurate than a pie chart in a population of cancer patients but not in a population of students.

For a simple numerical operation (Q3) and a task requiring an interpretation of the statement "larger than all other categories combined" (Q2), no differences in comprehension were found. To the best of our knowledge, analyses of these tasks have not yet been studied in the literature.

For identifying the largest difference, we found that approximate comprehension based on 3D pie charts in all rotations was worse than that of a 2D pie chart. To the best of our knowledge, an analysis of this task has not yet appeared.

Lastly, 3D format had no effect on comprehension for bar charts. This is consistent with the experimental literature for estimating absolute proportions (Spence 1990). Another study was inconclusive about the effect of 3D format for bar charts (Siegrist 1996); a third study found 3D format reduced accuracy (Zacks, et al. 1998) but did not evaluate approximate comprehension. Comprehension was reduced for some rotations of the 3D pie chart. The finding that rotations of the 3D pie chart matter is consistent with (Rangecroft 2003).

3. Study 2

Whereas most similar published studies have examined one or two variables, study 2 introduces a third variable. Its introduction increases complexity.

3.1 Experimental Design

The second experiment was conducted in the same web survey. In the second experiment, data about the prevalence of two unspecified diseases were shown by race and age group (children vs. adults). The experiment was introduced with the following sentence: "The graphs [or table] represent the number of children and adults out of every 1000 children and adults who have either Disease 1 or Disease 2. Please look at the graphs below and use them to answer the following questions." The data display was randomized into 5 arms: a 3-way table (Figure

9), a 2D clustered bar chart for children and adults (Figure 10), a 2D divided bar chart for children and adults (Figure 11), and corresponding 3D versions of the graphs (Figure 12 and Figure 13). As in the first study, each display consisted of a 9pt and a 12pt version. Each respondent saw only one of the 5*2 displays. We found no effect of font size on comprehension and therefore do not discuss font size further. Each random display was seen at least 440 times. All respondents were asked the same 8 comprehension questions (Table 2) in the same order with a wide range of tasks. The analysis was analogous to that of the first study.

	Children		Adults	
Group	**Disease 1**	**Disease 2**	**Disease 1**	**Disease 2**
all	18	0.5	21	8
White	24	0	25	4
Black	15	2	17	21
Hispanic	14	1	16	10
Asian	8	2	7	17
Am. Indian	5	0.5	5	35

Figure 9: Study 2: Display type table

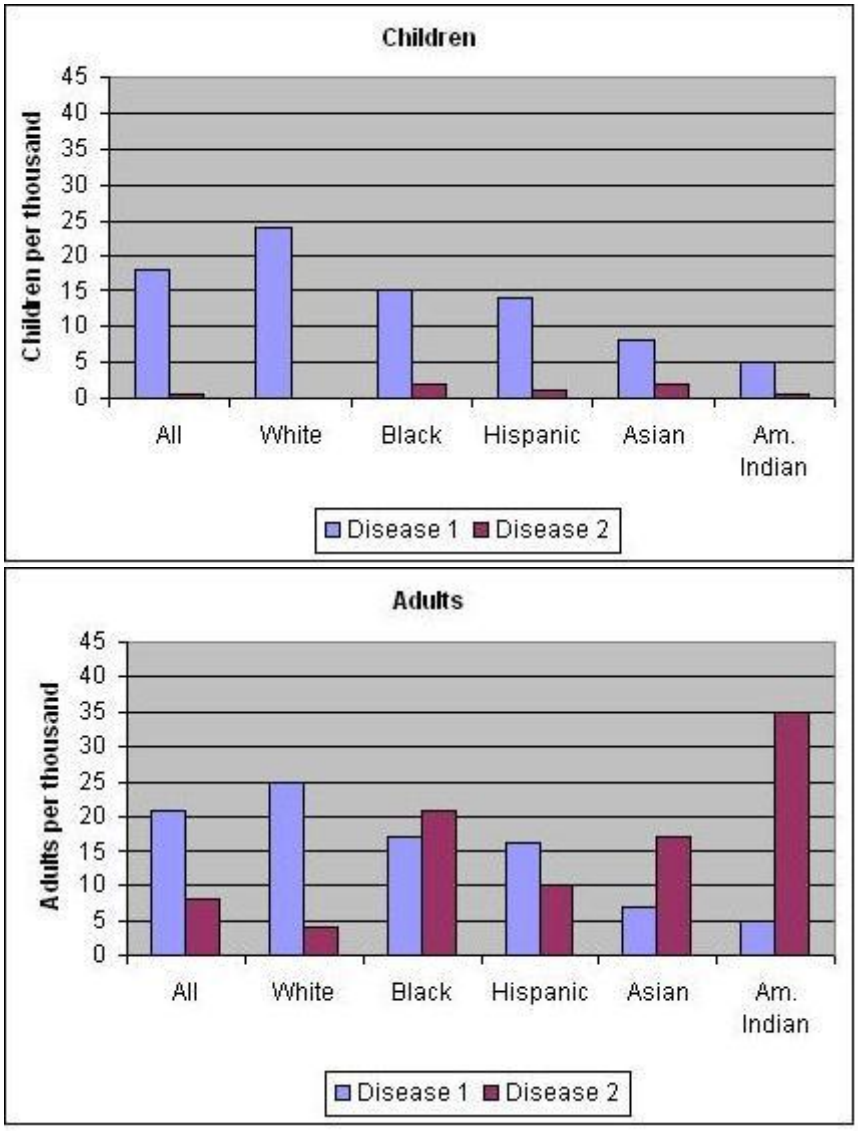

Figure 10: Study 2: Display type clustered bar charts

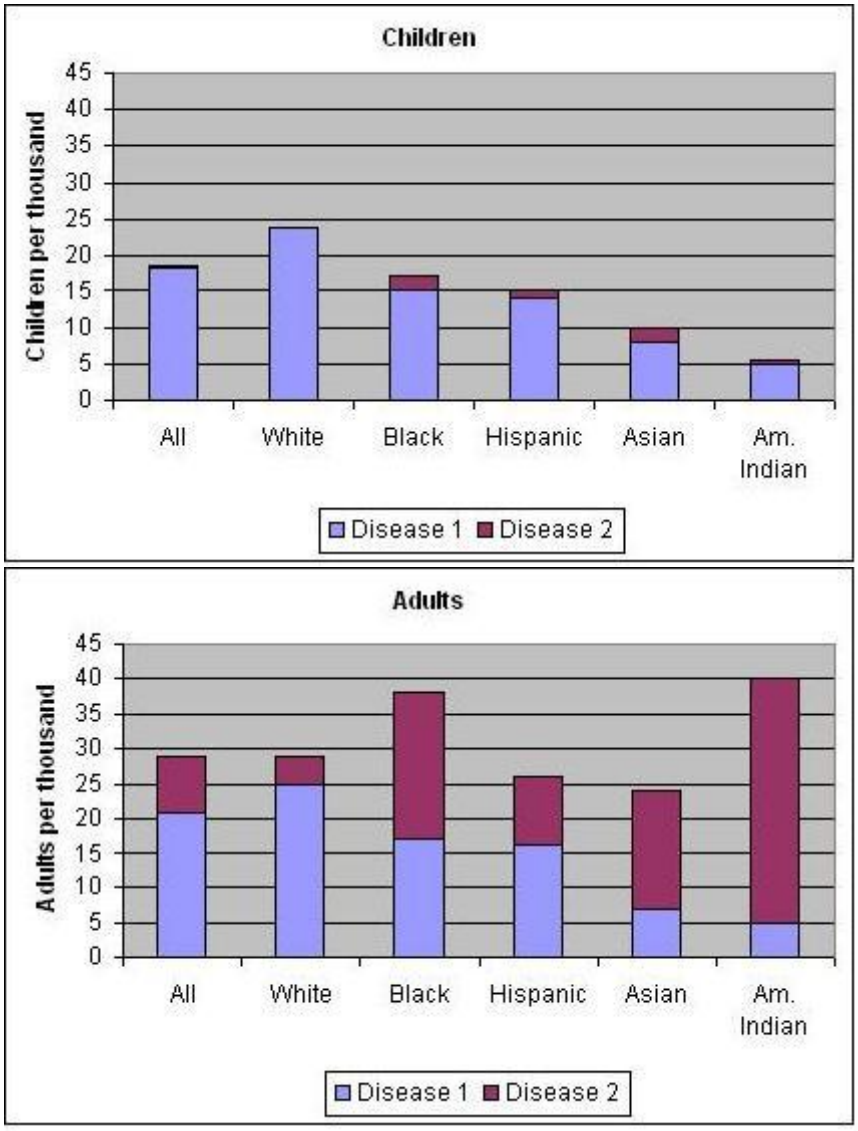

Figure 11: Study 2: Display type divided bar charts

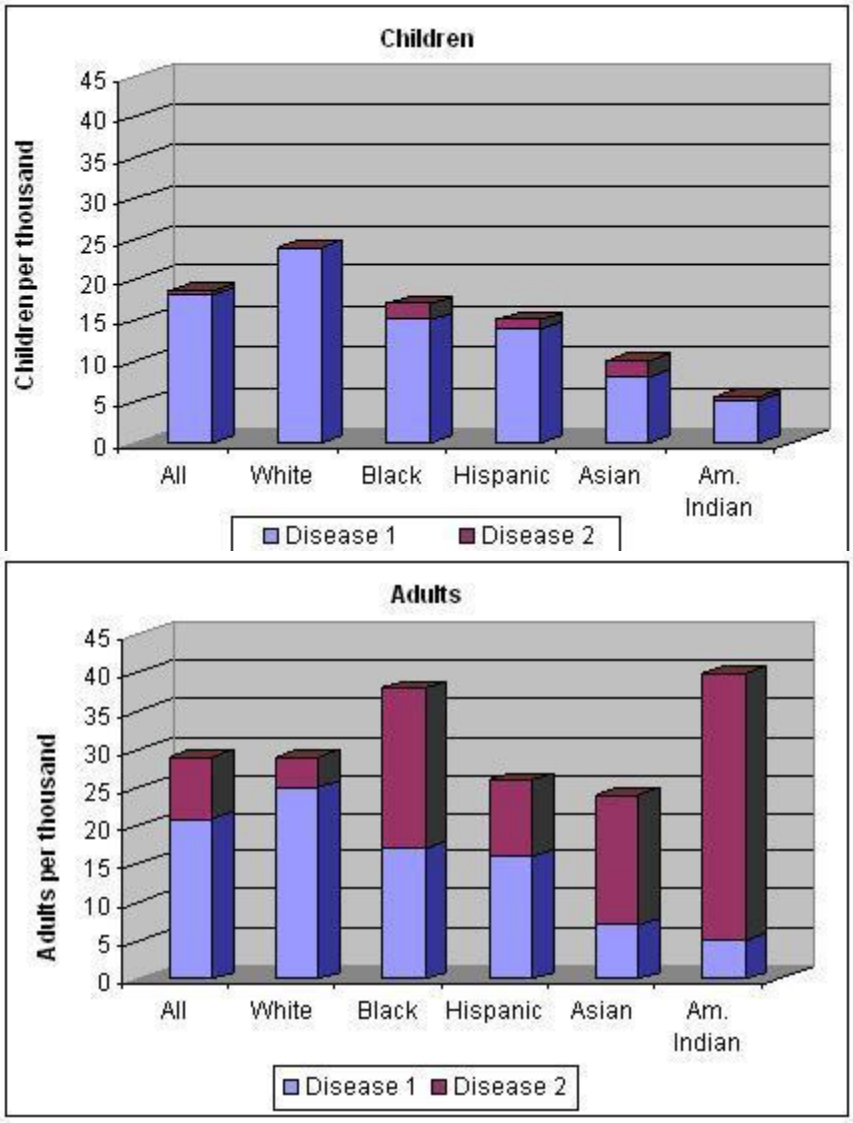

Figure 12: Study 2: Display type 3D divided bar charts

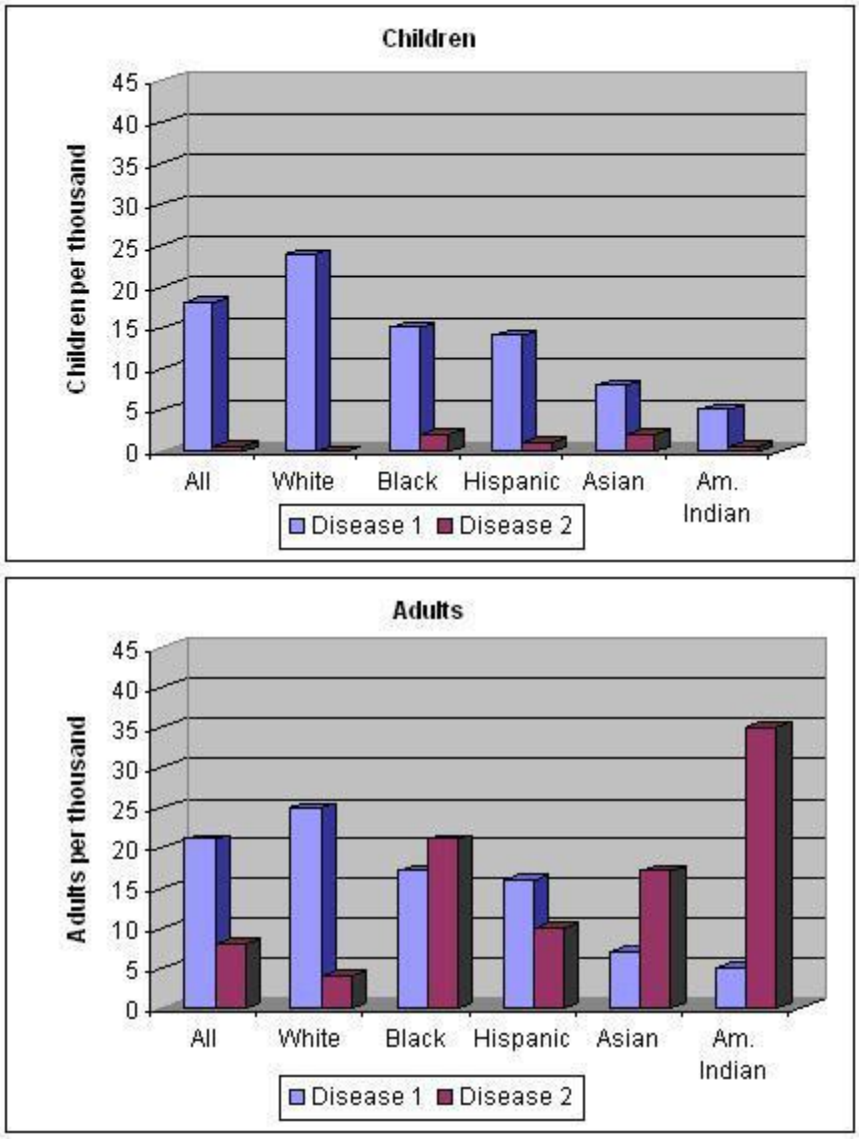

Figure 13: Study 2: Display type 3D clustered bar charts

Se-quence	Question Text	Percen-tage answered correctly
Q8	What children or adult group has the largest number of people (out of every 1000) with Disease 1? (1 White Children, 2 Black Children, 3 Hispanic Children, 4 Asian Children, 5 American Indian Children, 6 White Adults, 7 Black Adults, 8 Hispanic Adults, 9 Asian Adults, 10 American Indian Adults)	71%
Q7	What is the number of Hispanic children (out of every 1000 children) with either Disease 1 or 2? (Integer)	72%
Q6	For American Indians, Disease 1 is equally common among adults and children. (True/False)	78%
Q3	Among children, Disease 1 is more common in Asian children than in Black children. (True/False)	88%
Q4	Disease 2 is more common among adults than among children. (True/False)	89%
Q2	Among adults, Disease 1 is most common in American Indians. (True/False)	92%
Q5	Disease 1 more common for American Indian adults and children. (True/False)	92%
Q1	Among children, Disease 1 is more common than Disease 2. (True/False)	95%

Table 2: Study 2: Questions and percentages of correct answers.

3.2 Participants

The participants are the same as those in study 1.

3.3 Study 2: Overall Comprehension

The mean and median comprehension scores for Experiment 2 were 6.7 (sd=1.5) and 7, respectively. Scores covered the full range from 0 to 8. Overall comprehension scores by display format for study 2 are shown in Figure 14. Tables achieved the highest overall average comprehension scores followed by clustered bar charts and divided bar charts. 3D format seemed to have little effect.

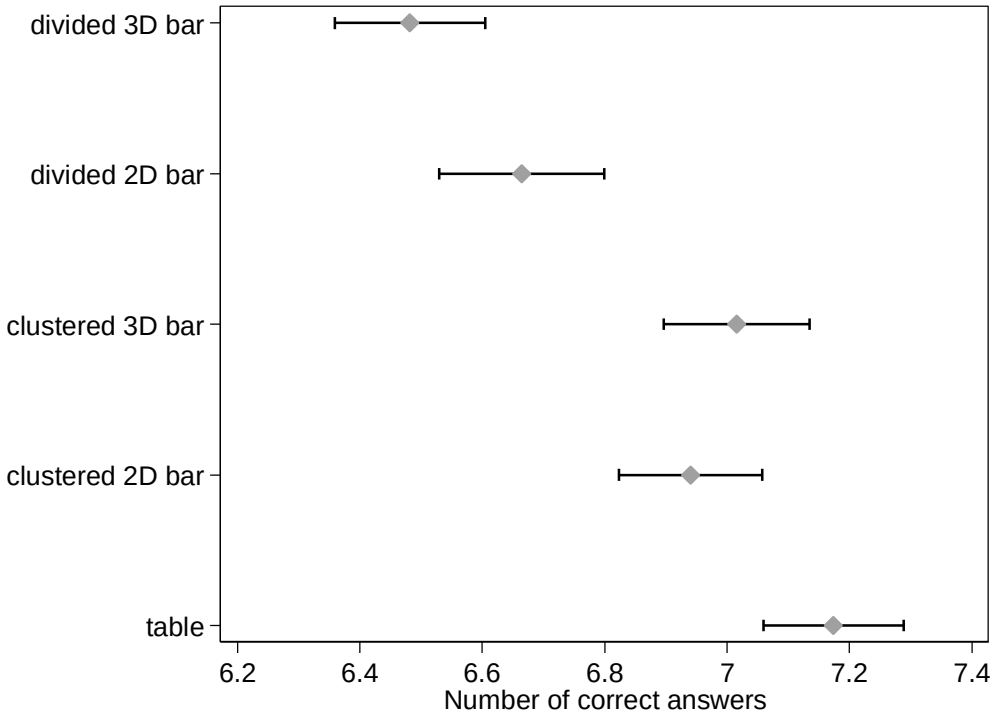

Figure 14: Study 2: Average number of correct answers (out of 8) by graph format. "bar" refers to "bar chart". 95% confidence intervals are shown.

3.4 Study 2: Comprehension of individual tasks

Figure 15 shows coefficients from logistic regressions of individual questions based on display format for Study 2. As before, individual comparisons are based on the significance of indicator variables or differences of indicator variables. Columns are sorted again by the percentage of correct answers. The task with the fewest correct answers, Q8, was difficult because the second largest category was only one percentage point smaller than the largest category (25% vs. 24%).

No display differences were found for comprehension for the easiest two tasks (Q1, Q5). Comprehension based on the table and the clustered bar chart was similar except for judging approximate equality (Q6) and to a lesser extent for estimating a sum (Q7) where comprehension based on tables was better. Comprehension was also significantly better based on the table for Q7 but the coefficient was small.

Comprehension based on the table was significantly better than the divided bar chart for all but the two easiest questions (Q1, Q5). Likewise, comprehension based on the clustered bar chart was usually better than that based on divided bar charts: contrasts (not shown) between the relevant coefficients in Figure 15 showed significant ($p<0.05$) differences for all questions except for the easiest questions (Q1, Q5) and question Q7. In addition, the negative interactions imply that comprehension of the 3D format for divided bar charts were even worse for two of the tasks.

The 3D display format had modest coefficients; comprehension of this format tended not to differ from the other formats. For two tasks (Q2, Q6 for clustered bar charts) comprehension was better for 3D format and, for two other questions, it was worse (Q8 and Q6 for divided bar charts).

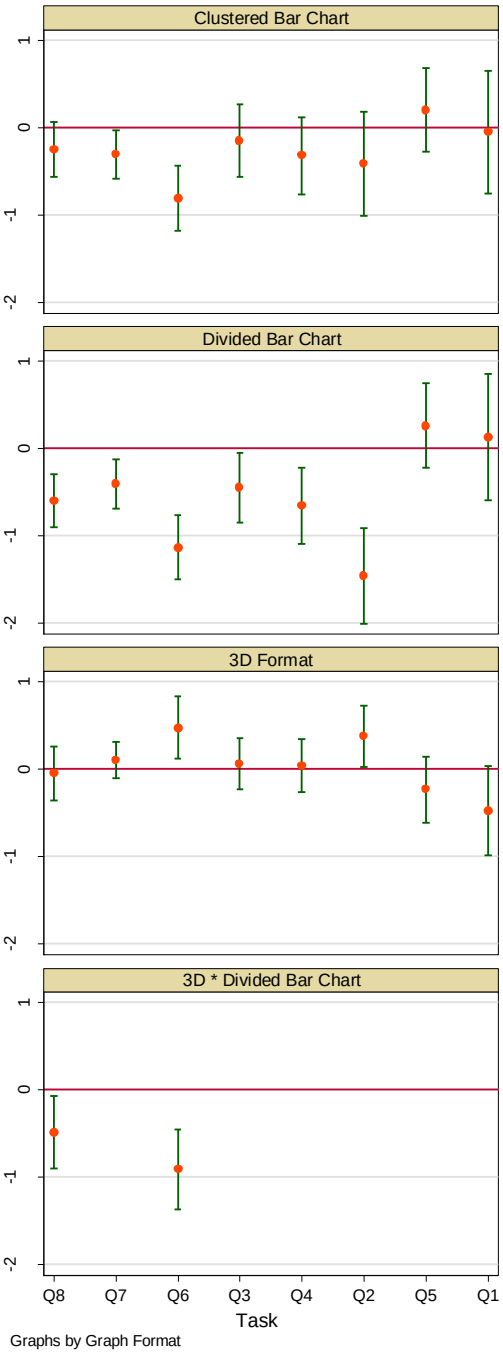

Figure 15: Study 2: Coefficients and 95% confidence intervals from 8 logistic regressions of correct responses based on display format for all 8 comprehension questions (tasks). Tasks are sorted by question difficulty (% incorrect answers) starting with the most difficult question. The reference category is the table. The interaction "3D* Divided Bar chart" was dropped from models where it was insignificant.

3.5 Study 2: Discussion

Comprehension of such displays for three variables has not yet been evaluated in the literature.[2] Our findings on three variables on a larger variety of tasks were largely consistent with those for one and two variables: For comparing the relative size of two categories, divided bar charts performed poorly as compared to clustered bar charts (Cleveland, et al. 1984). For estimating absolute proportions, comprehension based on a single divided bar was worse than that based on a single bar chart (Simkin, et al. 1987).

Comprehension of clustered and divided bar charts has also not been compared with comprehension based on tables in the literature. However, using a single display variable, tables have been reported to be most accurate with bar charts a close second (Spence 1990). We found that the 3-way table was similar in approximate comprehension to the clustered bar chart for most tasks. Therefore, increasing the number of variables to be displayed from one to three does not fundamentally shift comprehension of a table as compared to a (clustered) bar graph.

3D clustered and divided bar charts have also not yet been evaluated in the literature. We found that the 3D display had little effect on approximate comprehension.

4. Overall Discussion

We can comment on three general questions based on the present data. First, are visual displays inevitably better than tables? The answer is no. Graphs were better for estimating differences; however, tables were better when estimating equality and sums. Second, do 3D displays compromise comprehension? For bar charts, 3D displays appear to work as well as 2D displays, but they do compromise comprehension of pie charts. Finally, do pie charts worsen data understanding? In general, although pie charts never assisted comprehension, they impaired comprehension on fewer than half of the tasks.

In terms of the more detailed results, we conclude, first, that approximate comprehension for (clustered) bar charts and tables was comparable whether displaying data from two-way or three-way tables. This was not obvious as much of the literature focuses on the display of a single variable. In study 1, bar charts and tables led to equally good overall comprehension with one exception: Estimating the difference between two categories (Q4) was easier with bar charts than with tables. Calculating a precise difference as required in the table may be

---

[2] For pictographs, at least one study displays 3 variables: for each of 2*2 treatment combinations a pictograph is displayed related to the probability of survival.(Zikmund-Fisher, B. J., Angott, A. M., and Ubel, P. A. (2011), "The Benefits of Discussing Adjuvant Therapies One at a Time Instead of All at Once," *Breast cancer research and treatment*, 129, 79-87.)

harder than visually estimating the length difference between two neighboring categories. For judging equality on the other hand, comprehension was somewhat better for tables than clustered bar charts (Study 2, Q6). Another estimation task, Q7 (estimate a sum), also demonstrated a small superiority for tables over clustered and divided bar charts although the importance of such small effects may be questionable.

Two, comprehension for pie charts is task dependent. When estimating a percentage or a difference of percentages (even when allowing for rounding) and when testing for equality, comprehension based on pie charts was worse than that of a table. However, for judging which of two categories was larger, comprehension for pie charts and tables was comparable. Estimating even approximate proportions and differences of proportions may be a more complex task than deciding which of two slices is larger.

Three, divided bar charts should be avoided. Comprehension based on divided bar charts was worse for all but the easiest tasks. This is consistent with Cleveland's mantra that comparisons along a common scale are easier than length comparisons and is consistent with older adults' difficulties with similar divided bar charts often used to display quality-of-care data (Hibbard, Slovic, Peters and Finucane 2002).

Four, 3D format had no effect on comprehension for bar charts, but comprehension was reduced for some rotations of the 3D pie chart.

Clearly, some tasks were more difficult than others. In both experiments, the judgment of which of two numbers was greater (or smaller) had the largest percentage of correct answers. There were more correct answers in study 2 in part because there were more such questions in study 2. Judgment of approximate equality varied considerably by display format in study 1 (Q5) and 2 (Q6). Further research on task difficulty is needed.

This study is not without limitations. Because of practicalities with the web survey implementation, data formats were varied in each study but only a single set of numbers was used in each one. On the other hand, implementing the survey as a web survey enabled access to a large diverse population and testing of a larger variety of tasks; thereby substantially increasing external validity. Second, in Study 2 almost 7 out of 8 questions were answered correctly on average. This indicates a possible ceiling effect which might have disguised additional differences in performance. We also did not compare the present formats to pictographs, which have been found useful in risk communication (Hawley et al., 2008). We also note that the study was conducted on the web and was not paper-based.

Survey respondents were generally very interested in graphs. Nearly 30% of respondents left mostly positive comments in an open ended field. Some respondents volunteered advice such as "Please look at Edward Tufte's `The Visual Display of Quantitative Information' (1983) to learn how to display

quantitative information. Microsoft Excel's graphs are the worst." All comments related to 3D graphs were negative, for example: "I hope [the survey] will show that 3D graphs, while appealing are more difficult to read than 2D graphs." Such comments demonstrate that people care about graphs, presumably because graphs are, or could be, useful for them.

The display of graphs and tables matters in particular in the context of how they influence respondents' risk perceptions and decisions and in health literacy where graph comprehension is part of document literacy (Lipkus and Hollands 1999, Fagerlin, Ubel, Smith and Zikmund-Fisher 2007, Lipkus 2007, Peters, Dieckmann, Dixon, Hibbard and Mertz 2007, Peters, Hibbard, Slovic and Dieckmann 2007, Peters, et al. 2009). In both areas, the ability of people with lower education, lower numeracy, and older age to comprehend and use numeric information is particularly important and requires further research.

Making these kinds of choices about how to present information can be considered examples of "choice architecture," a term coined by Thaler and Sunstein (Thaler and Sunstein 2008) to reflect the notion that many choices depend upon how they are presented. The analogy is to the architect of a building who determines how people move throughout the building through his or her choices of where to place doors, hallways, offices, and bathrooms. Architects of choice can influence comprehension as we have focused in the present paper, but they also can influence what is chosen. Policy makers cannot present choices in a 'neutral' architecture, because all ways of presenting information ultimately influence the decision maker in some way. Thus, policy makers are already choice architects, with all of the ethical complications involved.

Let us go back to the original article in the New York Times that motivated the first experiment. Rather than displaying the divided bar chart shown in the original article (Weber, et al. 2008), the New York Times chose to display one pie chart for each of two years (the labels also contained the numerical percentage values for each slice). Was that a good choice? Pie charts are an improvement over divided bar charts which are the least desirable option (Figures 14 and 15). However, Figure 8 implies that, for several tasks, pie charts are not the best choice for approximate comprehension; a bar chart or a table might have been even better.

References

Ancker, J. S., Senathirajah, Y., Kukafka, R., and Starren, J. B. (2006), "Design Features of Graphs in Health Risk Communication: A Systematic Review," *Journal of the American Medical Informatics Association*, 13, 608-618.

Best, L., Smith, L., and Stubbs, D. (2001), "Graph Use in Psychology and Other Sciences," *Behavioural processes*, 54, 155-165.

Cleveland, W. (1984), "Graphs in Scientific Publications," *American Statistician*, 38, 261-269.

Cleveland, W., and McGill, R. (1984), "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, 79, 531-554.

Cleveland, W., and McGill, R. (1985), "Graphical Perception and Graphical Methods for Analyzing Scientific Data," *Science*, 229, 828-833.

Cleveland, W., and McGill, R. (1987), "Graphical Perception: The Visual Decoding of Quantitative Information on Graphical Displays of Data," *Journal of the Royal Statistical Society. Series A (General)*, 150, 192-229.

Fagerlin, A., Ubel, P., Smith, D., and Zikmund-Fisher, B. (2007), "Making Numbers Matter: Present and Future Research in Risk Communication," *American Journal of Health Behavior*, 31, S47-S56.

Fagerlin, A., Wang, C., and Ubel, P. A. (2005), "Reducing the Influence of Anecdotal Reasoning on People's Health Care Decisions: Is a Picture Worth a Thousand Statistics?," *Medical Decision Making*, 25, 398-405.

Feldman-Stewart, D., Brundage, M., and Zotov, V. (2007), "Further Insight into the Perception of Quantitative Information: Judgments of Gist in Treatment Decisions," *Medical Decision Making*, 27, 34-43.

Feldman-Stewart, D., Kocovski, N., McConnell, B., Brundage, M., and Mackillop, W. (2000), "Perception of Quantitative Information for Treatment Decisions," *Medical Decision Making*, 20, 228-238.

Fischer, M. (2000), "Do Irrelevant Depth Cues Affect the Comprehension of Bar Graphs?," *Applied Cognitive Psychology*, 14, 151-162.

Garcia-Retamero, R., and Galesic, M. (2009), "Communicating Treatment Risk Reduction to People with Low Numeracy Skills: A Cross-Cultural Comparison," *American journal of public health*, 99, 2196-2202.

Gelman, A., Pasarica, C., and Dodhia, R. (2002), "Let's Practice What We Preach," *The American Statistician*, 56, 121-130.

Haemer, K. (1951), "The Pseudo Third Dimension," *American Statistician*, 5, 28-28.

Hawley, S., et al. (2008), "The Impact of the Format of Graphical Presentation on Health-Related Knowledge and Treatment Choices," *Patient Education adn Counseling*, 73, 448-455.

Heer, J., and Bostock, M. (2010), "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design," in *CHI*, Atlanta, Georgia: ACM, pp. 203-212.

Hibbard, J., Slovic, P., Peters, E., and Finucane, M. (2002), "Strategies for Reporting Health Plan Performance Information to Consumers: Evidence from Controlled Studies," *Health Services Research*, 37, 291-313.

Hollands, J., and Spence, I. (1998), "Judging Proportion with Graphs: The Summation Model," *Applied Cognitive Psychology*, 12, 173-190.

Koch-Baumgarten, S., and Voltmer, K. (2010), *Public Policy and Mass Media: The Interplay of Mass Communication and Political Decision Making* (Vol. 66), London: Routledge.

Levy, E., Zacks, J., Tversky, B., and Schiano, D. (1996), "Gratuitous Graphics? Putting Preferences in Perspective," in *SIGCHI conference on Human factors in*

computing systems: common ground, Vancouver, BC: ACM New York, NY, USA, pp. 42-49.

Lipkus, I. (2007), "Numeric, Verbal, and Visual Formats of Conveying Health Risks: Suggested Best Practices and Future Recommendations," *Medical Decision Making*, 27, 696-713.

Lipkus, I., and Hollands, J. (1999), "The Visual Communication of Risk," *Journal of the National Cancer Institute Monographs*, 1999, 149-163.

Meyer, J., Shamo, M. K., and Gopher, D. (1999), "Information Structure and the Relative Efficacy of Tables and Graphs," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 41, 570-587.

Meyer, J., Shinar, D., and Leiser, D. (1997), "Multiple Factors That Determine Performance with Tables and Graphs," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39, 268-286.

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J., and Mertz, C. (2007), "Less Is More in Presenting Quality Information to Consumers," *Medical Care Research and Review*, 64, 169-190.

Peters, E., et al. (2009), "Bringing Meaning to Numbers: The Impact of Evaluative Categories on Decisions," *Journal of Experimental Psychology: Applied*, 15, 213-227.

Peters, E., Hibbard, J., Slovic, P., and Dieckmann, N. (2007), "Numeracy Skill and the Communication, Comprehension, and Use of Risk-Benefit Information," *Health Affairs*, 26, 741-748.

Rangecroft, M. (2003), "As Easy as Pie," *Behaviour and Information Technology*, 22, 421-426.

Reyna, V. (2008), "A Theory of Medical Decision Making and Health: Fuzzy Trace Theory," *Medical Decision Making*, 28, 850-865.

Shah, P., Freedman, E. G., and Vekiri, I. (2005), "The Comprehension of Quantitative Information in Graphical Displays," in *The Cambridge Handbook of Visuospatial Thinking*, eds. P. Shah and A. Miyake, Cambridge: Cambridge University Press.

Siegrist, M. (1996), "The Use or Misuse of Three-Dimensional Graphs to Represent Lower-Dimensional Data," *Behaviour & Information Technology*, 15, 96-100.

Simkin, D., and Hastie, R. (1987), "An Information-Processing Analysis of Graph Perception," *Journal of the American Statistical Association*, 82, 454-465.

Spence, I. (1990), "Visual Psychophysics of Simple Graphical Elements," *Journal of experimental psychology: Human perception and performance*, 16, 683-692.

Spence, I. (2005), "No Humble Pie: The Origins and Usage of a Statistical Chart," *Journal of Educational and Behavioral Statistics*, 30, 353-368.

Spence, I., and Lewandowsky, S. (1991), "Displaying Proportions and Percentages," *Applied Cognitive Psychology*, 5, 61-77.

Stewart, B., Cipolla, J., and Best, L. (2009), "Extraneous Information and Graph Comprehension: Implications for Effective Design Choices," *Campus-Wide Information Systems*, 26, 191-200.

Thaler, R., and Sunstein, C. (2008), *Nudge: Improving Decisions About Health, Wealth, and Happiness*, New Haven: CT: Yale University Press.

Tufte, E. (1983), *The Visual Display of Quantitative Information*, Cheshire, Conn. ,USA: Graphics Press.

Vessey, I. (1991), "Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature," *Decision Sciences*, 22, 219-240.

Weber, E. J., et al. (2008), "Are the Uninsured Responsible for the Increase in Emergency Department Visits in the United States?," *Annals of emergency medicine*, 52, 108-115.

Wilkinson, L. (1994), "Less Is More: Two-and Three-Dimensional Graphics for Data Display," *Behavior Research Methods, Instruments, & Computers*, 26, 172-172.

Wilkinson, L. (2001), "Presentation Graphics," *International Encyclopedia of the Social and Behavioral Sciences*, 9, 6369-6379.

Zacks, J., Levy, E., Tversky, B., and Schiano, D. (1998), "Reading Bar Graphs: Effects of Extraneous Depth Cues and Graphical Context," *Journal of Experimental Psychology: Applied*, 4, 119-138.

Zacks, J., Levy, E., Tversky, B., and Schiano, D. (2001), "Graphs in Print," in *Diagrammatic Representation and Reasoning*, ed. M. A. B. M. P. Olivier, London: Springer-Verlag, pp. 187-206.

Zikmund-Fisher, B. J., Angott, A. M., and Ubel, P. A. (2011), "The Benefits of Discussing Adjuvant Therapies One at a Time Instead of All at Once," *Breast cancer research and treatment*, 129, 79-87.