Recovering data from scanned graphs: Performance of Frantz's g3data software

BEN BAUER
Trent University, Oshawa, Ontario, Canada

AND

MICHAEL REYNOLDS
Trent University, Peterborough, Ontario, Canada

Recovering data points from scanned or photocopied graphs by hand is time consuming, prone to error, and it requires patience and a steady hand. An alternative is to use *g3data*—a program that reads many common image file types and allows accurate estimation of data point coordinates from the graphs contained in these files. Limits of the software are discussed, and validation of its accuracy using graphs with known data values is presented.

As science develops, new ways of looking at experimental and observational data arise. Consequently, the tomes of existing published data become a critical resource for new statistical analyses (e.g., meta-analyses, data mining, computational modeling, etc.). The practicality of such analyses is limited by access to the original data and the ease of getting the data into a form that is usable by statistical software.

Access to the original data from studies presented in journal articles, theses, and magazines is often difficult, and it becomes even more so with time. Moreover, research in many scientific journals spans decades or centuries, and older source data is likely to have been discarded or lost. More recent data may still be available on storage media, such as punch/magnetic tape, floppy discs, CD-ROMs, and so on, but these media may not be reliable over the long term. For instance, devices to restore the data may become rare, and, unfortunately, some of this information is stored in proprietary file formats with no software to access it, even if the medium and device were available (hopefully, the latter problem will become less prevalent now that the International Standards Association has accepted a nonproprietary format ISO/IEC 26300—OpenDocument Format—for electronic files, such as spreadsheets, text documents, and presentations). Consequently, the only traces that may exist for the data from these experiments are the printed pages or microform copies.

Recovering data points from the graphs and plots available in print copies (e.g., from original journal pages, photocopies, prints from optical scans, or microfiche) with ruler in hand and inputting them to data analysis software is problematic because it is labor intensive and of unknown accuracy. Consider a situation in which data points from a graph are recovered manually using a pencil and ruler to acquire the coordinates, which are then typed into a spreadsheet program; as the number of points increases, the probability of making a mistake—either in measurement or transcription—increases. The benefits of using a program such as g3data are amplified when one considers data recovery from a scatterplot containing scores of data points, from a graph containing many levels of many factors, or trying to quantify the magnitude of error bars from a multifactorial graph. The g3data program provides a way to increase efficiency and accuracy in accomplishing these tasks.

An alternative and more economical approach would be to use digital media to recover data. Indeed, the recent availability of online "e-journals" and subscription-based access to optical scans of older materials, as well as the existence of open-access archives—for example, the University of Michigan's Deep Blue (deepblue.lib.umich.edu/index.jsp), MIT's CSAIL Digital Archive (publications.csail.mit.edu), and open-access advocacy projects, such as Scientific Commons (www.scientificcommons.org, which recently announced its intention to support OpenDocument Format)—mean that a mass of electronic copies of scientific works is available. Furthermore, Jones, Warner, and Cross (1998) have estimated that in excess of 2×10^{12} new graphs are published each year. Hopefully, original data—or at least high quality images of these data—will be available.

The use of existing data retrieved from print sources can be an economical supplement to new investigations and can shorten the experiment piloting time. To update Westheimer's Discovery ("A couple of months in the laboratory can frequently save a couple of hours in the library"), a couple of hours recovering extant data might save a couple of weeks in the laboratory. Furthermore, data points from older studies can be recovered and plotted anew for inclusion in presentations and teaching mate-

---

B. Bauer, benbauer@trentu.ca

Table 1
Characteristics of the Six Graphs Examined in Detail

Source	Medium	Image Resolution	File Format	Color Depth	File Size (kB)	Data Scaling
Jevons (1871)	Replot	–	PNG	8-bit color	11	linear–linear
Taves (1941), Figure 15	Photocopy scan	200 dpi	PGM	250 grays	242	linear–linear
test2.png	In g3data	–	PNG	8-bit color	11	log–linear
Kaufman et al. (1949), Figure 7	Microfiche copy scan	–	PGM	238 grays	208	log–log
Taubman (1950), Figure 1	Photocopy scan	240 dpi	PGM	236 grays	593	linear–linear

Note—See Figures 1 and 5 for the reproduction of Jevons (1871) and the test2.png plot. The remaining three are presented in the Appendix.

rials, with the added bonus of plotting data from different studies (perhaps a century apart or even in different measurement units) on the same axes for visual comparison.

Recovery of data coordinates from image files (e.g., scans or original electronic files) using g3data Version 1.5.1 for Linux (Frantz, 2000; www.frantz.fi/software/g3data.php) is documented below. In brief, g3data is a graphical user interface (GUI) program that allows a user to process an image file of a graph by calibrating the axis scales of the plot and then clicking the data points (plotting symbols) to acquire estimates for the coordinates. The recovered points can be written to a text screen or saved in a file for subsequent analysis.

There are obviously many kinds of graphs that could be processed using this software. The goal of the tests presented below was to characterize the performance of g3data as a function of different kinds of axis scaling (e.g., linear–linear, log–log) and image quality (e.g., on the basis of the presence of skew or scanning artifacts) and to document the accuracy of the coordinates obtained with a view to deciding under which circumstances it is suitable for scientific work.

METHOD

Software

g3data (www.frantz.fi/software/g3data.php) is a free program that is available under the GNU General Public License, and it can be used to recover coordinates from graphs in many common image formats (e.g., grayscale or color: .png, .gif, .jpg, netpbm, .tiff, .bmp, .pcx). In theory, this task could be accomplished in an indirect way using many common image manipulation packages. Using the mensuration tools (calipers, rulers) in the package, one could record the screen coordinates (e.g., in pixel or millimeter units) and then convert them to graph coordinates. However, as with the ruler and pencil example considered previously, the labor required may be prohibitive, and the recovered values may be prone to error. Fortunately, g3data is designed to do these operations directly, and it handles linear–linear, semilog, and log–log axes that would otherwise require subsequent conversions prior to use, if done using other methods. There are three steps for obtaining the coordinates: (1) Calibration—indicate whether the axes are linear or log, and define the sizes and scales of the axes, (2) click on the data points to estimate, and (3) save the coordinates.

Clicking on the data points involves some subjective estimation in the placement of the crosshair cursor with respect to the plotting symbol. To help locate the cursor in the center of the plotting symbol, a secondary magnification window (called a *zoom area*) is built into the interface (see Figure 3). This window shows a live zoomed image of a small area containing the cursor. This is very useful when the image contains noise dots introduced in optical scanning and microfiche reproductions. Screenshots of the interface and usage are available on the g3data Web site.

In what follows, the performance of g3data (for Linux) is examined with respect to both the accuracy of the coordinates recovered and the ability of undergraduates to use the program. A set of graphs with known data values was selected to cover a variety of cases that might occur in using second- (or later-) generation copies. The original data (i.e., coordinates) were published in the journal article along with the graphs. The general properties of the image files discussed below are given in Table 1. In total, 16 graphs were processed; however, only five representative cases are presented in detail (some of the graphs processed were from a series in the original articles and were similar).

Apparatus

The g3data (for Linux) software was loaded onto a Dell Optiplex GX240 computer (1.8-GHz Pentium 4) that had a Logitech (M-S35) three-button optical mouse, a standard PC keyboard, and a 32-MB Geforce2 MX200 video card. This hardware configuration is very modest by 2008 standards; however, the interface was perfectly responsive, with no lag or jump. The monitor was a LaCie Electron Blue IV 19-in. CRT (with hood), set to 1,280 × 1,024 pixels at 85 Hz. The computer was running Slackware Linux 12.0.

Procedure

Eight student volunteers (observers) processed the graphs alone in a small classroom. The observers were all female undergraduates, and at least 4 of them had some office computer experience. They were told that they were helping the investigator to obtain data points from the graphs for use in a review article. The context was not described as an experiment testing their abilities or the properties of the software.

Each observer, in turn, sat at a large table with the monitor screen positioned approximately 60 cm from the front of the table, allowing ample room for the keyboard, mouse, and paper copies of the graphs on which they could note any information concerning problems, file names written, and so on.

A sample graph was displayed on the monitor (Bauer & Jolicœur, 1996, Figure 3, top), and each of the steps required to process the
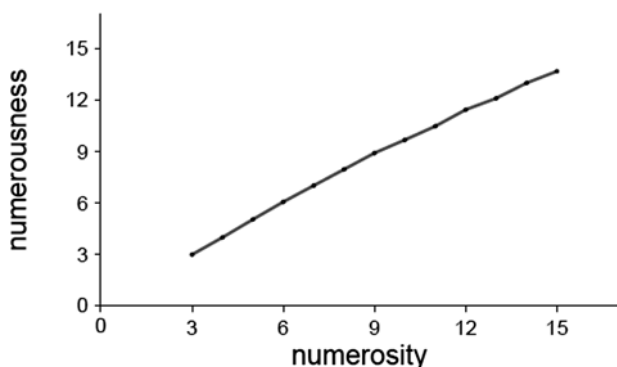

Figure 1. Geometric means computed from Jevons's (1871) data and replotted.

graph was shown to them by the investigator, who encouraged the observer to perform some of the steps. For each graph to be processed, a paper copy was available noting the nature of the axes (log–linear) and the number of data points to be located. The observer was told to ask the investigator for help at any time during the session if problems were encountered and to use the paper graphs to make notations or to circle/count the data points if needed. The entire session took between 60 and 80 min to complete the 16 graphs, which contained a total of 232 data points. The series of graphs to be processed was presented as a set of tabs in the interface.[1]

Observers were given instructions that contained five directions. First, observers were directed to make sure that they did not miss plotting symbols that fell on the axes. Second, observers were told to only process the data points and to ignore any "best-fit" lines. Third, when plotting symbols overlapped, observers were instructed to click the overlapped symbol location for each series so that the correct number of points was recorded. Fourth, the center of the plotting symbol was to be considered the location of the coordinate. Finally, if a plot contained more than one series of data, then the observer was told to look at the points corresponding to the lowest x-axis value. The series were then to be processed, starting with the series farthest from the x-axis and working downward.

RESULTS

Linear–Linear Scaling With Ideal Image Quality

A clean linear–linear graph was included as a baseline to discover how accurate the software can be with an ideal image. The graph was generated from a table in Jevons's (1871) article and replotted with very fine plotting symbols; then, it was converted directly to the image used (i.e., no printing or scanning steps). Therefore, there were no skew or scanning artifacts in the axes or labels, and there were no visual flaws such as noise dots or lint. The image (Figure 1) was 564×393 pixels (w $\times$ h) in size, and it was in ideal condition. In the version that the observers used, the line joining the data points was red.

Table 2
Original and Mean Recovered x- and y-Coordinates and Their Corresponding Standard Deviations, and Absolute Error in the Recovered Coordinates for Data Collected From Jevons (1871)

Original		Recovered				Absolute Error	
		x		y			
x	y	M	SD	M	SD	x	y
3	3.0	3.00	0.01	3.01	0.03	0.00	0.01
4	4.0	4.00	0.00	3.99	0.04	0.00	−0.01
5	5.1	4.99	0.00	5.05	0.01	−0.01	0.00
6	6.1	5.98	0.01	6.06	0.01	−0.02	−0.02
7	7.0	7.00	0.01	7.00	0.01	0.00	−0.03
8	8.0	7.99	0.00	7.94	0.01	−0.01	−0.03
9	8.9	8.98	0.00	8.89	0.03	−0.02	−0.03
10	9.7	9.97	0.00	9.68	0.03	−0.03	−0.01
11	10.5	11.00	0.00	10.48	0.04	0.00	−0.01
12	11.5	11.99	0.00	11.44	0.02	−0.01	−0.01
13	12.1	12.99	0.01	12.11	0.04	−0.01	−0.01
14	13.0	13.98	0.02	12.99	0.02	−0.02	−0.02
15	13.7	14.98	0.02	13.66	0.02	−0.02	−0.03

The analysis is based on results from 7 observers. The original coordinates and the mean recovered coordinates and standard deviations, with values truncated at 2 decimal places,[2] and collapsed across subjects, can be seen in Table 2. The graph reconstructed from recovered values is presented in Figure 2. The radius of the open circles represents a distance of 0.1 units on the x- and y-axes from the actual points in graph units. Standard deviation bars are too small to be seen. Clearly, for the purposes of establishing the values of the data points, g3data gives adequate results when recovered values are averaged over a small number of observers. Given that the standard deviations are so small, the recovered values from a single observer are likely accurate enough for scientific purposes.
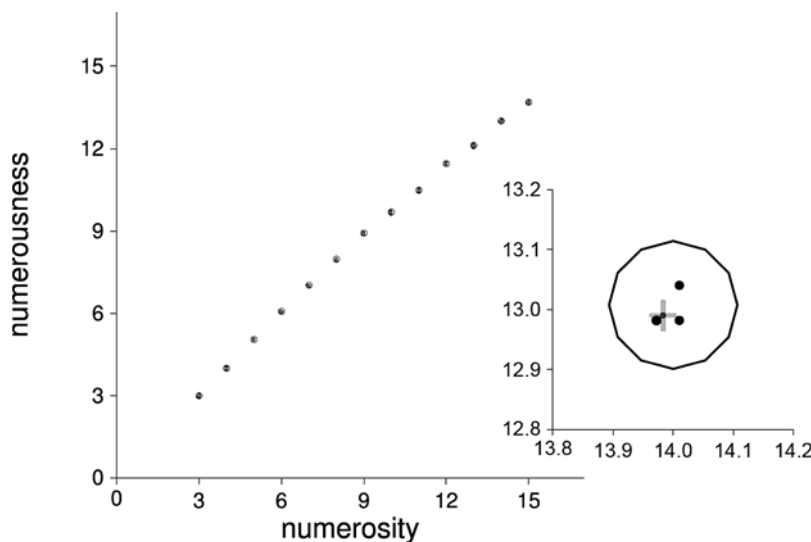

Figure 2. Reconstructed graph using coordinates estimated by g3data. Empty circles represent ±0.1 units on x and y around the original point. Individual observer coordinates are plotted as small black dots and those of the group mean are in gray. Inset: Enlargement of point x = 14.000, y = 13.007. Empty circle and black dots as in main graph. Gray crosshair is ±1 standard deviation around group mean coordinates.

Table 3
Linear Regression Slopes and Intercepts for Numerosity (*x*) and Numerousness (*y*) and the Regression Slopes Relating the Natural Logs of *x* and *y* for the Recovered Data As a Function of Observer and Group Mean and the Original Data From Jevons (1871)

Observer	Slope	Intercept	Slope ln(*x*), ln(*y*)
I	.892	.664	.935
II	.890	.641	.937
III	.891	.625	.941
IV	.892	.619	.941
V	.893	.594	.947
VI	.890	.633	.940
VII	.890	.620	.943
Recovered mean	.891	.628	.940
Original	.891	.631	.941

The data in the Jevons (1871) experiment was the estimated number of objects (numerousness) as a function of the number of objects presented (numerosity) under brief viewing conditions. Therefore, in order to test the accuracy of each observer's recovered data, linear regressions were computed on the original data, on each observer's observations (without truncation), and on the natural log of *x* and *y*. The rationale for using the natural log is that in psychophysical work, the slope of the linear regression of ln(*y*) with ln(*x*) is called the *Stevens exponent*, which captures the nature of the growth of sensation as a function of increase in stimulus intensity.

The data in Table 3 show that measures computed on the individual and group means are essentially interchangeable with those computed on the original data. Given that the plotting symbol was so small (3 pixels across in the g3data zoom window), the variance on the recovered points is expected to be small. In fact, this constraint resulted in several observers providing identical coordinates for some points (e.g., for original point $x = 15.000$, $y = 13.692$—the last point in Table 2—the recovered values for 4 observers were $x = 14.962$, $y = 13.655$; for 2 observers, they were $x = 15.000$, $y = 13.655$; and for the remaining observer, $y = 13.716$). This suggests that small plotting symbols will lead to less variance and better accuracy in recovery.

One might naturally ask why the *x* coordinates (independent measure) whose integer values are known from the stimulus description in Jevons's (1871) article were not substituted for the noninteger values recovered with g3data. In this case, there is little—if any—room to demonstrate improvement in the derived slopes or intercept. The value of substituting original for recovered values of the independent variable is tested in a following section. Under ideal conditions, the points recovered by a single observer or taken as the group means produced regression fits and plots that do not deviate in any practical way from the original data. Of course, the fact that the data were nearly linear worked in favor of this finding.

Linear–Linear Scaling
With Moderate Image Quality

A noncontrived plot with moderate image quality and linear–linear scaling (Taves, 1941, Figure 15) was used to determine whether the accuracy of recovered data declines as the image properties deviate from ideal. The original data are given in Taves's Table 7. This image was selected because it contains aliasing on the axes due to skewed reproduction (the plot appears to be turned slightly clockwise), the points are less systematic, and there are data points (large plotting symbols) on the *x*-axis, which obscure the reference point for calibrating the program (see screen capture from g3data interface in Figure 3 and full image in Figure A1 in the Appendix). The image is otherwise clean, with legible (but somewhat ragged looking) axis values.

The original graph (Figure A1 in the Appendix) is from a psychophysical study in which the observer responds "same" or "different" to the stimulus that changes in value from trial to trial. The plot shows the percent of time the observer said "different" as a function of the difference between a reference stimulus and the comparison stimuli. This explains why two points are at $y = 0$ (the observer never said "different" to that stimulus—see Figure 3 for an example) and why two points are at $y = 100$ (the observer always said "different" to that stimulus).
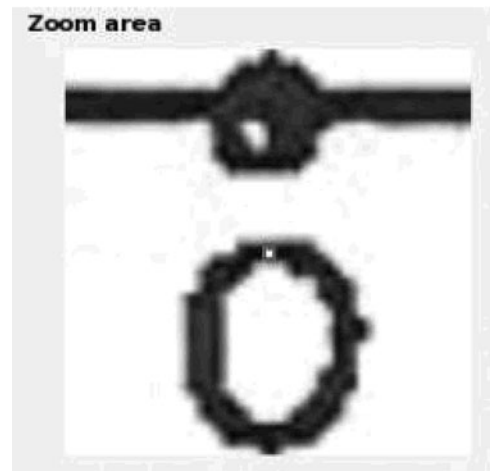

Figure 3. Screen capture from g3data GUI. This is the zoom area that provides a magnified version of the area under the cursor. The ragged 0 is the *x*-axis anchor, the horizontal line is part of the *x*-axis, and the spot on the axis is a plotting symbol.

Table 4
Original *x*- and *y*-Coordinates From Taves (1941), Mean Recovered *x*- and *y*-Coordinates With Their Corresponding Standard Deviations, and Original − Recovered Absolute Error Scores

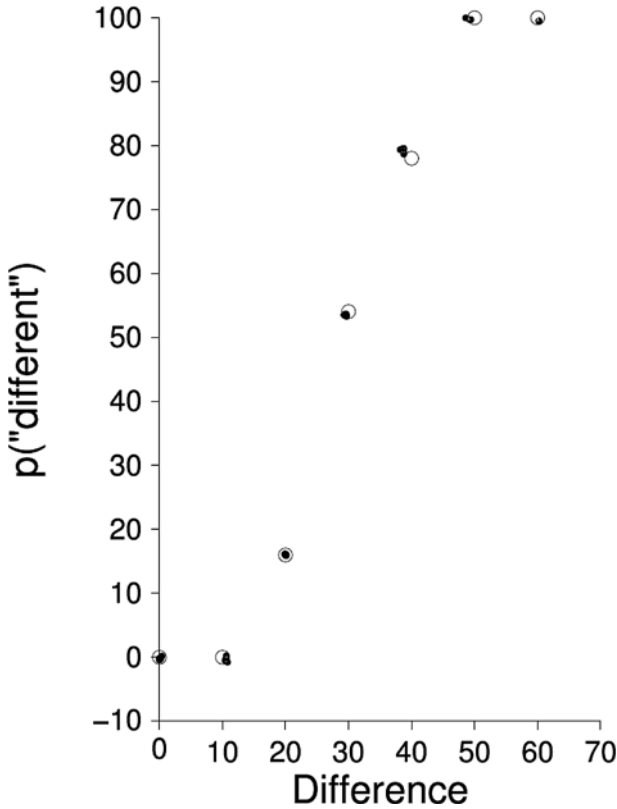| Original		Recovered				Absolute Error	
		x		*y*			
x	*y*	*M*	*SD*	*M*	*SD*	*x*	*y*
0.0	0	0.1	0.16	−0.2	0.21	0.1	−0.218
10.0	0	10.6	0.13	−0.3	0.35	0.6	−0.334
20.0	16	20.0	0.10	16.0	0.10	0.0	0.040
30.0	54	29.5	0.17	53.5	0.13	−0.5	−0.540
40.0	78	38.6	0.21	79.3	0.32	−1.4	1.256
50.0	100	49.0	0.30	100.2	0.53	−1.0	0.233
60.0	100	60.0	0.25	100.0	0.44	0.0	0.025

Figure 4. Reconstructed graph using coordinates estimated by g3data. Empty circles represent ±1.0 units on the *x*- and *y*-axes. Individual observer coordinates are plotted in small black dots and those of the group mean are in gray. These may be indistinguishable in the plot.

As can be seen in Table 4, the aliased *x*-axis resulted in a slight miscalibration of the *y*-axis, which was not observed with the ideal image used previously. Calibrating the *y*-axis requires clicking on two points: an upper point that should be at or higher than any of the data points (perhaps at the next *y*-axis checkmark with a printed value) and a bottom point that should be at or lower than any of the data points (often times at a value of 0 for *y*). The *x*-axis in the present graph is at least 3 pixels tall in the zoom area (see Figure 3), and setting the calibration point for the *y*-axis at three different pixel positions in the *x*-axis would produce three slightly different calibrations. This could result in variability in recovered values if observers differed in how they marked the points on the axes during calibration. Note that this can also create negative *y*-axis coordinates for the data points.

The reconstructed plot using the points from the 6 observers is presented in Figure 4. The radius of the open circles represents a distance of 1 unit on *x* and *y* in graph units. Part of the *y*-axis is negative in order to accommodate the small negative values recovered. For this type of experiment, the point at which the observer responds equally often "same" and "different" is referred to as the *difference limen* (DL) (threshold). This point—the 50:50 point—is estimated from the data in various ways. Taves (1941) reported a DL of 31.0, which is close to the value one might

estimate by eye from Figure 4 by observing that the middle data point is at approximately *y* = 50 and *x* = 30.

To determine whether the DL for the original data is similar to the DL for the recovered points, both were computed using psignifit Version 2.5.6 (Wichmann & Hill, 2001), assuming cumulative Gaussian distributions. For the original data (left two columns in Table 4), the DL is 30.18, and for the means of the recovered *x* and *y* values, the DL is 29.49 after limiting the *y* values to between 0 and 100 (the psignifit program correctly rejects values outside this range).

Table 5
Recovered Difference Limens Computed for Each Observer With the *p*("Different") Range Restricted Between 0 and 100 (See Text) As a Function of the Recovered and the Original *x*-Coordinates

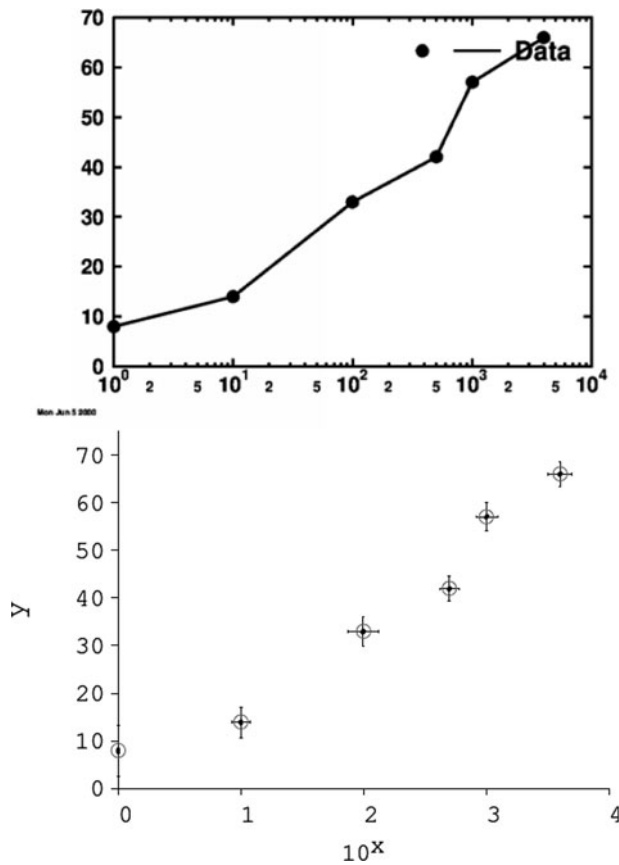| Observer	Recovered *x*	Original *x*	Difference
I	29.24	29.97	−0.73
II	29.50	29.97	−0.47
IV	29.42	29.97	−0.55
V	29.84	30.18	−0.34
VI	29.63	29.97	−0.33
VIII	29.53	29.97	−0.44

Figure 5. Top: Original log–linear graph as provided in the g3data install package. Bottom: Reconstructed graph using coordinates estimated by g3data. Empty circles represent ±1.0 unit on *y*. Individual observer coordinates are plotted in small black dots, and the group mean is in gray. These may be indistinguishable in the plot. The error bars are 20 times the standard deviations for visibility.

Table 6
Original and Mean Recovered x- and y-Coordinates With
Corresponding Standard Deviations From the test2.png
Graph Provided With g3data

Original			Recovered				
				x		y	
$\log_{10}(x)$	x	y	$\log_{10}(x)$	M	SD	M	SD
0.000	1	8	0.000	1.00	0.00	7.92	0.27
1.000	10	14	1.000	9.99	0.00	13.91	0.16
2.000	100	33	1.996	99.10	0.01	32.91	0.15
2.699	500	42	2.699	499.80	0.00	41.89	0.13
3.000	1,000	57	3.002	1,004.38	0.00	56.98	0.15
3.602	4,000	66	3.601	3,985.66	0.00	65.94	0.13

Table 7
Linear Least Squares Fits to Recovered Data As a Function of
Each Observer, to the Recovered Observer Group Mean Data,
and the Original Data From test2.png Provided With
the g3data Software

Observer	Slope	Intercept	r^2	$F(1,4)$	$p <$
I	.01152	25.74	.61	6.32	.066
II	.01148	25.91	.61	6.32	.066
III	.01173	25.73	.62	6.39	.065
IV	.01183	25.55	.62	6.41	.065
V	.01175	25.61	.61	6.35	.066
VI	.01174	25.77	.62	6.45	.065
VIII	.01180	25.45	.61	6.35	.066
Group mean	.01169	25.68	.61	6.37	.062
Original	.01163	25.79	.61	6.33	.066

Replacing the recovered x values with their original values (0–60 in steps of 10) results in a DL of 29.97. Because the values of the independent variable are often known from the source articles, this seems like a valid substitution. At least one negative value was present in each observer's data set. The DL was also computed for the data points recovered by each of the 6 observers after limiting the percentages to the 0–100 range. The DLs are presented in Table 5.

Given the small magnitude of the standard deviations in Table 4 and the close correspondence between the original and recovered coordinates, it appears that g3data is adequate for estimating the original values, even under conditions of moderate distortion. Furthermore, the derived measures—slopes and intercept in the Jevons (1871) data and the DLs computed on Taves's (1941) data—were very close to the expected values. This was true for the group mean coordinates and those for individual observers.

The graphs explored thus far were linear–linear, contained a small number of points, and were free of noise spots. To verify the accuracy of g3data for log–linear and log–log plots, two more graphs were tested.

Log–Linear Scaling With Ideal Image Quality

In order to discover how accurate the g3data software can be at recovering nonlinear data, a clean log–linear graph was used to establish a baseline. The graph is a log–linear plot shipped along with original coordinates in the g3data install package (test2.png). This 600×423 pixel (w × h) image is clean and contains six data points joined by lines (see Figure 5). On the printed sheet available to the observers, the phrase "log–linear" was provided to in-

dicate that the scaling was logarithmic on x and that they should set the "x-axis is logarithmic" box in the interface, but leave the "y-axis is logarithmic" box unchecked. Also on the printed sheet, the base10 value for each of the exponential values was provided (e.g., $10^2 \rightarrow 100$), and it was these latter values that were used as input to the calibration step. Results from 7 observers were analyzed.

This semilog plot produced results very consistent with the original data for both the observers' means and individual observer data. There is little evidence for systematic bias in the recovered coordinates other than perhaps a slight underestimation of the y coordinates (see Table 6). Similarly, the linear regressions on the base10 values of x and y were very consistent (see Table 7).

Log–Log Scaling With Poor Image Quality

To determine whether the accuracy of the data points recovered using g3data declines as the image properties deviate from ideal, performance was examined using a log–log plot taken from Kaufman, Lord, Reese, & Volkmann (1949, Figure 7) that contains 35 points with far-from-ideal image quality. Indeed, this was by far the worst image processed. It contained streaks that were presumably from the microfilm reader or printer, noise dots (some as large as the plotting symbols), and a slight clockwise skew (see Figure A2 in the Appendix). The independent variable was again the number of items presented, and the dependent variable was the median reported number of items. Both axes were logarithmic and labeled in linear values (i.e., not as powers of 10).

Two steps were taken to reduce error. First, the phrase "log–log" was written on the paper copy. Second, Data Point 1,1 was circled on the paper copy and pointed out to the observers to make sure it was not missed. This was
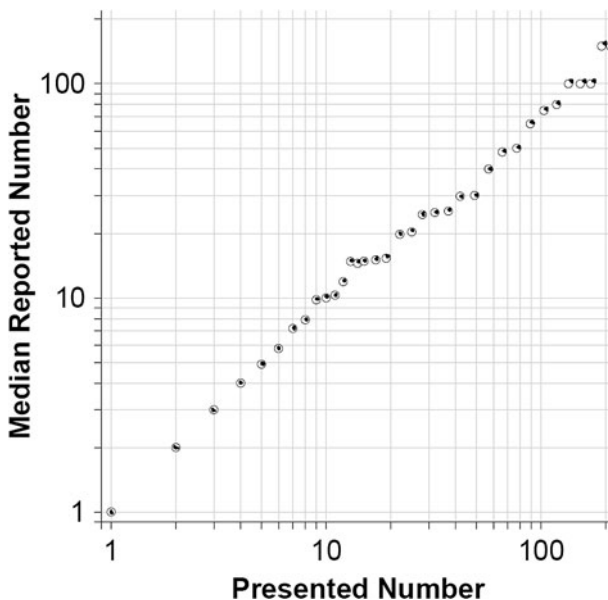

Figure 6. Reconstructed graph using coordinates estimated by g3data. Scaling of empty circles is arbitrary and is used to show the location of observed points with respect to the position of original points. Individual observer coordinates are plotted in small black dots.

Table 8
Original and Mean Recovered *x*- and *y*-Coordinates With Their Corresponding
Standard Deviations and Signed Absolute and Proportion Error
From Kaufman et al. (1949, Figure 7)

		Recovered				Absolute Error		Proportion Error	
Original		*x*		*y*					
x	*y*	*M*	*SD*	*M*	*SD*	*x*	*y*	*x*	*y*
1	1.00	1.00	0.00	1.00	0.01	0.00	0.00	.002	−.002
2	2.00	1.99	0.03	2.00	0.01	−0.01	0.00	−.004	.001
3	3.00	2.98	0.03	2.99	0.03	−0.02	−0.01	−.007	−.002
4	4.00	4.00	0.03	4.02	0.03	0.00	0.02	.000	.005
5	4.90	5.06	0.04	4.93	0.04	0.06	0.03	.013	.007
6	5.80	6.01	0.02	5.83	0.05	0.01	0.03	.002	.006
7	7.20	7.09	0.04	7.24	0.08	0.09	0.04	.013	.006
8	7.90	8.09	0.05	7.93	0.05	0.09	0.03	.011	.004
9	9.80	9.10	0.07	9.88	0.04	0.10	0.08	.011	.009
10	10.00	10.06	0.08	10.15	0.08	0.06	0.15	.006	.015
11	10.30	11.10	0.06	10.33	0.08	0.10	0.03	.009	.003
12	11.90	12.16	0.06	12.03	0.07	0.16	0.13	.013	.011
13	14.80	13.29	0.42	14.93	0.13	0.29	0.13	.022	.009
14	14.50	14.03	0.38	14.77	0.08	0.03	0.27	.002	.019
15	14.84	15.15	0.09	14.92	0.09	0.15	0.08	.010	.005
17	15.05	17.18	0.09	15.22	0.15	0.18	0.17	.010	.011
19	15.31	19.31	0.16	15.63	0.13	0.31	0.32	.016	.021
22	19.82	22.19	0.12	19.95	0.16	0.19	0.13	.009	.006
25	20.34	25.21	0.14	20.68	0.10	0.21	0.34	.009	.017
28	24.50	28.30	0.16	24.62	0.29	0.30	0.12	.011	.005
32	25.05	32.57	0.19	25.16	0.17	0.57	0.11	.018	.004
37	25.42	37.48	0.22	25.75	0.18	0.48	0.33	.013	.013
42	29.84	42.60	0.26	29.78	0.21	0.60	−0.06	.014	−.002
49	30.11	50.11	0.41	30.22	0.22	1.11	0.11	.023	.004
57	40.00	58.03	0.56	40.07	0.30	1.03	0.07	.018	.002
66	48.00	67.55	0.47	48.60	0.38	1.55	0.60	.023	.013
77	50.00	78.73	0.56	50.58	0.40	1.73	0.58	.023	.012
89	65.00	90.84	0.95	66.22	0.55	1.84	1.22	.021	.019
103	75.00	105.73	0.83	76.32	0.66	2.73	1.32	.027	.018
118	80.00	120.15	0.95	81.46	0.72	2.15	1.46	.018	.018
134	100.00	138.36	1.16	103.20	0.98	4.36	3.20	.033	.032
152	100.00	159.21	1.34	103.18	0.94	7.21	3.18	.047	.032
170	100.00	176.33	1.48	103.30	0.95	6.33	3.30	.037	.033
191	150.00	198.04	1.82	154.42	1.47	7.04	4.42	.037	.029
210	150.00	219.32	2.01	153.48	1.75	9.32	3.48	.044	.023

Log−Log Regressions	Slope	Intercept	r^2
Mean recovered *x* with mean recovered *y*	.888	.177	.993
Original *x* with original *y*	.890	.174	.993
Original *x* with mean recovered *y*	.895	.170	.993

done because the data point was obscured by the intersection of the two axes and the best-fit line that ran through it. Note that in Figure 6 below, the 1,1 point has been moved off the axis for clarity.

The original and recovered data can be seen in Figure 6 and Table 8. Overall, the accuracy of the recovered data was quite high; however, there appears to be a systematic overestimation of the *x* and *y* values toward the high end of the scale. This is easily seen in Figure 6, where the center of each open circle represents an original data point and the solid circles represent the recovered data; these tend to be high and to the right of the center of the open circle. The choice of size of the circle around the original points was arbitrarily set to show the difference in accuracy as the true coordinates increase in value. The increase in absolute error of the recovered data points as the original *x*- and *y*-coordinates increased can be seen in Table 8. Mean and maximum proportion error [(re-

covered − original)/original] were .016 and .047 for the *x*-axis and .012 and .033 for the *y*-axis.

As in the Jevons (1871) data set, a typical use for this type of data is to compute the Stevens exponent (the slope of the linear least squares fit for the log–log data). The slight concave-downward trend to the data indicates a slope of less than 1. The intercepts are of little theoretical value, but are shown as evidence of the consistency. The slopes and intercepts computed on the group mean data, the original data, and on the group mean *y* values regressed with the original *x* values (see Table 8, bottom) are all the same to the second decimal place (rounded).

Even with the systematic bias in the larger values, the reconstructed graph compares favorably with the original, and the derived slopes and intercepts correspond quite well. It appears that the g3data program performs adequately for scientific work. The question of whether the observers perform adequately can be answered by looking at the slopes

Table 9
Recovered Log–Log Slope Calculated for Each Observer Using Either the Recovered or Original x-Coordinates From Kaufman et al. (1949, Figure 7)

Observer	Recovered x	Original x	Difference
I	.8893	.8944	−.0051
III	.8896	.8965	−.0069
IV	.8834	.8919	−.0085
V	.8945	.8982	−.0037
VI	.8870	.8959	−.0089
VII	.8848	.8934	−.0086
VIII	.8869	.8935	−.0066
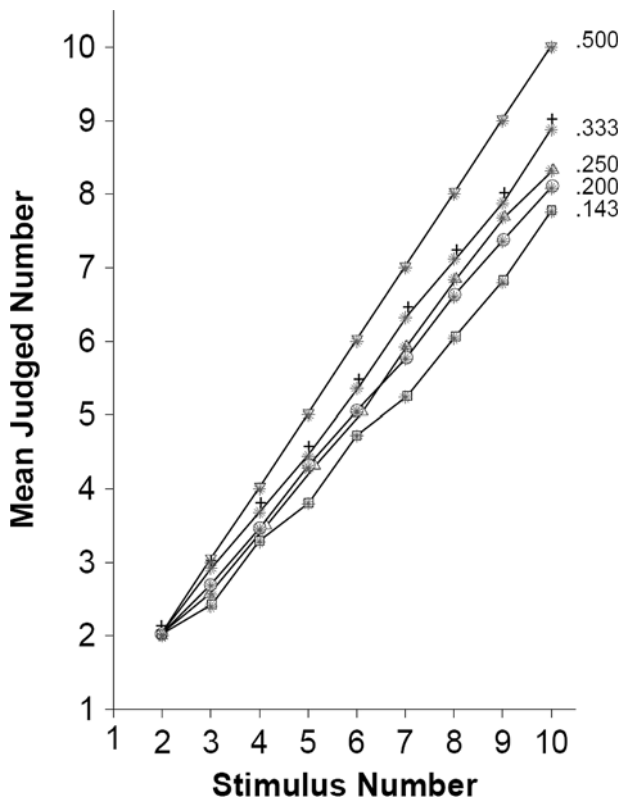
Figure 7. Reconstructed graph using coordinates estimated by g3data. Gray spoke symbols are centered at the original coordinates. The decimal numbers beside each data set reflect the conditions in the original Taubman (1950) article. See text for plotting issue with plus ("+") symbol.

for each observer computed with their obtained y and either their obtained x or the original x values (see Table 9). Again, no worrisome deviations are evident.

Linear–Linear Scaling With Missing Information and a Factorial Design

The final type of graph investigated is a relatively clean linear–linear plot with only a few noise specks and a slight clockwise skew (Taubman, 1950, Figure 1 and Table 2, Observer BS). This graph was included in testing for two reasons: First, a complex factorial design is used in which there are five data sets, each containing nine data points. Second, it has overlapping points, several different plotting symbols, and the axis lines do

not meet (see Figure A3 in the Appendix for the image used).

Given the complexity of the graph, students were instructed to process the data points starting on the extreme left of the graph and to process the values working downward (because all values overlapped at $x = 2$, this only begins to apply when the values for $x = 3$). They were told to use the center of the symbols (∇, 0, +, Δ) as the location of each data point, and, when dealing with coincident data points, they were also told to click the overlapping symbol once for each line it belonged to.

Only 2 of the students followed the instructions concerning the processing order for the five series. The other 6 processed top-down, but they started with the highest series at the value $x = 10$. This error was corrected in the students' data files. Clearly, the instructions given by the investigator were at fault. This example does highlight the value of clear directions; perhaps the order of processing should be indicated on the printed sheets. Knowing the order of processing is critical, because the data

Table 10
Linear Regression Slopes and Intercepts and Linear Least Squares Fits (r^2) to Unscaled Coordinates and Linear Regression Slopes to Log–Log Transformed Data As a Function of Series for Original and Mean Recovered Data

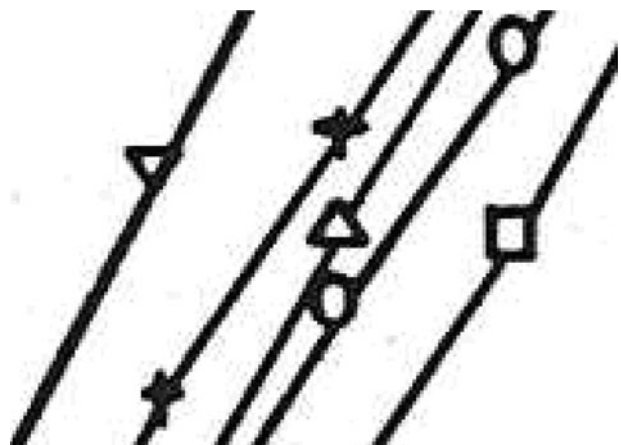| Series	Slope	Intercept	r^2	Slope (Log–Log)
		Original Data		
.500	1.000	.000	1.000	1.000
.333	.853	.284	.999	.921
.250	.818	.212	.998	.923
.200	.769	.410	1.000	.884
.143	.720	.351	.995	.863
		Recovered Data		
.500	.997	.049	1.000	.987
.333	.849	.307	.999	.911
.250	.813	.218	.997	.908
.200	.767	.447	1.000	.874
.143	.719	.371	.994	.854

Figure 8. Enlargement of a small section of the original Figure 1 in Taubman (1950). Note how the line joining the symbols (this was not a best-fit line, clearly a line connecting the symbols) seems to connect the bottom edge of the 0 and ∇ plotting symbols within a series.

files contain only a list of coordinate pairs acquired, and the values from multiple sets might become confused. Perhaps a way to add an annotation to the saved file would help here. Because the data files are plain text, there is nothing preventing the addition of comments with a text editor. At the same time, a facility in the GUI for adding comments or notes might prove helpful.

The instruction to use the center of each symbol as the correct coordinate was also problematic. The fine spoke symbol used in Figure 7 is centered at the correct coordinate. Its radius covers 0.2 in x-axis units. Notice how the location of the spoke symbol changes in comparison with the other plotting symbols. For instance, the plus sign used for the plotting symbol in condition .333 of the recreated graph does not plot with the intersection at the data points. Rather, the data points are located at the bottom of the plus sign. This brings up an important issue, especially since many of the plots in journal articles older than 25 years or so might have been hand-drawn on acetate. The information given to the observers was to consider the data point to be at the center of the plotting symbol. If the convention for the hand-drawn graph or the computer plotting package does not respect this rule, then error may be introduced into the recovered values. Given the large number of points on this plot, only summary data will be given.

A repeated measures ANOVA with the absolute unsigned error on the y-coordinate as the dependent measure, and the factors of series and x value as independent measures, revealed a rank ordering of error from worst to best of the series .500, .200, .333, .143, .250, or in terms of the plotting symbol, ∇, 0, +, or Δ [$F(4,28) = 38.97$, $MS_e = .0001$, $p < .001$]. Obviously, the plotting symbol is confounded with the series, so other factors could be at play. However, taken at face value, the ranking is consistent with the differences in plotting symbol positioning on the lines joining the symbols within a series in the original graph (see Figure A3 in the Appendix). For example, the line for the .500 series in the original joins the bottom vertices of the ∇ symbol (see Figure 8 for an enlarged section from the original graph). The 0 series (.200) seems to follow the same convention, and it was second to the ∇ in error. The center-of-symbol instruction given to the observers might explain the small discrepancies between the slopes computed as compared with those computed on the original data. Thus, center-of-symbol instruction may not be ideal for some types of graphs.

CONCLUSION

For the range of graphs tested here, g3data performed well in the hands of a small sample of undergraduate observers. They were able to process over a dozen graphs in about an hour, including the time to learn how to use the interface and save data. In general, there is little difference between group performance and individual performance, suggesting that a single conscientious research assistant would probably produce accurate data. Of course, there are many kinds of plots untested in the present sample, and the ability of g3data to recover data points would require

some testing. At the same time, it performed very well for the linear–linear, semilog, and log–log plots tested here. Even derived measures, such as slopes, psychophysical DL, and intercepts, fared well. There appeared to be little practical benefit to replacing the recovered values for the x-axis with the true values. Again, this was not tested extensively, so at this point, the researcher would be prudent to verify that both lead to a similar result, especially if derived measures might propagate error. In addition, the substitution is likely to make the generation of new plots easier and more accurate, given common x-axis values. Any data recovered by g3data or any other method for use in subsequent publications should be accompanied by notification that the coordinates are now secondhand with declaration of any particular extraction details (e.g., center of symbol estimation, repeated measures) so that it is clear that the generated data are estimation. In short, researchers should feel comfortable using data recovered from simple graphs—log or linear—by g3data, provided that the person operating the software is methodical, understands when to indicate log scaling, and correctly saves the obtained coordinates.

REFERENCES

BAUER, B., & JOLICŒUR, P. (1996). Stimulus dimensionality effects in mental rotation. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 82-94.

FRANTZ, J. (2000). g3data (for Linux) (Version 1.5.1) [Computer software]. Retrieved from www.frantz.fi/software/g3data.php.

JEVONS, W. S. (1871). The power of numerical discrimination. *Nature*, **3**, 281-282.

JONES, R. W., WARNER, J. W., & CROSS, C. L. (1998). How are graphs read? An indication of sequence. *Behavior Research Methods, Instruments, & Computers*, **30**, 238-245.

KAUFMAN, E. L., LORD, W. M., REESE, T. W., & VOLKMANN, J. (1949). The discrimination of visual number. *American Journal of Psychology*, **62**, 498-525.

TAUBMAN, R. E. (1950). Studies in judged number: II. The judgment of visual number. *Journal of General Psychology*, **43**, 195-219.

TAVES, E. H. (1941). Two mechanisms for the perception of visual numerousness. *Archives of Psychology*, **265**, 47.

WICHMANN, F. A., & HILL, N. J. (2001). The psychometric function: I. Fitting, sampling and goodness-of-fit. *Perception & Psychophysics*, **63**, 1293-1313.

NOTES

1. Some observers neglected to process some of the graphs (or failed to save their data), so some of the analyses presented are based on fewer than 8 observers. Perhaps a warning built into the software to remind that there are unsaved coordinates would avert this error in the future.

2. The data files generated by g3data contained between 0 and 13 decimal places for various points. Obviously, that level of accuracy is subpixel and therefore meaningless. There is an estimate of spatial error in the coordinate values provided in the interface, and these estimates can be exported along with the point values.

APPENDIX
Vignettes of Three of the Four Plots Characterized

See text for further information. Note that skew and noise artifacts are present in the scanned image files.
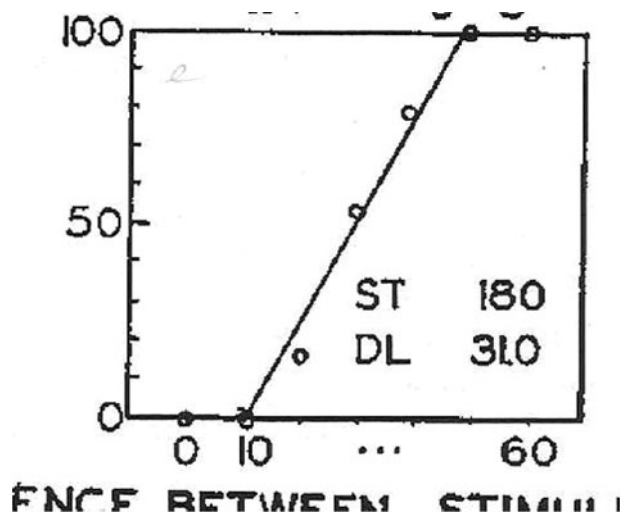

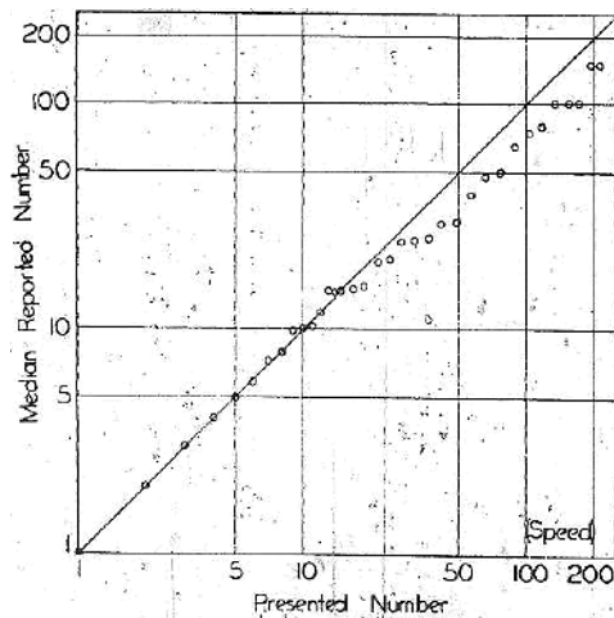**Figure A1. Taves (1941), Figure 15, bottom; 549 × 441 pixels (w × h).**

Figure A2. Kaufman et al. (1949), Figure 7; 561 × 499 pixels (w × h).
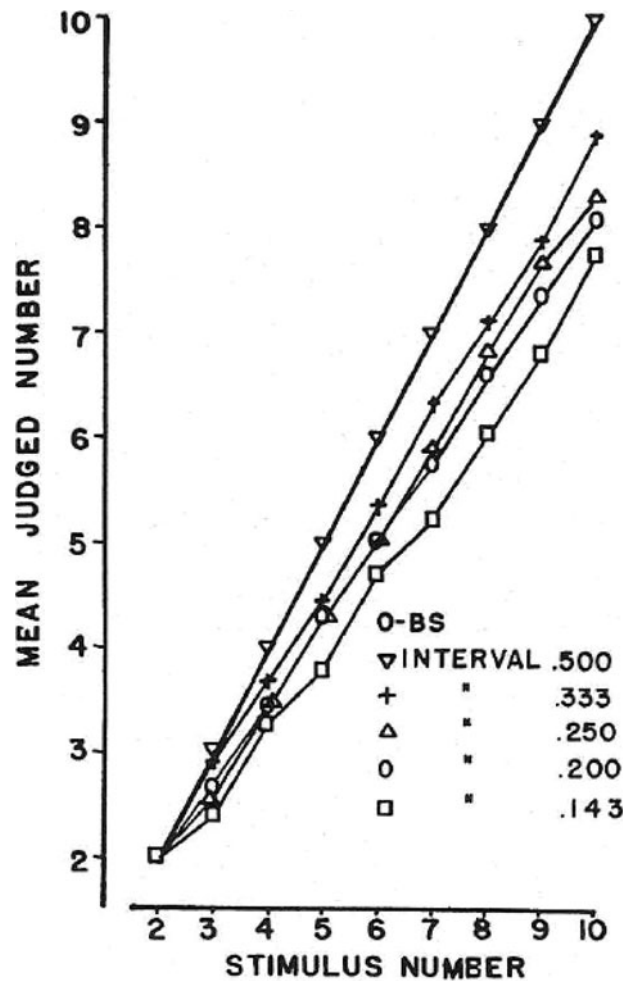
APPENDIX (Continued)

Figure A3. Taubman (1950), Figure 1; 661 × 898 pixels (w × h).

The following graphs were also processed but are not presented in the present article.

Casperson, R. C., & Schlosberg, H. (1950). Monocular and binocular intensity thresholds for fields containing 1–7 dots (Figures 2A–2E; Table 1). *Journal of Experimental Psychology*, **40**, 81-92.

Hanes, R. M. (1949). The construction of subjective brightness scales from fractionation data: A validation (Figure 1; Table 1). *Journal of Experimental Psychology*, **39**, 719-728.

Kaufman, E. L., Lord, W. M., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number (Figure 6; Table 1). *American Journal of Psychology*, **62**, 498-525.

Taves, E. H. (1941). Two mechanisms for the perception of numerousness (Figure 15: ST 15, ST 25, ST 50, ST 100; Table 7 Observer BT). *Archives of Psychology*, **37**, 625.

Thurstone, L. L. (1929). Fechner's law and the method of equal-appearing intervals (Figure 3; Table 2). *Journal of Experimental Psychology*, **12**, 214-224.