Graphical Exploration of Statistical Assumptions: Can We?

Lena Brandl

University of Twente

1st Supervisor: Dr. Martin Schmettow

2nd Supervisor: Dr. Stéphanie van den Berg

24th of June 2016

Table of Contents

Abstract

In statistical inference, failure to control for violations of statistical assumptions increases the risk of committing a type I (rejecting a true null hypothesis) or type II error (failure to reject an untrue null hypothesis). Graphical exploration of statistical assumptions is advocated. The objective of the current study was to confirm the usefulness and superiority of statistical graphics in data exploration. A complete within-subject design was employed, exposing participants to 100 simulated dot-histograms and 100 box-jitter plots. Participants' ability to visually detect violations of the normality assumption and homoscedasticity was assessed. Results revealed that participants were not able to validly detect violations of statistical assumptions. Exploratory data analysis and a general linear mixed-model in form of a logistic regression further uncovered that participants did not inform their choices by objective criteria. However, our results are ambiguous. Participants varied in their baseline tendency to reject stimuli and in how much they were influenced by objective criteria. Stimuli varied in which kind of response they provoked, even after objective criteria were taken into account. We conclude that the scientific community is not ready for a methodological shift from conventional statistical techniques to graphical data exploration. A temporary joint usage of statistical tests and graphics is advocated while further research investigates whether our results can be replicated with subjects of higher statistical proficiency. Drastic changes to the psychological curriculum and empirical research into the effectiveness of competing statistical graphics shall prepare the community for the supersession of conventional statistical techniques.

In their protocol for data exploration, Zuur, Ieno, and Elphick (2010) firmly point out the importance of data exploration prior to conducting statistical analyses. The authors note that their students in ecology frequently neglect to examine data with regard to the underlying assumptions of the statistical techniques employed. While some violations of statistical assumptions may have little effect on statistical outcomes, others can distort results dramatically, leading to wrong conclusions and poor recommendations. In statistical inference, violations of statistical assumptions can, among other things, increase type I (rejecting a true null hypothesis) and type II errors (failure to reject an untrue null hypothesis). As the authors put it: "All statistical techniques have in common the problem of 'rubbish in, rubbish out'." (p.3).

Multiple authors advocate the use of statistical graphics in data exploration (Behrens & Yu, 2003; Chatfield, 1985; Gelman, Pasarica, & Dodhia, 2002; Gelman, 2011; Kline, 2008; Marmolejo-ramos & Valle, 2009; Nolan & Perrett, 2015; Zuur et al., 2010). As Tufte (2007) stated: "(…) of all methods for analyzing and communicating statistical information, well-designed graphics are usually the simplest and at the same time the most powerful." (p.9). According to Cleveland (1984), the natural pattern recognition abilities of the human brain lie at the heart of efficient graphical communication. However, in the literature on the use and design of statistical graphics (e.g. Marmolejo-ramos & Valle, 2009; Tukey, 1977; Zuur et al., 2010) we could find no empirical evidence that well-designed graphics indeed enable humans to arrive at valid inferences about scientific data. Therefore, the objective of the current study is to confirm the promised merits of statistical graphics for data exploration. We test in how far humans are indeed able to detect violations of statistical assumptions by means of graphical perception.

Theoretical Background

Statistical Assumptions in Analysis of Variance

It has been established that the field of ecology is no exception when it comes to statistical problems. According to Leys & Schumann (2010), experimental social psychologists tend to apply Analysis of Variance (ANOVA) to study associations between factors, even when underlying assumptions are not met. This severely compromises the validity of their outcomes.

Howell (2012) designates that ANOVA "has long enjoyed the status of being the most used (some would say abused) statistical technique in psychological research." (p.320). ANOVA is a variant on regression and as such, its logic is rooted in regression (Field, Miles, & Field, 2012). Basically, ANOVA tests how well a certain model fits the observed data. In basic experimental research, the effect of some predictor variable is examined on an outcome variable. To this end, multiple experimental conditions are devised in which participants are manipulated to the extent that they differ in the amount of the predictor variable, but ideally, in no other regard. ANOVA is then applied to test whether group means in the outcome variable significantly differ from each other. ANOVA produces a F-statistic or F-ratio which is the ratio of the explained to the unexplained variation in the outcome variable. Variation refers to differences between observed values and predicted values of the outcome variable. Values are predicted on the basis of the model at hand. The F-ratio is calculated using the following formula:

$$F = \frac{MS_M}{MS_R} \qquad (1)$$

Where MS_M is the average amount of variation explained by the model, and MS_R the average amount of unsystematic variation which cannot be accounted for by the model. As the mathematical calculation of the F-statistic lies beyond the scope of this paper see for example Field et al. (2012) for a more detailed description. When the F-value exceeds a critical value, one may conclude that the examined predictor variable indeed affects the outcome variable in some way. Otherwise, it would be highly unlikely to get the obtained F-value.

ANOVA is a parametric test based on the normal distribution and as such, it comes with a number of statistical assumptions, ensuring the reliability of the F-statistic (Field et al., 2012). A first requirement is that the dependent variable has to be measured on an interval scale at minimum (Leys & Schumann, 2010). Secondly, individual observations should be independent. Thirdly, residuals have to be normally distributed. In ANOVA, the assumption of normality boils down to the requirement that the outcome variable within groups has to be normally distributed (Field et al., 2012; Grace-Martin, 2012). A final statistical assumption of ANOVA is homogeneity of variance or homoscedasticity. This assumption holds that at each level of the predictor variable, the variance of the outcome variable is the same. When one collects groups

of data it means that group variances have to be sufficiently similar across groups (Field et al., 2012).

Whenever one or more of these assumptions is violated, the chance of committing a type I or type II error increases (Leys & Schumann, 2010; Zuur et al., 2010). Given ANOVA's omnipresence in psychological research, ensuring the efficient and reliable screening of data for violations of its assumptions is of crucial importance. For normality, the Shapiro-Wilk test (Shapiro & Wilk, 1965) is a conventional statistical technique for testing whether a distribution significantly deviates from normal (Field et al., 2012). Alternatively, the assumption of normality can be checked graphically using histograms (Field et al., 2012; Zuur et al., 2010). For checking homoscedasticity, Levene's test (Levene, 1960) is a conventional statistical technique (Field et al., 2012; Marmolejo-ramos & Valle, 2009). Again, the assumption of homoscedasticity can be examined visually. Zuur et al. (2010) advise and demonstrate the use of boxplots for this purpose. At this point it should be pointed out that there is a great deal of unclarity among researchers about how one should best control for violations of statistical assumptions (Field et al., 2012; Zuur et al., 2010). Some researchers even debate whether parametric tests should be replaced by nonparametric techniques to avoid the trouble of violating underlying assumptions (Johnson, 2009; Läärä, 2009).

Statistical Graphics in Data Exploration

Multiple authors advocate the use of statistical graphics in data exploration (Behrens & Yu, 2003; Chatfield, 1985; Gelman et al., 2002; Gelman, 2011; Kline, 2008; Marmolejo-ramos & Valle, 2009; Nolan & Perrett, 2015; Zuur et al., 2010). At the same time, Behrens and Yu (2003) note that graphical data exploration is rarely performed as a current conventional research practice in psychology. Kline (2008) states that blind reliance on test statistics discourages researchers from actually looking at their data, which is a crucial first step in any data analysis (Kline, 2008; Wilkinson, 1999). Marmolejo-ramos and Valle (2009) demonstrate how a combination of graphical methods and formal statistical tests outperforms the application of the very same statistical tests in isolation in preventing type II errors. The authors explore and analyze simulated data sets using conventional approaches, i.e. homogeneity of variance and normality tests,

graphical techniques proposed by Exploratory Data Analysis (EDA; Tukey, 1977), and data transformations. Specifically, the authors created two data sets that provoke committing a type II error. The observed means of the two data sets (group A and group B) are indistinguishable from each other by a parametric test while actually being different. Both data sets were drawn from a normal distribution, however, in one of them (group B), two normal observations were replaced with two outliers. In a first step, the authors applied the Lilliefors (Kolmogorov-Smirnov) normality test (Lilliefors, 1967) to examine whether the given samples were drawn from a normally distributed population. The test indicated no problem regarding normality. To check whether the variances of the two groups were homogenous, the Levene's test (Levene, 1960) was applied. Again, no problem was detected by a conventionally used preliminary test. In accordance with a provided research scenario, an independent-samples t-test was conducted to examine whether there was a statistically significant difference between the two group means. Unfortunately, the test detected no significant difference between means when in actuality, the data sets were created with different means. At this stage, a researcher who blindly relies on the outcomes of the applied statistical techniques would commit a type II error by wrongly rejecting the hypothesis that group means are different. In a second step, however, the authors estimated group densities using a kernel (Silverman, 1986; Wilcox, 2004). The kernel density plots revealed that the two groups had similar variances, but different distributions, possibly caused by outliers. Outliers are observations with less than 5% frequency (Cowles & Davis, 1982), and it is common practice to use a two standard deviations cut-off to determine them. A second test of normality was administered to re-examine sample distributions and to confirm what was made evident by the density plots. The Shapiro-Wilk test of normality is said to be more sensitive than the Lilliefors test (Field, 2009). And indeed, the test indicated that group B significantly departed from being normally distributed. Further investigation through application of outlier z tests (Shiffler, 1988) confirmed that there were two outliers present in group B. In a last step, the two outliers were removed and another two-tailed t-test was run. This time, the difference in means was correctly detected by the test. The simulation study of Marmolejo-ramos and Valle (2009) makes two important points. Firstly, the reliability of preliminary tests should be double-

checked, even more so if visual representations cast doubt on the reliability of the test. And secondly, graphical methods can spot problematic features of data that may not be detectable by means of conventional tests. The authors make a case for the joint application of statistical and graphical techniques in data exploration.

Exploratory Data Analysis. The pioneering work of John W. Tukey (1977) is closely related to the origins of statistical graphics and today's use of visual displays in data exploration. Tukey advocates exploratory data analysis (EDA) as a first step, as "numerical detective work (…) or graphical detective work" (p.1). This detective work precedes any kind of confirmatory data analysis (i.e. hypothesis testing), providing it with indications. According to Turkey, without these indications, confirmatory data analysis has nothing to consider. EDA promotes a very different philosophy of analyzing data. Instead of imposing a model on the data, as in confirmatory data analysis, EDA allows the data itself to reveal its structure. Tukey (1977) developed several statistical graphics on which data exploration, as proposed by Zuur et al. (2010), heavily draws. Inspired by Tukey (1977), Hartwig and Dearing (1979) further elaborate EDA techniques. The authors are convinced that knowing as much as possible about one's data by employing EDA will result in sounder data analyses compared to when EDA is omitted. To them, the exploratory perspective is a state of mind that includes both, skepticism towards summarizing statistics, and openness to unexpected patterns in the data. Hartwig and Dearing (1979) firmly hold that visual analysis should precede statistical analysis.

Merits of graphical communication. Statistical graphics greatly profit from the merits of graphical communication. The enormously powerful pattern recognition abilities of the human brain underlie the efficiency of graphical communication. Thanks to pattern recognition, graphs are capable of efficiently communicating vast amounts of quantitative data. At the same time, they make trends and other salient features pop out (Cleveland, 1984). Indeed, as Gelman et al. (2002) and Tukey (1990) recognize, the main interest in scientific research lies in comparison, not in absolute numbers. And graphs continue to outperform other modes of presentation, tables in particular, in this regard (Feliciano, Powers, & Kearl, 1963; Meyer, Shamo, & Gopher, 1999; Washburne, 1927). However, even though graphical communication has numerous merits for

the exploration and analysis of scientific data, there is much lack of clarity and empirical evidence concerning what graphics should be employed in which instances. Evermore graphics are being developed. At the same time, authors rather advocate their own preferences when it comes to the use of statistical graphics than support their choices with sound scientific test (e.g. Haughton & Haughton, 2011; Hintze & Nelson, 1998; C. Johnson, 2004; Marmolejo-ramos & Valle, 2009; Zuur et al., 2010).

Cleveland and McGill (1984) long recognized that graphical design is largely a matter of general agreement rather than sound empirical test. The authors react to this circumstance by initializing a scientific foundation for graphical design. Specifically, Cleveland and McGill provide guidelines on how to design graphs based on ten elementary perceptual tasks. According to the authors, the viewer performs certain perceptual tasks while decoding graphical information. Among these tasks are: position on a common scale, length, area, volume, shading, and direction. For example, the authors explain that viewers of a bar chart extract the values of the data by judging position on a common scale, in this case, the y-axis, combined with judgments of area and length. Cleveland and McGill (1984) advocate that graphical designers should take into account a ranking of the proposed elementary tasks according to the accuracy with which viewers perform these tasks. According to the authors, graphs that use elementary tasks as high in the hierarchy as possible will facilitate pattern detection and extraction of quantitative information. By re-designing commonly used statistical graphics in a way that converts low-hierarchy elementary tasks inherent in the graphics to high-hierarchy elementary tasks the authors demonstrate the effectiveness of their findings. The research of Cleveland and McGill (1984) aims at optimizing graphical design for efficient communication of statistical information, using human graphical perception, and thus, human ability as a starting point. Their research can be understood as one element of many in an emerging movement in the scientific community away from the focus on statistical tests as the dominating approach to scientific data.

Pattern Recognition and Expertise

As stated above, the merits of graphical communication are grounded in human pattern recognition (Cleveland, 1984). Pattern recognition has been identified as a major contributor to

expert performance in multiple disciplines (Chase & Simon, 1973; Klein & Hoffman, 1992; Loveday, Wiggins, & Festa, 2013; Regehr, Cline, Norman, & Brooks, 1994; Waters, Underwood, & Findlay, 1997). For instance, Loveday et al. (2013) were able to distinguish between competent non-expert diagnostic practitioners and expert diagnosticians by assessing pattern recognition performance during domain-relevant tasks in two settings, medicine and power control. According to the authors, competent non-experts make use of prior cases and heuristics, while true experts utilize reliable and efficient cognitive shortcuts. The authors define pattern recognition as "the non-conscious recognition of problem-states based on patterns of features that prime appropriate scripts in memory" (p.1). Highly specified and automated feature-outcome associations ("cue associations") in memory reduce processing load for experts without sacrificing depth of processing, and enable them to arrive at rapid and accurate diagnoses. The authors concluded that rather than years of experience, pattern recognition based assessments should be used to identify experts within samples of experienced diagnosticians. Expertise is pattern recognition.

Expertise research in chess (Chase & Simon, 1973; de Groot, 1965, 1966) illustrates how pattern recognition arises. In contrast to popular belief prior to de Groot's (1966) work, the distinguishing feature between expert and novice players is not thinking ahead. Neither experts nor novices think ahead more than a few moves. What expert chess players can do better is temporarily memorizing chessboard positions (Lesgold, 1983). De Groot (1965; 1966) and Chase and Simon (1973) demonstrated that chess experts have the same short-term memory constraints (Miller, 1956) as non-experts. Experts could recall as many positions as chess novices when trying to recall randomly scrambled chessboards following brief exposure. The superiority of master chess players derives from their ability to encode positions into larger perceptual chunks, each chunk consisting of familiar piece configurations. Chase and Simon (1973) identified the patterns by which pieces are bound in chunks: proximity, attack over small distances, mutual defense, common color, and type. Chunking strategies have been found to be similar for players of different skill levels. Despite being restricted by the same memory constraints as novices, master chess players could recall more and larger chunks when briefly

exposed to meaningful positions. Chase and Simon (1973) concluded that experts organize chunks hierarchically, enabling them to infer sub chunks from recalled chunks, increasing the total number of chunks recalled. Furthermore, the authors postulate that master chess players can "see" the right move solely on the basis of immediate perceptual processing, whereas less experienced players need to undergo slow and conscious reasoning.

Expertise research has shown that pattern recognition contributes to the superior performance of experts across multiple disciplines. As stated by Cleveland (1984), the human brain naturally possesses powerful pattern recognition abilities. For instance, Ehman et al. (2007) describe biomedical researchers as drawing on humans' natural way of making sense of the world. Biomedical researchers interpret complex multidimensional data by means of biomedical imaging, which facilitates rapid and accurate interpretation of information. Likewise, the analysis of scientific data is likely to profit with regard to accuracy and efficiency from integrating this powerful ability more extensively into formal analysis procedures.

In sum, data exploration and the examination of underlying statistical assumptions has been established to be a vital first step in any statistical analysis. Several authors (Gelman et al., 2002; Hartwig & Dearing, 1979; Kline, 2008; Marmolejo-ramos & Valle, 2009; Zuur et al., 2010) favor the use of visual displays in conducting data exploration. This preference is based on the premise that, provided with visual displays, human beings make better inferences concerning violations of statistical assumptions than statistical tests do. This assumption, however, has not been thoroughly examined yet. Are humans really able to detect violations of underlying assumptions by means of graphical perception? And, are they better at it than statistical tests? Research in pattern recognition and expertise draws a hopeful picture of efficient and accurate visual analysis. However, the reviewed expertise research also stresses the importance of specific domain knowledge underlying expert performance (Lesgold, 1983; Loveday et al., 2013). The current study draws on the faith advocators of graphical data exploration put in humans' visual detection abilities. We test these abilities experimentally in comparison to the performance of statistical techniques. The underlying statistical assumptions featured in this study are important assumptions of widely used parametric techniques, such as

ANOVA. Specifically, the current study examines subjects' ability to validly detect deviations from normality and homoscedasticity by means of graphical perception.

Method

Participants

In total, 33 subjects participated in the current study. 17 participants were male and 16 were female. Their age ranged from 19 to 32 (M = 22.78, SD = 2.768). Nine participants were Dutch, and 24 participants were German. As for inclusion and exclusion criteria, anyone who either followed or had completed their statistical education for psychologists at the University of Twente (UT), Enschede, The Netherlands could participate. This was necessary to ensure that participants possessed the required statistical knowledge to complete the experiment. Participants had to be 18 years old or older. Participants of the specific target population were recruited in three different forms. First, researchers approached their own acquaintances and fellow students directly. Second, participants signed-up for the experiment via the university-intern cloud-based subject management software. At the UT, psychology and communication science students are required to earn 15 subject hour credits within the first two years of their undergraduate studies. These credits are earned by participating in the research of fellow students. And finally, flyers were distributed and placed in the Cubicus, the behavioral sciences building of the UT. Potential participants could react to the flyers by either making an appointment with the researchers directly or by signing up via the online subject management software. Regardless of the specific form of sampling, potential participants received a short briefing about the goal and nature of the experiment on the basis of which they could decide whether to participate or not. The briefing included deceptive elements concerning the goal of the study. Participants were told that in order to improve the statistical training for psychologists, the UT intends to implement a serious game aimed at enhancing students' attitude towards statistics. Potential participants were made believe that by participating in the experiment, they would help with the development of this game. We chose to make use of participant deception to make the nature of our research more appealing to potential participants. After the experiment,

participants were informed that they had been deceived concerning the goal of the experiment and our true intentions were revealed. The current research received ethical approval by the Ethics Committee for Behavioral and Management Sciences at the University of Twente (Request Nr: 16073).

Materials

Statistical game - Game Mechanics

The software was programmed by the conducting researchers in Python 2.7. In order for the program to run on a computer, Python 2.7 (or a newer version) and the PyGame module are needed. The researchers used their own laptops to test subjects. One laptop had a resolution of 1366x768 pixels, the other a resolution of 1920x1080 pixels. Both laptops had a 15.6" screen. The program window itself had a resolution of 1024x700 pixels, regardless of the laptop used. Thus, the program window turned out to be much smaller on the second laptop. It was assumed that the difference in screen resolution did not have any serious confounding effects. However, in future research differences in screen resolution should be avoided or at least controlled statistically. Participants used the laptop's keyboard to interact with the software. When the program started up, a welcome screen appeared. When the return key was pressed, participants were prompted to fill in their participant ID and a number of demographic data, including gender, age, nationality, and study year. Apart from that, participants could optionally fill in their last known statistics grade. Afterwards, the rules of the statistical game were printed to the screen. The program included the following game mechanics: For each correct answer the participant received one point, for each incorrect answer, a point was lost. After a streak of five consecutive correct answers, two points were received for each following correct answer. After at least 15 consecutively correct answers, three points were granted per correct answer. If an incorrect answer was given, the score mechanics were reset to one point per correct answer. The program consisted of two rounds. Round one (normality) tested participants' ability to detect violations of the normality assumption. Round two (homogeneity of variance) tested participants' ability to detect violations of homoscedasticity. At the end of each round, a fictive leaderboard was printed to the screen. After the normality part, the participant's score was reset to zero. Both rounds

consisted of 100 experimental trials plus five practice trials which did not add to the participant's score. During the trials, participants were shown two plots on the screen, a stimulus plot and an "ideal". Depending on the construct being tested, participants were asked to react to a respective yes/no question that was printed to the screen by either pressing the <y> or <n> key on the keyboard. For normality, the question was "Are these scores normally distributed? (Y/N)", and for homoscedasticity, "Are the variances homogenous? (Y/N)". Answers were recorded in the form of numerical values. "0" marked a "correct" answer and "1" an "incorrect" answer. Instantly, the program provided feedback by printing "Correct!" or respectively, "Incorrect!" to the screen. Within the computer program, correctness was determined by the outcome of specific statistical tests. For normality, the Shapiro-Wilk test of normality (Shapiro & Wilk, 1965) was applied to the 100 normality samples, with an alpha level of .05. For homoscedasticity, Levene's test (Levene, 1960) at alpha level .05 served as an objective criterion. Thus, for example, if the Shapiro-Wilk test detected deviation from normal for a certain stimulus, and for the same stimulus the participant replied "Yes" to the question "Are these scores normally distributed?", then "1" was recorded. Participants' replies were recoded as preparation for further data analyses. Using Boolean algebra, their raw answers (rejecting or accepting a stimulus) could be recovered from the recorded 0-1 responses. Taking the respective test's judgment concerning the stimulus as a starting point (accepting or rejecting normality, that is homoscedasticity), and comparing whether the participant's answer agreed with the test's judgment (recorded as "0" if the participant agreed with the test and as "1" if the participant and the test disagreed), it could be determined whether the participant rejected or accepted normality or respectively, homoscedasticity. The recoding was performed for each participant and each stimulus. The trial number of each judging event was also recorded. Both rounds of the experiment were introduced by an introductory screen which asked participants to either read instruction 1 for normality or instruction 2 for homoscedasticity respectively. All participants completed the two construct rounds in the same order. The instructions were provided on two separate pieces of paper which remained covered until participants were prompted to read them. Instruction 1 and 2 contained fictive research scenarios in order to make the plotted data more tangible for participants. We anticipated that

providing participants with plausible research scenarios would help to minimize bias from confusion. Instruction 1 and 2 are included in Appendix A. Upon completion of the program, a .csv file was automatically saved for later analysis purposes. The file contained the following information per participant: participant ID, gender, age, nationality, study year, last known statistics grade (if applicable) and per trial, stimulus ID and whether the given answer was correct or not.

Statistical Game - Stimuli

Stimuli plots were generated on the basis of 200 simulated data sets. Our supervisor took over the task of simulating the data sets in R (R Core Team, 2015). Specifically, for normality, data sets were simulated by drawing from the Ex-Gaussian distribution. 100 samples were generated with fixed $\mu = 10$ and varied in the extent to which they were affected by the Gaussian component (σ) in relation to the exponential component (λ). λ produced skewness. σ varied between 1 and 4, λ between 0.25 and 1, and sample size between 20 and 200. The Ex-Gaussian distribution was chosen for simulation purposes as common psychological measures, such as response times, best fit the Ex-Gaussian distribution and thus, the chosen distribution is omnipresent in psychological research. For homogeneity of variance, 100 samples were created drawing from a linear model, featuring three groups with fixed means (μ), respectively 1, 3, and 4. Sample size varied between 20 and 80, but was balanced across groups. Residuals were normally distributed. A scale parameter (ϕ) was applied to the standard deviation, letting it vary with the mean: $\sigma_i = \sigma + \mu_i\phi$. This accurately reflects the typical empirical relation between sample mean and variance. ϕ varied between 0 and 1.5 and σ between 2 and 6. In effect, more pronounced heteroscedasticity emerged with increasing ϕ. The simulation parameters of the 100 normality and the 100 homoscedasticity samples are included in Appendix B.

On the basis of the 200 simulated data sets, 200 stimuli plots were created for the purpose of the study at hand. Our choice of plots was on the one hand inspired by the suggestions of Zuur et al. (2010) and on the other hand, by the kinds of plots our subjects were used to on the basis of their education at the psychology department of the University of Twente. However, we modified the suggested statistical graphics so that they additionally communicated sample size

in the normality plots and group size in the homoscedasticity plots. For normality, 100 dot-histograms were generated using R. On the x-axis "total score" was displayed with values between 10 and 50. Each dot in the graph represented a data point. This way an impression of sample size was included in the dot-histogram. The "ideal" accompanying each stimulus dot-histogram in the game was chosen out of the pool of available dot-histograms. The dot-histogram that visually came closest to being normally distributed, was chosen. A normal density curve was added to the "ideal" using R to facilitate recognition of a normal distribution in the plots. We included "ideals" in the experiment to facilitate decision making for participants, and to ensure the validity of our measurements. By providing an "ideal", participants' task was reduced to comparing two simultaneously presented graphs. They did not need to produce and maintain a mental image of an "ideal" against which our stimuli could be compared. Thereby, we could compensate for potential lack of sufficient statistical background knowledge. For homoscedasticity, 100 box-jitter plots were generated using R. Each plot consisted of three box-jitters, each representing one group. Group labels (Risky, Safer, Extremely cautious) were displayed on the x-axis. Each dot in the graph represented a data point. The "ideal" homogeneity plot was chosen out of the pool of available box-jitter plots. All stimuli plots and both ideals had a resolution of 420x300 pixels. During all trials, the stimulus was shown to the right of the ideal with its center at a length of 450 pixels and at a height of 50 pixels. The ideal's center was at a length of 250 pixels and at a height of 50 pixels in all cases respectively. Example stimuli for normality and homoscedasticity, as well as the ideal for normality and the homoscedasticity ideal are included in Appendix C. The complete R code used to simulate data, generate plots, and analyze results is included in Appendix D.

Design

The current study employed a complete within-subject experimental design. Manipulation occurred by exposing participants first to 100 dot-histograms and afterwards, to 100 box-jitter plots of 200 simulated data sets by means of a computer program. Presentation order of individual stimulus plots was randomized across participants. The accuracy and extent to which participants' judgments were informed by rational criteria was examined.

Procedure

All participants received the same briefing concerning the nature and purpose of the experiment. During data collection, it was not feasible to conduct the experiment at exactly the same place in all cases. However, all locations fulfilled the following requirements: isolation from other people, quietness, and a laptop to run the program. Locations included study places in the library of the UT, the laboratory of the behavioral sciences on campus of the UT, and the private homes of the conducting researchers. An appointment was made between researcher and participant. The conducting researcher was responsible for arranging a suitable experimental location. Upon arrival of the participant, he or she was greeted by the researcher and led to the experimental location. The participant sat down in front of the laptop with three pieces of paper lying on the table, two of them being covered. The experimenter asked the participant to read the provided research summary, that is, the cover story, and afterwards, whether the participant still had questions about the experiment. The provided cover story is included in Appendix E. When all questions of the participant were clarified, the participant signed the informed consent form. The program was opened and the experimenter filled in the participant's ID. The researcher informed the participant that he or she should let the researcher know in case of any problems with the software. The participant was then left alone in the room to complete the program.

Upon completion of the program, the experimenter rejoined the participant. It was first checked whether the participant's data was successfully saved by the program. The participant was then debriefed. Debriefing included revealing the true purpose of the experiment and clarifying participant questions about the experiment and program, if applicable. Finally, the participant was seen off.

Data Analysis

Data analysis was performed in two steps. First, during visual exploratory data analysis, the statistical outcomes (accept or reject normality/homogeneity of variance) of the respectively administered statistical test, and participants' judgements were compared with regard to the extent to which they were informed by objective criteria. For the normality data sets for the objective criteria were: sample size and the amount of skewness in the sample. For the

homoscedasticity data sets the objective criteria were: the amount of scale, relative to σ, and sample size. In a second step, a generalized linear mixed-model in the form of a logistic regression was built to further examine the influence of the above mentioned objective criteria on participants' judgement. Skew and sample size, that is scale and group size, were used to predict how likely a participant rejected normality or respectively, heteroscedasticity. We considered examining the interaction of skew, sample size and trial. That way we could have examined whether participants adjusted their usage of skew and sample size depending on the progress of the experiment, that is, whether participants' performance improved. Including trail number in the normality model led to model saturation. However, the interaction between scale, group size and trial number could be studied in an early version of the homoscedasticity model. Effects were minuscule which is why we pruned both models and trial number was ultimately removed as a predictor.

Results

Normality

EDA. Figure 1 and Figure 2 show the relation between the response (accept or reject normality) on the one hand, and sample size and skew on the other hand for both, the Shapiro-Wilk test of normality and the participants of the current study. As shown in Figure 1, for the Shapiro-Wilk test a clear pattern emerges. With increasing skew, the test correctly rejects normality. Sample size appears to be of little influence on the test's judgment.

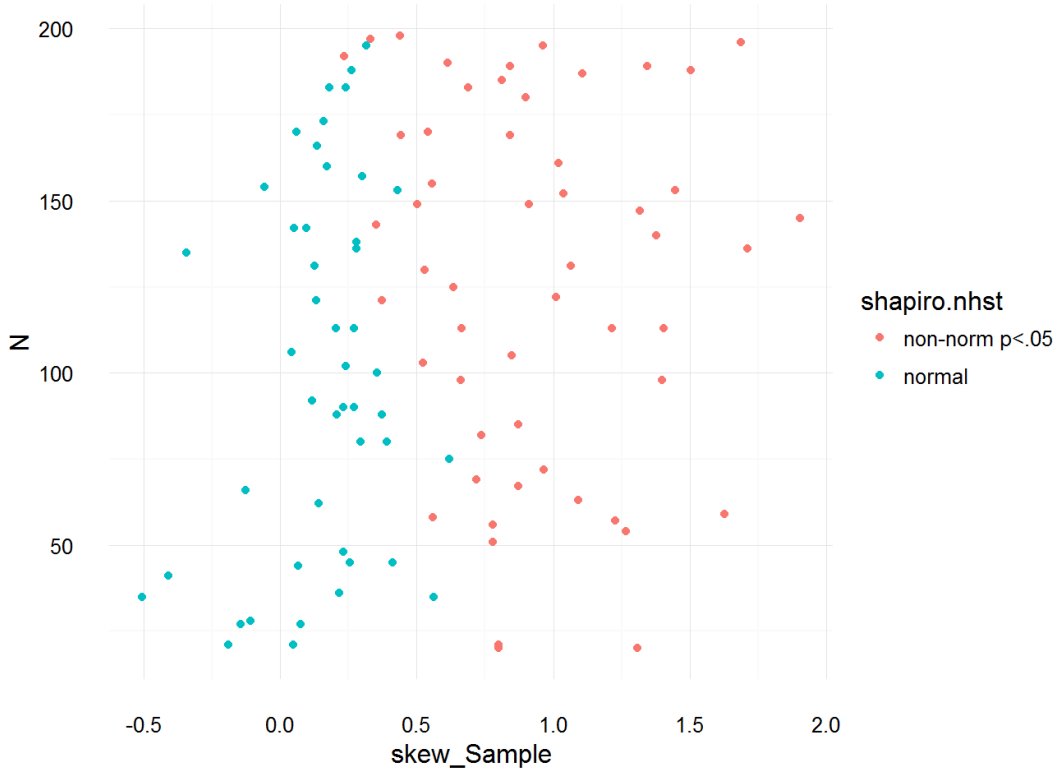

Figure 1. Association between statistical outcome (normal or non-norm p<.05), sample size (N) and skew in the sample for the Shapiro-Wilk test of normality.

Figure 2 shows the reject normality responses of each participant in relation to sample size and skew in the sample. The response pattern of participants is less clear than the test's outcome pattern, and shows great variation across participants. Apparently, participants had difficulties judging normality. Some participants even rejected samples with zero skew (see for example participant 19 and 21), others accepted samples with more than 1.0 skew (see for example participants 3, 14, and 15). However, the same pattern as found in Figure 1 is slightly visible in some participants, most clearly in participants 1 and 29, suggesting that at least to some extent, these participants made use of objective criteria.

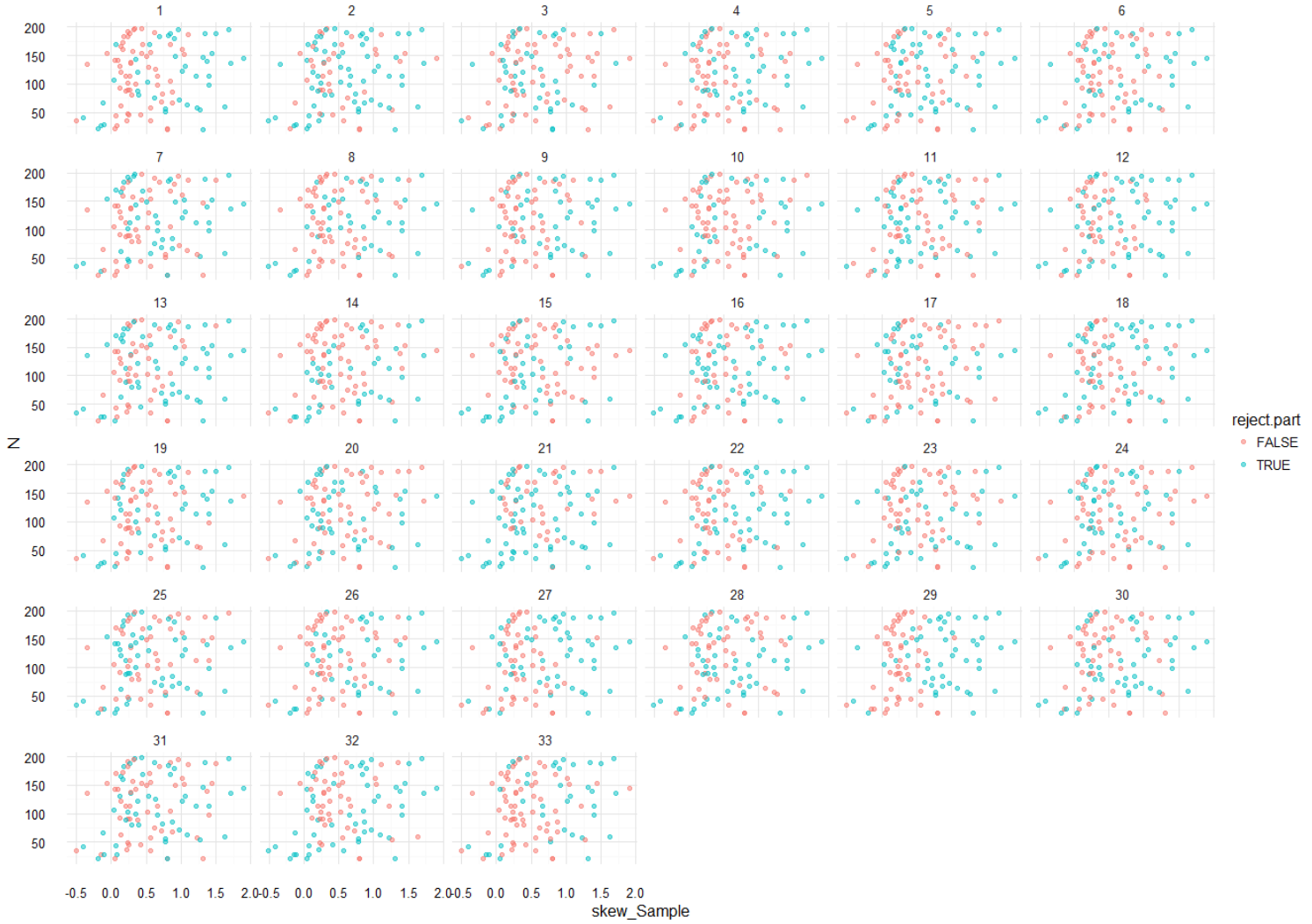

Figure 2. Association between reject normality response (TRUE or FALSE), sample size (N) and skew in the sample for each participant.

Logistic regression model. Table 1a shows fixed effects of skew in the sample, sample size (N), and the interaction effect of skew and sample size on the probability that participants reject normality. Values in Table 1a are on the logit scale.

Table 1a
*Fixed effects of skew, sample size, trial number and the skew * sample size interaction on participants' judgments*

Parameter	Point Estimate	Lower *	Upper*
Intercept	-.098	-.934	.661
Skew	.488	-.580	1.650
Sample size	-.007	-.012	.000
Skew*Sample size	.006	-.002	.014

*95% credibility limits

To obtain interpretable probabilities, the logit values first have to be transformed into linear predictors (η_i) using the following formula:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i \qquad (2)$$

Where η_i is the linear predictor of values x_{1i} and x_{2i}. x_{1i} can be any value for skew between 0 and 1, and x_{2i} can be any specific sample size. β_0 is the intercept value. The intercept in our model represents the situation that skew is 0 and N = 0. A more useful baseline rate for rejection rate would be at skew = 0 and the smallest sample size N = 10. β_1, β_2, and β_3 are the regression coefficients for respectively, skew, sample size and the interaction effect of skew and sample size. β_0, β_1, β_2, and β_3 are the logit values as shown in Table 1a. ε_i represents unknown random error. Once η_i is obtained, it is transformed into a probability μ_i using the following formula:

$$\mu_i = \frac{e^{\eta_i}}{(1+e^{n_i})} \qquad (3)$$

Where μ_i is the predicted probability that participants reject normality given η_i. μ_i can be used as a measure of impact by comparing how different values for skew and sample size affect the probability that participants reject normality. For example, the baseline probability that participants reject normality with skew = 0 and a sample size of N = 0 is calculated as follows:

1. Calculation of the linear predictor η_i

$$\eta_i = -0.098 - 0.488 * 0 - 0.007 * 0 + 0.006 * 0 * 0 = -0.098 \qquad (4)$$

2. Calculation of the probability μ_i

$$\mu_i = \frac{e^{-0.098}}{(1+e^{-0.098})} = 0.48 \qquad (5)$$

The baseline probability that participants reject normality is thus 0.48, almost chance level, which is logical as there is nothing to judge in the plot if N = 0. However, this can only be concluded with high uncertainty, 95% CI [-.934, .661]. Likewise, the adjusted baseline probability that participants reject normality when skew = 0 and N = 10 is 0.46, still close to chance level. Respectively, when skew = 1 and N = 10 the probability that participants reject normality is 0.59. Ideally, the probability that participants reject normality increases drastically with increasing skew. However, according to our model, participants tend to judge at chance level in both cases, when there is little skew in the sample, and when the sample is extremely skewed. The fact that the probabilities for the most extreme values for skew (0 and 1) are both fairly close to chance level shows that participants did not make proper use of the objective criterion. However, it has to be noted that there is considerable uncertainty concerning the impact of skew, 95% CI [-.580, 1.650]. As earlier revealed by visual data exploration, sample size is of little influence on participants judgments according to our model. Increasing sample size to N = 50 and keeping skew constant at skew = 0 yields a probability of 0.39 that participants reject normality, compared to a probability of 0.46 when skew is 0 and N = 10. This can be concluded with fair certainty, 95% CI [-.012, .000]. Moreover, there is only a miniscule interaction effect between skew and sample size. Disregarding the interaction effect when skew = 1 and N = 10 yields a probability of 0.58, compared to a probability of 0.59 when the interaction is taken into account. This can be concluded with fair certainty, 95% CI [-0.002, 0.014]. It makes sense that there is only a small interaction between skew and sample size if participants were neither able to make use of skew nor of sample size during their judgments.

Table 1b shows random effects of skew, sample size, the interaction between skew and sample size, and stimulus. Random effects show how much variation across units exist in the data set for a particular effect (Table 1a). Relevant units are the participant and stimuli. For instance, the large coefficient for participant intercept confirms what was earlier seen in the visual EDA: participants differ greatly in their base rejection rate. This means that some participants were more inclined to reject our stimuli than others to begin with. However, the 95%

CI [.355, 1.074] renders the estimate value of .680 considerably uncertain. The great coefficient for skew (.780) indicates that participants differed largely in how much their responses were influenced by skew in the sample. Being strongly influenced by skew is desirable, under the condition that skew is correctly interpreted. The large 95% CI [.188, 1.362] renders the estimate value of .780 considerably uncertain, though. The great coefficient for stimulus (.891) means that, even when skew and sample size are taken into account, stimuli still systematically differed in which response they provoked (rejecting or accepting normality). Apparently, they have additional unknown characteristics that either promote rejecting or accepting normality. The 95% CI [.737, 1.074] shows that the estimate value of .891 is considerably uncertain. Participants did not differ remarkably in how much they were influenced by sample size and the interaction between skew and sample size.

Table 1b
*Random effects of skew, sample size, the skew*sample size interaction, and stimulus*

Parameter	Point Estimate	Lower *	Upper*
Participant Intercept	.680	.355	1.074
Skew	.780	.188	1.362
Sample size	.002	.000	.006
Skew*Sample size	.004	.000	.009
Stimulus Intercept	.891	.737	1.074

*95% credibility limits

Homoscedasticity

EDA. Figure 3 and Figure 4 show the relation between the statistical outcome (accept or reject heteroscedasticity) on the one hand, and group size and amount of scale on the other hand for both, Levene's test and the participants of the current study. As shown in Figure 3, for the Levene test, a clear pattern emerges. With increasing scale, the test correctly accepts heteroscedasticity. Sample size has apparently little influence on the test's judgment.

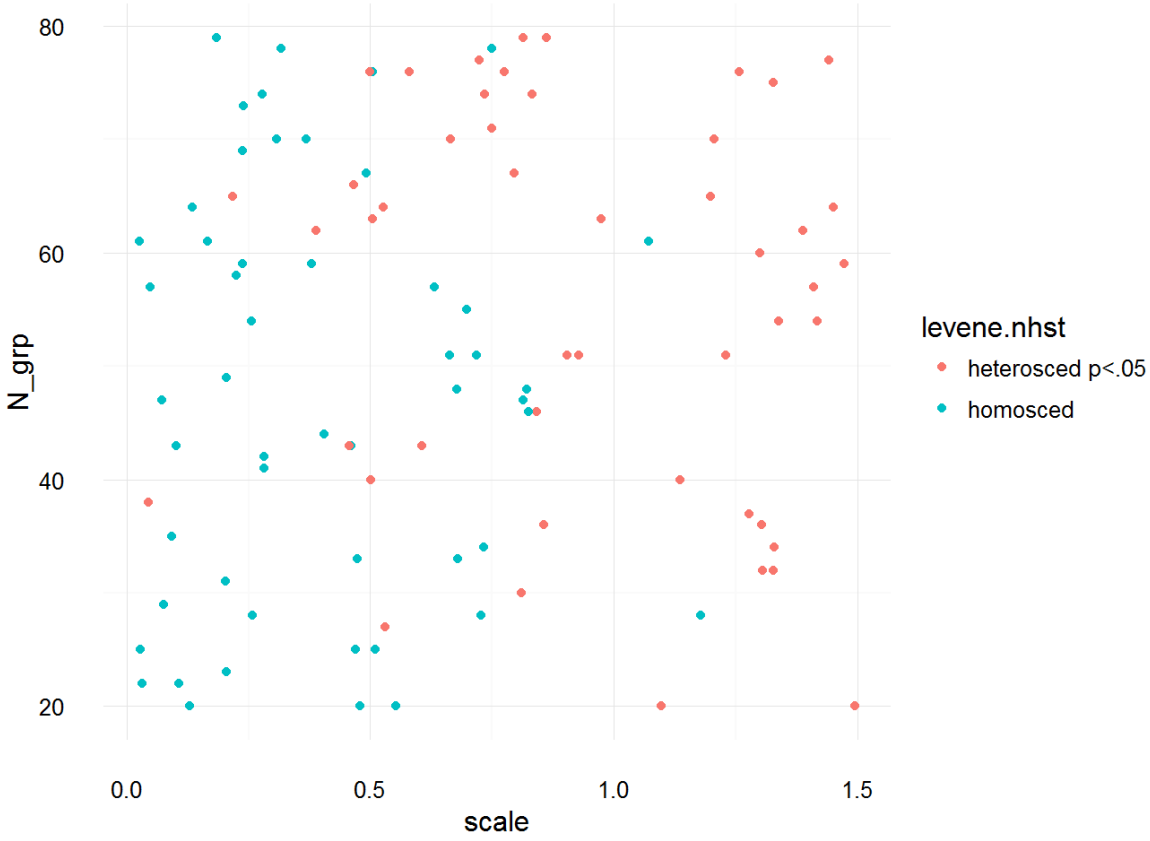

Figure 3. Association between statistical outcome (heterosced p<.05 or homosced), group size (N_grp) and amount of scale in the sample for Levene's test.

Concerning participants' responses (accept or reject heteroscedasticity), Figure 4 shows their reject heteroscedasticity responses in relation to sample size and amount of scale in the sample. There is no clear pattern in participants' judgements and large variation across participants. This variation refers to, for instance, a general tendency to accept or reject heteroscedasticity. For example, participants 22 and 27 seem to have a much higher general tendency to accept heteroscedasticity compared to the rest of the participants. Participants apparently had difficulties judging homoscedasticity. As for most participants responses are almost randomly scattered, regardless of the amount of scale in the sample or group size, it becomes evident that they made little use of objective criteria to support their judgements. There is one exception, though. Of all participants, the response pattern of participant four appears to be most informed by objective criteria.

Figure 4. Association between reject heteroscedasticity response (TRUE or FALSE), group size (N_grp) and amount of scale in the sample for each participant.

Logistic regression model. Table 2a shows fixed effects of scale, group size, and the interaction effect of scale and group size on the probability that participants reject heteroscedasticity. Values in Table 2a are on the logit scale. They are transformed in the same manner as values in Table 1a by first obtaining a linear predictor (η_i), and afterwards a measure of the probability that participants reject heteroscedasticity (μ_i). See Equation 2 and 3 for a detailed description of how η_i and μ_i are obtained. In Equation 2, the scale parameter replaces the skew parameter, group size replaces sample size, and the scale*group size interaction is included instead of the skew*sample size interaction. Also, unlike skew, scale varies between 0 and 1.5.

Table 2a
*Fixed effects of scale, group size and the scale * group size interaction on participants' judgments*

Parameter	Point Estimate	Lower *	Upper*
Intercept	.828	-.293	1.918
Scale	.560	-.945	2.165
Group size	-.031	-.053	-.010
Scale*Group Size	.014	-.018	.040

*95% credibility limits

Again, the intercept in our model represents the situation that scale is 0 and N_grp = 0. The baseline probability that participants reject heteroscedasticity is therefore 0.70. This is not logical as there is nothing to judge if N_grp = 0. Seemingly, our participants had a general tendency to reject our stimulus plots. A more useful baseline for rejection rate would be at scale = 0 and the smallest group size set to N_grp = 10. The suggested baseline probability that participants reject heteroscedasticity is 0.63, quite close to chance level. When scale is increased to 1.5 and group size is kept constant at N_grp = 10, the probability that participants reject heteroscedasticity is 0.83, according to our model. Again, this is worrisome. Participants apparently judge close to chance level when scale is low and reject heteroscedasticity when scale is high, which is ill-founded. Participants were not able to make good use of the objective criterion of scale. It has to be noted, though, that there is considerable uncertainty concerning the impact of the predictor scale, 95% CI [-.945, 2.165]. In fact, there is so much uncertainty that it is difficult to arrive at clear general conclusions. According to our model, group size is of medium influence on

participants' judgments. The probability that participants reject heteroscedasticity is 0.33 for scale = 0 and N_grp = 50, compared to a probability of 0.63 when scale = 0 and N_grp = 10. Participants apparently tend to accept heteroscedasticity with larger group sizes, independently of amount of scale. However, this can only be concluded with fair uncertainty, 95% CI [-.053, -.010]. Our model shows that there is only a weak interaction between scale and group size. For example, the probability that participants reject heteroscedasticity when scale = 1.5 and N_grp = 10 is 0.80 when disregarding the interaction effect between scale and group size, compared to a probability of 0.83 when the interaction is taken into account. This can be concluded with fair certainty, 95% CI [-.018, .040]. It makes sense that there is only a small interaction between scale and group size if participants were not able to make use of scale during their judgments.

Table 2b shows random effects of scale, group size, the scale*group size interaction, and stimulus. Random effects show how much variation across units exist in the data set for a particular effect (Table 2a). Relevant units are the participant and stimuli. According to our model, there is great variation between participants' tendency to reject heteroscedasticity. This is represented by the relatively large estimate value of .666. However, this can only be concluded with considerable uncertainty, 95% CI [.123, 1.178] The large value for scale (.746) means that participants differed greatly in how much they were influenced by scale. The large 95% CI [.127, 1.266], renders this conclusion quite uncertain, though. The considerable value of .672 for stimulus shows that, similar to the stimuli employed for normality, stimuli for homogeneity of variance differed in what response they provoked after taking scale and group size into account. There is less variation compared to the stimuli employed for normality, though. Still, this means that the stimuli for homoscedasticity have unknown characteristics that trigger a certain response (accept or reject heteroscedasticity). The 95% CI [.549, .831] sheds fair uncertainty on the estimate value of .672. Participants differed little in how much they were influenced by group size and the interaction of scale and group size.

Table 2b
*Random effects of scale, group size, the scale*groups size interaction, and stimulus*

Parameter	Point Estimate	Lower *	Upper*
Participant intercept	.666	.123	1.178
Scale	.746	.127	1.266
Group size	.015	.006	.025
Scale*Group size	.006	.000	.019
Stimulus intercept	.672	.549	.831

*95% credibility limits

Further Data Exploration

As mentioned above, our stimuli apparently had unanticipated characteristics that made them differ systematically in which response (accepting or rejecting normality, respectively heteroscedasticity) they provoked, even when the respective intended objective criteria were taken into account. Our results suggest that objective criteria were largely ignored by our participants, but the stimuli intercept random effects indicate that stimuli still differed systematically in how often they were rejected. Apparently, the answers of our participants were not completely random. Rather, instead of making decisions informed by objective criteria, participants appear to have made use of other, unknown "rules" to judge the stimulus plots. Presumably, participants fell back into heuristic decision making. In our further data exploration we try to identify these heuristics by comparing plots of high and low rejection rates. To that end, random effects on stimulus-level were extracted. They show how much a plot differs from the average rejection rate. Figure 5 shows the random effect intercepts and 95% CIs for normality stimuli, ordered by point estimate. Even though the estimates are quite uncertain, considerable variance is visible: stimuli vary in how often they were rejected. The precise stimulus random intercepts and 95% CIs for normality stimuli are included in Appendix F.

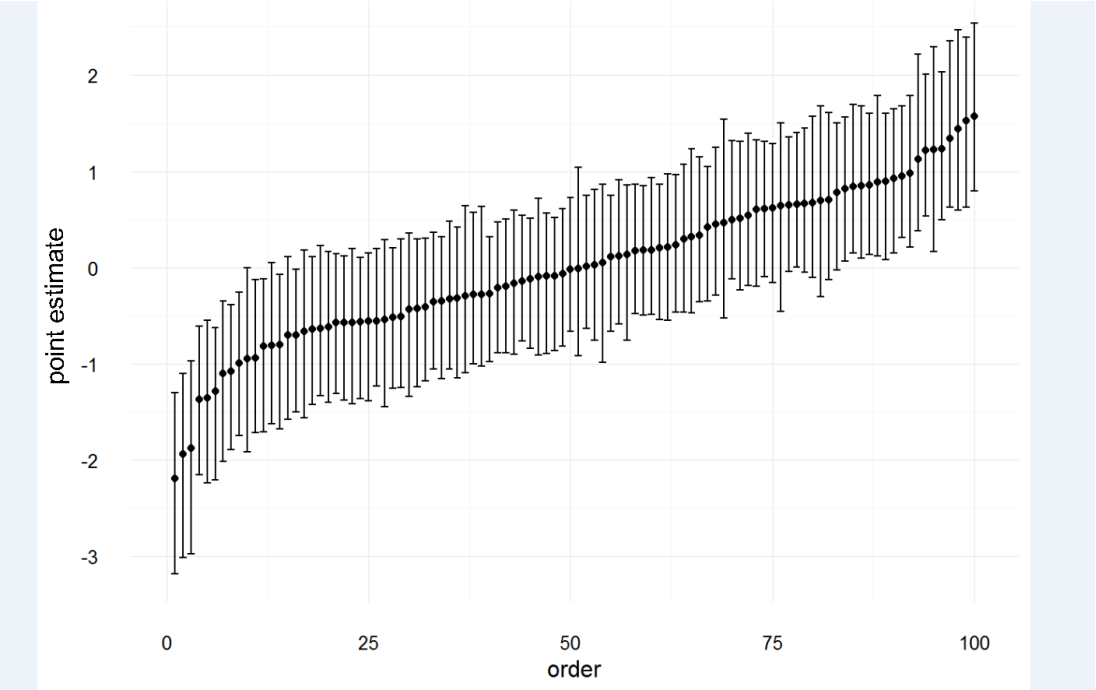

Figure 5. Normality stimuli random intercepts and 95% CIs, ordered by point estimate.

To examine properties that are related to high rejection rate every fifth normality stimulus plot ordered by rejection rate is shown in Figure 6.

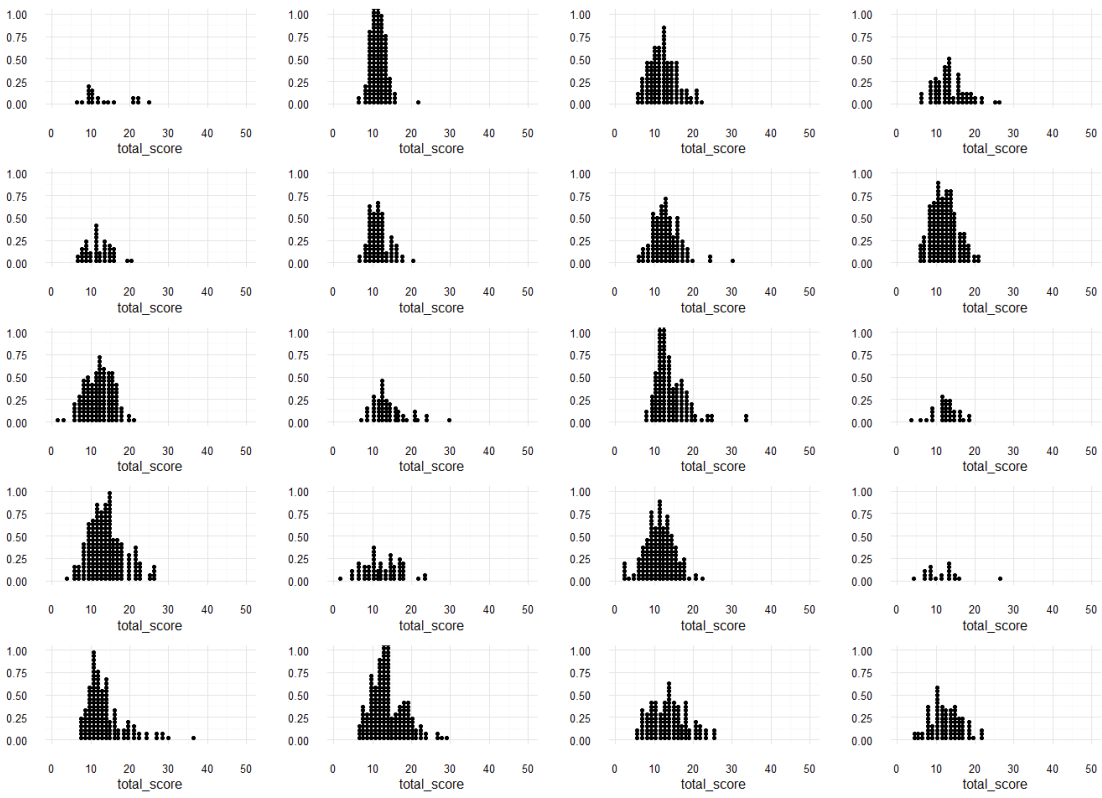

Figure 6. Every fifth normality stimulus plot ordered by rejection rate. At the top-left corner the stimulus with the highest rejection rate is displayed.

Apparently, there are no consistent properties that triggered a reject response. For instance, the first plot has strikingly few data points plus a visible gap between core observations and a few "outlier observations" on the right side. However, if these were heuristics our participants made use of, the fourth plot in the fourth row should have been rejected more often than our ranking reveals. Likewise, the second plot shows an extreme slope with a strikingly high peak. However, if this was a property by which our participants consistently judged the plots, the third plot in the third row should have been rejected more often. Ruggedness does not seem to be a property that is generally associated with a reject response. It seems as if participants judged each plot individually and neither made use of objective criteria nor consistent heuristics in the course.

Concerning the homoscedasticity plots, Figure 7 shows the random effect intercepts and 95% CIs for homogeneity of variance stimuli, ordered by point estimate.

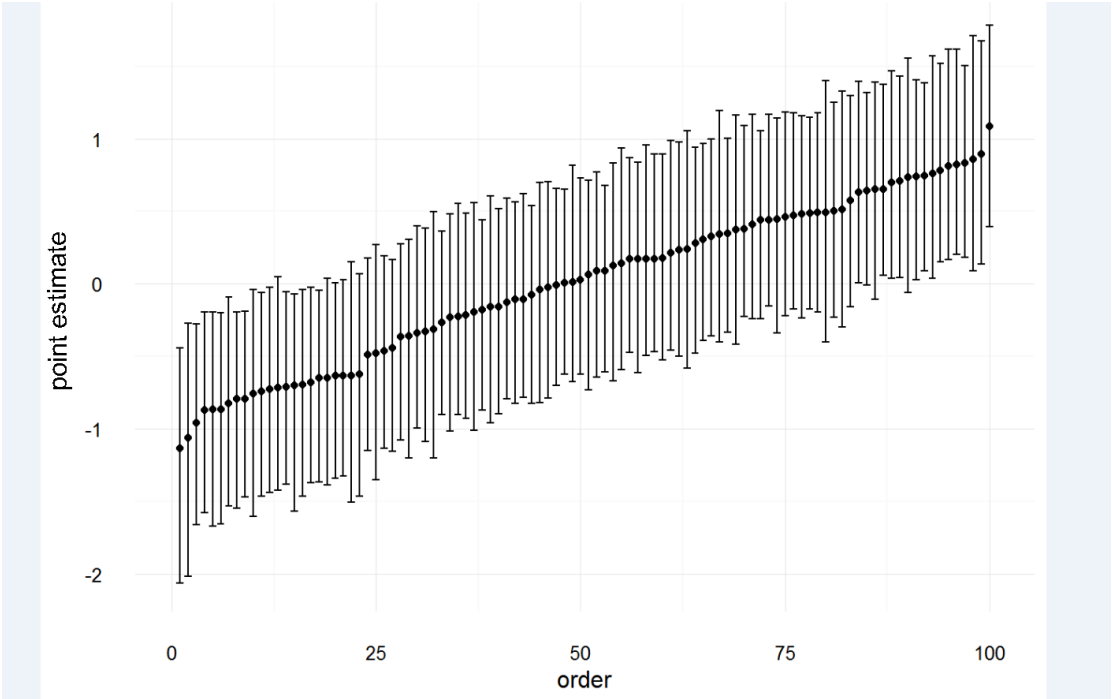

Figure 7. Homoscedasticity stimuli random intercepts and 95% CIs, ordered by point estimate.

Again, although estimates are uncertain, considerable variance is visible: stimuli systematically vary in how often they were rejected. The precise stimulus random intercepts and 95% CIs for homoscedasticity stimuli are included in Appendix F. As with the normality stimulus plots, every fifth box-jitter plot, ordered by rejection rate, is shown in Figure 8 to examine the heuristics employed by our participants.

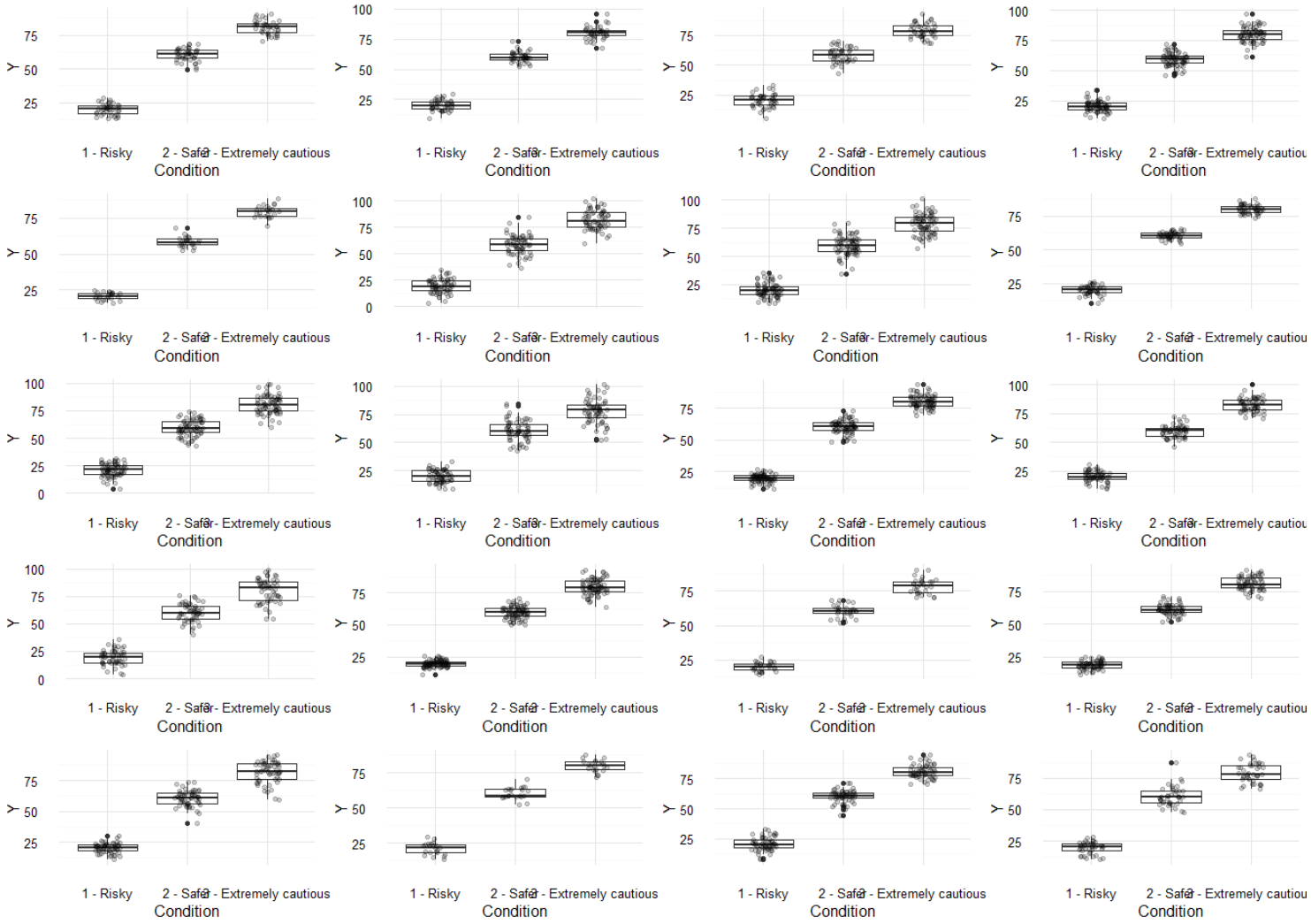

Figure 8. Every fifth homogeneity of variance stimulus plot ordered by rejection rate. At the top-left corner the stimulus with the highest rejection rate is displayed.

.

As with the normality plots, there are seemingly no consistent properties that led to a reject response. The equality of box length across groups, the location of the median line within the box, or the number of outliers may have served as cues. However, our ranking does not reveal a consistent reject pattern for these properties. The first plot of the third row and the second plot of the first row should both have been rejected more often, whereas the second plot of the last row should have been placed higher in the reject heteroscedasticity ranking if outliers, the position of the median, and equality of box length were consistent heuristics employed by our participants. Even a sophisticated combination of the above mentioned properties cannot account for the obtained rejection ranking.

Discussion

The current study questioned the faith advocators of graphical data exploration put in humans' visual detection abilities by testing these abilities experimentally. Specifically, the current study exposed participants to 100 dot-histograms and 100 box-jitter plots of 200 simulated data sets by means of a computer program. Participants indicated whether they accepted or rejected normality, respectively homoscedasticity, for each stimulus plot. Our participants were apparently not able to reliably and validly detect violations of the normality assumption and homoscedasticity visually. Furthermore, our participants did not make judgments informed by rational objective criteria. Skew and sample size served as objective criteria for the normality data sets, and scale and group size for the homoscedasticity data sets. Instead, our participants appeared to judge in a rather random and ill-founded fashion. Also, great variation between participants was revealed regarding their baseline rejection rate and in how much they were influenced by objective criteria, which makes drawing general conclusions difficult. However, this also leaves a spark of hope that under the general picture of poor performance some competent individuals are covered. The current study provides strong reason for concern regarding the question whether professionals can cope without statistical tests when examining statistical assumptions as suggested by Zuur et al. (2010). The question arises why the participants of the current study, all being students of the psychology department of the

University of Twente and thus, assumed to have had substantial experience with statistics, had so much difficulty making visual judgements on whether statistical assumptions were violated.

Limitations

One possible explanation lies in the formulation of the questions we asked the participants. In particular, for the normality data sets we asked "Are these scores normally distributed? (Y/N)". A better formulation may have been "Is this distribution asymmetrically skewed? (Y/N)" or "Are these scores randomly generated from a normal distribution? (Y/N)". The normality assumption refers to residuals and not to raw scores, as suggested by the question we asked. However, there were no questions and no remarks on the part of our participants concerning what exactly should be normally distributed. This suggests that the ambiguity we left our participants with remained unnoticed. Some other limitations mentioned were initial unclarity as to which of the two plots presented during trials was the ideal and which was the stimulus plot. Furthermore, it was criticized that the normal density curve in the ideal for normality was cut off at the y-axis intercept, which does not accurately represent a normal density curve. All mentioned limitations of the current study potentially confused participants, thereby distorting their response pattern. However, our results suggest that regardless of the above mentioned limitations, our participants performed poorly.

Choice of Statistical Graphics

Another possible explanation for why our participants had so much difficulty judging the plots lies in the choice of the administered plots. Indeed, random effects revealed that the stimulus plots for both, normality and homoscedasticity, differed greatly from each other in the kind of response they provoked. Our stimulus plots apparently had unknown characteristics that provoked a certain response, even after the respective objective criteria had been taken into account. However, when we examined random effects on individual stimulus level, that is, how much a stimulus differed from the average rejection rate, there were no clear properties visible that might be associated with a reject response. The results suggest that our participants did not only fail to inform their decisions by objective criteria, but also, that they did not even make use

of consistent heuristics. However, it has to be mentioned that the performed analyses do not allow any conclusion on whether individual participants did make use of consistent heuristics. The consistent use of heuristics by individual participants may have been lost in averaging rejection rates over participants. Thus, in future research it would be more productive to examine judgement heuristics on the level of the individual participant to obtain an understanding of how participants judge stimuli plots. This is in accordance with what Molenaar and Campbell (2009) call a necessary "Kuhnian paradigm shift" (p.1) from interindividual to intraindividual variation analyses whenever person-specific psychological processes, such as information processing and perception, are the subject of interest.

As mentioned above, our plots turned out to have some shortcomings. The question arises what kind of plots are optimal for visually testing for violations of statistical assumptions in general, and for the assumptions of normality and homogeneity of variance in particular. As mentioned above, there is little clarity in the relevant scientific literature when it comes to the question of how statistical graphics should best be designed and administered. As Johnson (2004) notes, the visualization community still relies too much on ad hoc techniques and rules of thumb. In the current study, the suggestions of Zuur et al. (2010) were chosen as a guidance. Our results question the appropriateness of the suggested graphics. However, there are plenty of alternatives. For instance, the violin plot is a useful graphical technique that has received considerable attention in recent years (Hintze & Nelson, 1998; Marmolejo-ramos & Valle, 2009). The core feature of the violin plot is that it shows the same information about the center, spread, asymmetry, and outliers of a variable as a boxplot, and additionally, a smoothed histogram, a density estimate, also providing information about the shape in a single plot. An example of a violin plot is included in Appendix G. The single plot structure facilitates comparison of the distributions of several variables (Hintze & Nelson, 1998). However, according to Haughton and Haughton (2011), violin plots reduce the impact of an otherwise well-designed box plot and lack the ease of interpreting a horizontal kernel density. Another frequently used technique for checking the normality assumption is the Q-Q plot (quantile-quantile plot). Essentially, when a Q-Q plot is used to check for normality, observed values are plotted against the expected values

the scores should have if the data was normally distributed. If the data is indeed normally distributed then the observed scores will have the same distribution as the predicted scores and the Q-Q plot shows a straight diagonal. Deviations from the diagonal indicate deviation from normal (Field & Field, 2012). For instance, Figure 9 shows the normal Q-Q plot of one of the simulated normality samples of the current study, S01_1.

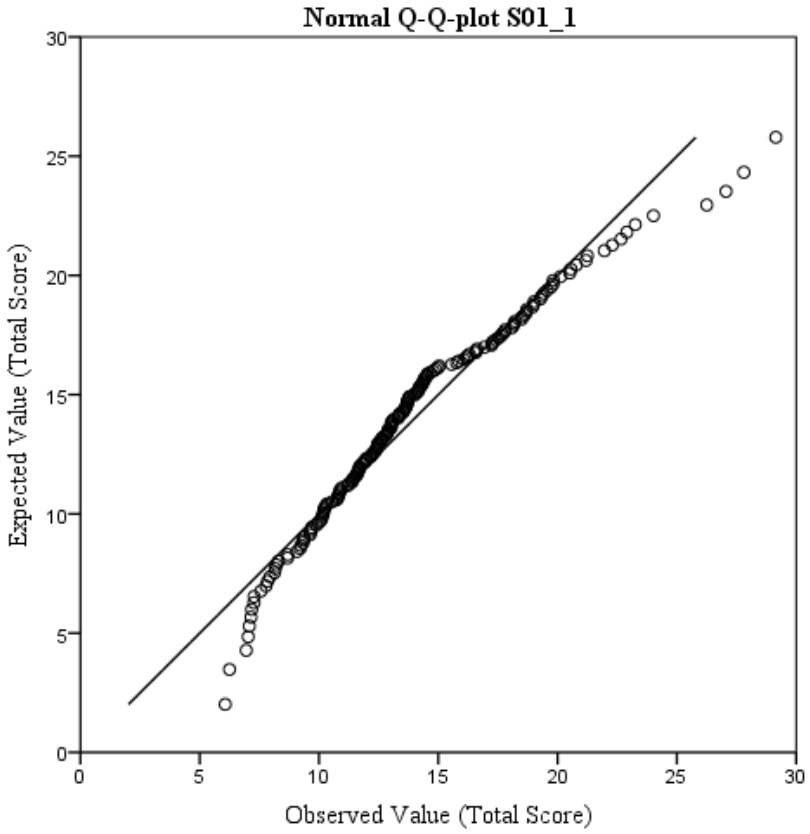

Figure 9. Normal Q-Q plot of one of the simulated normality samples of the current study (S01_1). Observed values of Total Score are on the x-axis and expected values of Total Score given normality on the y-axis.

Although Q-Q plots give a clearer picture of whether a certain set of data is normally distributed or not, they are conceptually and graphically less intuitive. Indeed, the important point is that whatever graphical technique is to be preferred should be decided from the viewer's point of view. Statistical graphics are only useful if the viewer is able to extract the relevant information with ease. Thus, the viewer and her/his visual and theoretical capacities have to be taken into account (Chen, 2005; Johnson, 2004). The research of Cleveland and McGill (1984) on elementary perceptual tasks is a good starting point, but provides little practical value when it comes to selecting one technique out of the many available and commonly used techniques.

Empirical support for the usefulness of existing graphical techniques is desirable Johnson (2004). In the context of evaluating competing graphics, Tukey 1990) proposed a method to examine which of two or more graphics best shows a "phenomenon". A phenomenon is a potentially interesting thing that can be described non-numerically. There are three phases in Tukey's suggested experimental design. In phase one, each participant receives information about a fair amount of likely phenomena and what they look like in each style of presentation, thus, in each graphical technique being tested. In the second phase, participants are briefly exposed to a data visualization using one style of presentation applied to one set of data. Participants are asked whether a certain phenomenon was present in the visualization. Presence and nonpresence, as well as the different phenomena should be balanced and randomized across trials. It should be recorded whether the participant gave a correct or incorrect answer. Finally, in phase three, the data of phase two is analyzed. The time of graphic display for 90% right should be determined for each presentation style. Then, time is to be compared across data sets and style. The graphical technique in which 90% right is seen most rapidly is to be preferred. Tukey encourages modification and extension of his suggested experimental design.

Criticism of Psychological Education

Apart from the need to quantify the effectiveness of competing graphical techniques in future research, Nolan and Perrett (2015) advocate the early incorporation of statistical graphics into the undergraduate statistical curriculum. According to the authors, creating informative statistical graphics can be rewarding for students at all levels of proficiency as it emphasizes statistical thinking over calculations. Also, the creative expression of statistical findings may help to overcome the difficulties inherent in learning computational thinking. Nolan and Perrett (2015) developed several assignments that exemplify possibilities for incorporating graphics into curricula in a pedagogically meaningful way. Among these assignments are: deconstructing and reconstructing plots, converting tables into graphics and copying expert graphs. Indeed, the above mentioned considerations on the limitations of the current study and the choice of statistical graphics are based on the premise that the participants of the current study are sufficiently proficient in statistics. However, the experimenters expressed their doubt about

participants' competence. During debriefing of the experiment, experimenters frequently received basic statistical questions. For instance, participants asked what was meant by "homogeneity of variance", a term that should be well known to any student in the psychology department. Also, participants reported that they experienced judging our stimuli plots to be extremely difficult. Indeed, as mentioned earlier, the reviewed expertise research (Lesgold, 1983; Loveday et al., 2013) stressed the importance of domain knowledge for expert performance. A lack of relevant domain knowledge will prevent subjects from performing well on pattern recognition tasks as featured in our experimental setup. To rule out the possibility that our results simply reflect incompetence on behalf of our participants, future research should include professional statisticians and researchers in the subject pool. Generally speaking, the quality and competence of the methodological and statistical education of psychology departments and psychological researchers in practice have received much criticism in recent years (e.g. Aiken et al., 1990; Hoekstra, Morey, Rouder, & Wagenmakers, 2014). Aiken et al. (1990) conducted a detailed survey to assess the extent to which advances in statistics, measurement, and methodology had been implemented into the doctoral training of all PhD programs in psychology offered in the United States and Canada. They revealed that the methodological curriculum had advanced little in the previous 20 years, despite major advances in statistics, methodology, and measurement. Aiken et al. (1990) describe three major costs of not being familiar with advances in the field. A first cost is failure to utilize designs that optimize the tests of theory. A second cost is failure to gather relevant data that would be most suitable for answering a research question at hand. And thirdly, another cost is failure to draw correct conclusions due to the misanalysis of data. The authors stress that serious errors will increasingly flaw the psychological literature if researchers do not become better educated in methodology. Aiken et al. (1990) plead for the reformation of the methodological curriculum to prevent the current situation from further deteriorating. After all, today's students are tomorrow's researchers. Specifically, the authors' considerations include the retraining of faculty members, a stronger undergraduate preparation in mathematics and statistics, and the individual tailoring of statistical, methodological, and measurement skills to the needs of students in their specific

substantive areas. A second major criticism of the methodological and statistical education in the field of psychology is formulated by Hoekstra et al. (2014). The authors examined the interpretations of confidence intervals (CIs) of first-year, and graduate psychology students, and researchers in the field of psychology. Specifically, participants were given six particular statements involving different interpretations of a single CI, and asked to rate each statement as either true or false. Though all six statements were false, both, researchers and psychology students endorsed on average more than three statements, exposing their lack of understanding of CIs. Remarkably, neither self-declared experience with statistics nor having undergone substantial statistical training protected researchers and students alike from endorsing false statements. Hoekstra et al. (2014) stress the lack of understanding of CIs among psychological professionals and the apparent difficulty of the concept, especially since CIs are among the main tools by which psychologists draw inferences from data. In sum, the reviewed critique of the methodological and statistical education in the field of psychology is two-folded. Aiken et al. (1990) revealed a reluctance among American and Canadian psychology departments to incorporate advances in methodology, statistics, and measurement into their doctoral trainings, whereas Hoekstra et al. (2014) exposed considerable drawbacks in the statistical competence of both, psychology students and researchers in practice. This is worrisome.

Implications and Future Research

Zuur et al. (2010) advocated the use of statistical graphics in data exploration. Our research objective was to confirm that practitioners of research statistics are actually able to detect violations of statistical assumptions visually. Unfortunately, on the basis of the obtained results we have to conclude "No, they cannot detect violations of statistical assumptions visually.". So, what does this insight leave us with? Apparently, there is need for more research into the following subjects: First of all, as earlier advocated by Tukey (1990) and Johnson (2004), the effectiveness of competing statistical graphics needs to be established empirically once for all. We suggest using Tukey's (1990) method as a starting point to this end. One could, for instance, select a number of graphics suitable for checking normality, such as Q-Q plots, kernel densities (Marmolejo-ramos & Valle, 2009) and histograms, and examine experimentally what graphic

yields the most correct detections of violations in the shortest amount of time. Secondly, the apparent lack of basic statistical competence of our participants and the criticism on the curriculum of psychology departments in general, indicates that students should be taught methodology and statistics differently. This is crucial if the field of psychology does not wish to end up with a bunch of poorly educated researchers in the near future. Research should dig into how to teach students methodology and statistics properly. Nolan and Perrett (2015) provide suggestions regarding the incorporation of statistical graphics into the curriculum, teaching them the proper use of them and promoting statistical thinking. Another option would be to dig deeper into what constitutes good feedback in an experimental setup comparable to the one employed in the current study. Our feedback was not directed at learning how to judge our stimulus plots correctly. Feedback directed at insight instead of a short comment on whether a given reply was correct or not may promote learning and would yield valuable insights into how students can be taught how to correctly interpret statistical graphics. Given the great merits visualization holds for data exploration and data analysis this appears to be a promising course of action. In the meantime, one could settle for a joint usage of graphical and statistical techniques as advocated by some (Behrens & Yu, 2003; Marmolejo-ramos & Valle, 2009). After all, contrary to common criticism of null hypothesis testing techniques (e.g. Läärä, 2009), the statistical tests employed in the current study displayed considerable robustness against fluctuations in sample size. Conventional tests may still hold (temporary) merits for assumption checking. Once there are sufficient researchers who received proper education in (the application and interpretation of) statistical graphics, the scientific community may finally be ready for a methodological shift from statistical techniques to the use of statistical graphics as the main tool for data exploration. Fact is, we are not ready for it yet.

Conclusion

In conclusion, the results of the current study provide reason for serious concern regarding the abolishment of confirmatory statistical techniques for examining violations of statistical assumptions as advocated by Zuur et al. (2010). We plead for a stronger concern for actively tackling visualization problems (Chen, 2005; C. Johnson, 2004), partly through the reformation

of the current psychological curriculum (Nolan & Perrett, 2015). Furthermore, the improvement of psychological training in methodology, statistics, and measurement emerged as an important end in itself since the participants of the current study, all being students of the psychology department of the University of Twente, displayed devastating competence in statistics. Taking current drawbacks in the design and application of statistical graphics, the competence of students and researchers in psychology, and the advantages of both, statistical techniques and visual EDA, into account, we advocate a temporary joint usage of confirmatory and graphical exploratory techniques in psychological research. In addition, psychological curricula have to undergo a reformation and should incorporate and emphasize statistical graphics early in their training. If no measures are being taken the quality and reproducibility of psychological research will further deteriorate in the near future.

References

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., Scarr, S., … Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, *45*(6), 721–734. http://doi.org/10.1037/0003-066X.45.6.721

Behrens, J., & Yu, C. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology Volume 2: Research Methods in psychology* (pp. 33–64). New Jersey: John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0471264385.wei0202/full

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. http://doi.org/10.1016/0010-0285(73)90004-2

Chatfield, C. (1985). The initial examination of data. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *148*(3), 214–253. Retrieved from http://www.jstor.org/stable/2981969

Chen, C. (2005). Top 10 unsolved information visualization problems. *IEEE Computer Graphics and Applications*, *25*(4), 12–16. http://doi.org/10.1109/MCG.2005.91

Cleveland, W. S. (1984). Graphs in Scientific Publications. *The American Statistician*, *38*(4), 261–269. http://doi.org/10.2307/2683400

Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, *79*(387), pp. 531–554. http://doi.org/10.2307/2288400

Cowles, M., & Davis, C. (1982). On the origins of the. 05 level of statistical significance. *American Psychologist*, *37*(5), 553–558. Retrieved from http://psycnet.apa.org/journals/amp/37/5/553/

Ehman, R. L., Hendee, W. R., Welch, M. J., Dunnick, N. R., Bresolin, L. B., Arenson, R. L., … Thrall, J. H. (2007). Blueprint for imaging in biomedical research. *Radiology*, *244*(1), 12–27. http://doi.org/10.1148/radiol.2441070058

Feliciano, G., Powers, R., & Kearl, B. (1963). The presentation of statistical information. *Educational Technology Research and Development*, *11*(3), 32–39. Retrieved from http://www.springerlink.com/index/R687U7G22X0W6378.pdf

Field, A. (2009). *Discovering Statistics Using SPSS*. *Sage Publication* (Vol. 58). London: Sage. http://doi.org/10.1234/12345678

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. *Statistics*. London: Sage. http://doi.org/10.1111/insr.12011_21

Gelman, A. (2011). Why Tables Are Really Much Better Than Graphs. *Journal of Computational and Graphical Statistics*, *20*(1), 3–7. http://doi.org/10.1198/jcgs.2011.09166

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's Practice What We Preach: Turning tables into graphs. *The American Statistician*, *56*(2), 121–130. http://doi.org/10.1198/000313002317572790

Grace-Martin, K. (2012). Checking the Normality Assumption for an ANOVA Model. Retrieved May 9, 2016, from http://www.theanalysisfactor.com/checking-normality-anova-model/

Groot, A. de. (1965). *Thought and choice in chess*. The Hague: Mounton.

Groot, A. de. (1966). Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.), *Problem solving*. New York: Wiley.

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis*. Newbury Park, CA: Sage.

Haughton, D., & Haughton, J. (2011). Graphical Methods. In *Living Standards Analytics* (pp. 1–22). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4614-0385-2_1

Hintze, J. L., & Nelson, R. D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, *52*(2), 181–184. http://doi.org/10.1080/00031305.1998.10480559

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 1–7. http://doi.org/10.3758/s13423-013-0572-3

Howell, D. (2012). *Statistical methods for psychology*. Pacific Grove, CA: Duxbury/Thomson Learning.

Johnson, C. (2004). Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, *24*(4), 13–17. http://doi.org/10.1109/MCG.2004.20

Johnson, D. H. (2009). Statistical Sirens : The Allure of Nonparametrics. *USGS Northern Prairie Wildlife Research Center*, *76*(6), 1998–2000.

Klein, G., & Hoffman, R. (1992). Perceptual-cognitive aspects of expertise. In M. Rabinowitz (Ed.),

Cognitive science foundations of instruction (pp. 203–226). Mahwah, NJ: Eribaum. Retrieved from http://cmapsinternal.ihmc.us/rid=1217527241618_235135118_3994/Seeing the Invisible-1992.pdf

Kline, R. (2008). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York: Guilford Press.

Läärä, E. (2009). Statistics: reasoning on uncertainty, and the insignificance of testing null. *Annales Zoologici Fennici*, *46*(2), 138–157. Retrieved from http://www.bioone.org/doi/abs/10.5735/086.046.0206

Lesgold, A. (1983). *Acquiring expertise*. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA124876

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in Honor of Harold Hotelling* (pp. 278–292). Stanford: Stanford University Press.

Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology*, *46*(4), 684–688. http://doi.org/10.1016/j.jesp.2010.02.007

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance unkown. *Journal of the American Statistical Association*, *62*(318), 399–402. http://doi.org/10.1080/01621459.1967.10482916

Loveday, T., Wiggins, M., & Festa, M. (2013). Pattern recognition as an indicator of diagnostic expertise. *Pattern Recognition - Application and Methods*, *204*, 1–11. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-36530-0_1

Marmolejo-ramos, F., & Valle, U. (2009). Getting the most from your curves : Exploring and reporting data using informative graphical techniques. *Tutorials in Quantitative Methods for Psychology*, *5*(2), 40–50.

Meyer, J., Shamo, M., & Gopher, D. (1999). Information structure and the relative efficacy of tables and graphs. *Human Factors*, *41*(4), 570–587. Retrieved from http://hfs.sagepub.com/content/41/4/570.short

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *65*(2), 81–97. Retrieved from http://psycnet.apa.org/journals/rev/63/2/81/

Nolan, D., & Perrett, J. (2015). Teaching and Learning Data Visualization: Ideas and Assignments, *80639*, 7–8. Retrieved from http://arxiv.org/abs/1503.0781

Regehr, G., Cline, J., Norman, G., & Brooks, L. (1994). Effect of processing strategy on diagnostic skill in dermatology. *Academic Medicine*, *69*(10), 34–36. Retrieved from http://journals.lww.com/academicmedicine/Abstract/1994/10000/Effect_of_processing_strategy_on_diagnostic_skill.34.aspx

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591–611. http://doi.org/10.1093/biomet/52.3-4.591

Shiffler, R. E. (1988). Maximum Z Scores and Outliers. *The American Statistician*, *42*(1), 79–80. http://doi.org/10.1080/00031305.1988.10475530

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Team, R. (2015). R: A language and environment for statistical computing . Vienna, Austria: R Foundation for Statistical Computing; 2013. *Freely Available on the Internet at: Http://www. R-Project.*

Tufte, E. R. (2007). *The visual display of quantitative information* (2nd ed., Vol. 16). Cheshire, Connecticut: Graphics Press Cheshire, CT.

Tukey, J. (1990). Data-based graphics: visual display in the decades to come. *Statistical Science*, *5*(3), 327–339. Retrieved from http://projecteuclid.org/euclid.ss/1177012101

Tukey, J. W. (1977). Exploratory Data Analysis. In *Exploratory Data Analysis* (1st ed., pp. 5 – 23). Upper Saddle River, NJ: Prentice Hall. http://doi.org/10.1007/978-1-4419-7976-6

Washburne, J. (1927). An experimental study of various graphic, tabular, and textual methods of presenting quantitative material. *Journal of Educational Psychology*, *18*(7), 465–476.

Waters, A., Underwood, G., & Findlay, J. (1997). Studying expertise in music reading: Use of a pattern-matching paradigm. *Perception & Psychophysics*, *59*(4), 477–488. Retrieved from http://link.springer.com/article/10.3758/BF03211857

Wilcox, R. R. (2004). Kernel Density Estimators: An Approach to Understanding How Groups Differ. *Understanding Statistics*, *3*(4), 333–348. http://doi.org/10.1207/s15328031us0304_7

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. http://doi.org/10.1037/0003-066X.54.8.594

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. http://doi.org/10.1111/j.2041-210X.2009.00001.x

Appendix A

Instruction 1

As part of a nationwide online-survey on student satisfaction with the services provided by their higher educational institutions, students were asked to evaluate the library of their institution. This was done by means of a 10 item questionnaire. Example items included “The last time I asked for help, the librarians working at the library were able to answer my questions competently.”, and “The last online catalogus reservation I made was processed in due time.” For each item, the participants replied by marking their preference on a 5-point-Likert-scale (1 = completely unsatisfactory, 2 = partly unsatisfactory, 3 = neutral, 4 = partly satisfactory 5 = completely satisfactory). The obtained answers of the participants yielded one total score per participant on the scale.

The obtained data was read into an spss file.
‘Higher educational institution’ was added as a grouping variable to distinguish samples. Each sample represents the students of one specific educational institution.

As part of data exploration prior to conducting statistical analyses on the data, you take a look at how total scores are distributed in the samples. The following graphs show the distribution of participant total scores. Each graph shows a specific sample, thus the total scores of the students of a specific educational institution on the questionnaire.

We would like you to answer the following question per graph presented:

Are the total scores normally distributed?

Press <y> on the keyboard for “yes”.

Press <n> on the keyboard for “no”.

For your convenience, each sample graph will be accompanied by a graph of an ideal normal distribution. You may refer to this “ideal” as a means for comparison.

Also, there is no need to think long before answering. Your intuitive answer will usually be the best one.

There will be 5 practice trials. Upon completion of the practice trials, your score will be reset to 0 and the actual game begins.

Press <ENTER> to start the practice trials

Instruction 2

A questionnaire has been sent to a randomized sample of car drivers. They were asked, among other questions, how they would rate their own driving style (risky, safer, extremely cautious). They were also asked how close they pull up to cars that braked when driving on a highway before stopping or steering around (numerical in meters).

The data has been transformed into a data file for SPSS. Per group of drivers (risky, safer, extremely cautious) it has been examined more closely how far they stay away from other drivers when braking on a highway.

Imagine you want to check with an Analysis of Variance-method if there is an effect of self-reported driving style on the space they keep between themselves and other drivers.

In this case, you need to check whether the data fulfills the assumption of homogeneity of variance. You will do that with the help of the following box-jitter plots. The dots in the graphs represent data points. The following 100 graphs are possible representations of the aforementioned data.

We would like you to answer the following question per graph presented:

Are the variances homogenous?

Press <y> on the keyboard for “yes”.
Press <n> on the keyboard for “no”.

For your convenience, each sample graph will be accompanied by a graph of ideal homogeneity of variance. You may refer to this “ideal” as a means for comparison.

Also, there is no need to think long before answering. Your intuitive answer will usually be the best one.

There will be 5 practice trials. Upon completion of the practice trials, the actual game will continue (i.e. during the trials your score will be frozen).

Press <ENTER> to start the 5 practice trials.

Appendix B

Table 3

Simulation parameters of the 100 samples in the normality data set (S01)

Stimulus	N	Mu (μ)	Sigma (σ)	Lambda (λ)	Skew sample	Skew population
S01_1	185	10	2,878736	0,273571	0,812284	0,969772
S01_2	189	10	1,651473	0,391952	1,343313	1,183202
S01_3	72	10	1,649702	0,281231	0,966427	1,492908
S01_4	169	10	2,166835	0,429482	0,842331	0,784597
S01_5	136	10	3,827367	0,678592	0,280207	0,092779
S01_6	113	10	3,887824	0,429743	0,202992	0,270908
S01_7	153	10	3,219566	0,2562	1,446748	0,918159
S01_8	44	10	3,199738	0,407537	0,066133	0,450688
S01_9	138	10	2,607284	0,56903	0,278418	0,349199
S01_10	147	10	1,006819	0,562091	1,316279	1,318365
S01_11	102	10	2,826812	0,3808	0,240919	0,630561
S01_12	149	10	3,510405	0,339026	0,501082	0,532454
S01_13	188	10	3,254568	0,497696	0,262385	0,289935
S01_14	66	10	2,358195	0,845416	-0,12656	0,180254
S01_15	103	10	2,60737	0,424821	0,522726	0,601828
S01_16	189	10	2,61213	0,284393	0,841175	1,034553
S01_17	196	10	1,004143	0,367119	1,688208	1,652047
S01_18	41	10	2,066998	0,485425	-0,4094	0,703539
S01_19	105	10	2,836399	0,378461	0,848321	0,633383
S01_20	121	10	3,486826	0,271875	0,373522	0,764462
S01_21	183	10	2,070166	0,404863	0,687506	0,900351
S01_22	45	10	2,231905	0,660111	0,412704	0,354252
S01_23	198	10	2,720428	0,380359	0,437793	0,67121
S01_24	190	10	2,769035	0,25744	0,61302	1,079829
S01_25	35	10	3,158972	0,515185	-0,50501	0,286972
S01_26	113	10	2,184919	0,288888	1,214952	1,209422
S01_27	90	10	3,757612	0,520518	0,269276	0,188672
S01_28	183	10	3,887711	0,642525	0,240861	0,10267
S01_29	100	10	1,700571	0,873332	0,354775	0,348453
S01_30	170	10	3,173493	0,575698	0,059535	0,221371
S01_31	153	10	3,710904	0,487016	0,429385	0,226968
S01_32	166	10	2,810422	0,677017	0,134388	0,201385
S01_33	90	10	2,894522	0,522932	0,229631	0,334979
S01_34	143	10	3,812158	0,949988	0,350959	0,037713
S01_35	21	10	3,551448	0,250648	0,800552	0,833452
S01_36	170	10	2,739463	0,292982	0,540507	0,948643
S01_37	21	10	3,464212	0,793813	-0,19143	0,079828
S01_38	57	10	1,341156	0,277024	1,227312	1,647388
S01_39	183	10	3,293523	0,37541	0,179793	0,497365
S01_40	130	10	2,87084	0,44167	0,529859	0,474937
S01_41	88	10	1,44534	0,831317	0,371203	0,523554

S01_42	98	10	1,240793	0,372535	1,399658	1,495829
S01_43	27	10	2,392209	0,82497	-0,14576	0,184689
S01_44	195	10	3,338104	0,611929	0,314249	0,170009
S01_45	98	10	3,200584	0,377521	0,660245	0,518369
S01_46	192	10	3,451691	0,40884	0,233128	0,38655
S01_47	180	10	1,510487	0,676403	0,899819	0,684465
S01_48	135	10	3,834161	0,690256	-0,34467	0,088318
S01_49	195	10	1,880872	0,400349	0,96232	1,019575
S01_50	131	10	1,447216	0,261664	1,064623	1,635805
S01_51	80	10	3,158136	0,499327	0,39081	0,307184
S01_52	82	10	1,972258	0,638854	0,736849	0,480501
S01_53	92	10	3,336428	0,552746	0,117509	0,216617
S01_54	161	10	2,183323	0,385768	1,019378	0,894883
S01_55	27	10	3,035779	0,93954	0,073284	0,072435
S01_56	155	10	3,327475	0,29444	0,555333	0,728923
S01_57	142	10	1,563607	0,751312	0,049658	0,544691
S01_58	51	10	1,087257	0,381768	0,779714	1,575712
S01_59	67	10	1,407141	0,3685	0,873134	1,39927
S01_60	113	10	3,040493	0,350035	0,663019	0,642152
S01_61	142	10	3,804469	0,318009	0,096369	0,517171
S01_62	197	10	2,651482	0,729979	0,329627	0,19342
S01_63	157	10	2,805299	0,517291	0,300062	0,365392
S01_64	122	10	1,590983	0,260606	1,010747	1,576483
S01_65	173	10	2,60571	0,399988	0,158951	0,663694
S01_66	54	10	1,538667	0,711394	1,266012	0,613685
S01_67	69	10	2,355659	0,277179	0,718426	1,174085
S01_68	169	10	1,95116	0,619138	0,440696	0,518559
S01_69	145	10	1,348524	0,264891	1,902866	1,67031
S01_70	63	10	1,558306	0,273205	1,092599	1,557817
S01_71	28	10	3,18919	0,70978	-0,10807	0,131971
S01_72	45	10	2,235616	0,297969	0,2543	1,152901
S01_73	59	10	2,242149	0,423745	1,62735	0,762041
S01_74	106	10	2,44093	0,710617	0,042004	0,249184
S01_75	56	10	2,282483	0,273548	0,779311	1,220631
S01_76	149	10	1,409471	0,497478	0,90978	1,097812
S01_77	21	10	3,474038	0,510772	0,048707	0,236686
S01_78	88	10	2,776913	0,452213	0,205641	0,48348
S01_79	113	10	3,383191	0,410302	0,269616	0,399409
S01_80	20	10	3,307097	0,475352	1,309295	0,309238
S01_81	125	10	3,754169	0,417176	0,632713	0,311723
S01_82	48	10	3,587889	0,86981	0,232471	0,056829
S01_83	85	10	1,950926	0,640176	0,872332	0,488327
S01_84	136	10	1,777782	0,253294	1,71185	1,516195
S01_85	160	10	3,226799	0,503823	0,17042	0,287632
S01_86	121	10	3,242083	0,660943	0,130993	0,151255
S01_87	62	10	3,753712	0,405718	0,141111	0,330709
S01_88	36	10	3,379574	0,946914	0,215358	0,053066

S01_89	35	10	1,399989	0,495428	0,561958	1,109597
S01_90	75	10	1,863249	0,918282	0,620444	0,256956
S01_91	140	10	1,584028	0,277648	1,377146	1,534041
S01_92	20	10	3,352328	0,312977	0,799919	0,656819
S01_93	58	10	1,386616	0,513932	0,558306	1,080186
S01_94	188	10	1,387268	0,4631	1,503383	1,191076
S01_95	187	10	1,216759	0,500666	1,107249	1,245719
S01_96	152	10	1,159388	0,787815	1,037491	0,805073
S01_97	80	10	2,595623	0,305701	0,295041	0,961393
S01_98	113	10	1,336925	0,356011	1,402999	1,472343
S01_99	154	10	3,229563	0,696282	-0,05781	0,13418
S01_100	131	10	3,193946	0,911118	0,124385	0,068645

Table 4

Simulation parameters of the 100 samples in the homogeneity of variance data set (S02)

Stimulus	Group size	Sigma	Scale
S02_1	75	4,504981	1,327677
S02_2	76	2,868631	0,775667
S02_3	37	2,866269	1,277896
S02_4	70	3,55578	0,664194
S02_5	59	5,769823	0,23682
S02_6	51	5,850432	0,663487
S02_7	64	4,959421	1,451601
S02_8	28	4,932984	0,726882
S02_9	59	4,143045	0,378688
S02_10	62	2,009092	0,389535
S02_11	47	4,43575	0,813024
S02_12	63	5,347206	0,974814
S02_13	76	5,00609	0,504629
S02_14	35	3,810926	0,091425
S02_15	48	4,14316	0,676966
S02_16	76	4,149507	1,258133
S02_17	79	2,005523	0,861956
S02_18	27	3,422664	0,530026
S02_19	48	4,448532	0,821139
S02_20	54	5,315769	1,339078
S02_21	74	3,426888	0,734986
S02_22	28	3,642541	0,257448
S02_23	79	4,293904	0,814546
S02_24	77	4,358713	1,442202
S02_25	25	4,878629	0,470526
S02_26	51	3,579892	1,230772
S02_27	43	5,676816	0,460582
S02_28	74	5,850281	0,27818
S02_29	47	2,934094	0,07252
S02_30	70	4,89799	0,368511
S02_31	64	5,614538	0,52666

S02_32	69	4,413896	0,238534
S02_33	43	4,526029	0,456147
S02_34	61	5,749543	0,026322
S02_35	20	5,401931	1,494829
S02_36	70	4,319284	1,20659
S02_37	20	5,285616	0,129871
S02_38	32	2,454874	1,3049
S02_39	74	5,058031	0,831879
S02_40	57	4,494454	0,632068
S02_41	43	2,593786	0,101455
S02_42	46	2,321058	0,842157
S02_43	22	3,856278	0,106083
S02_44	78	5,117473	0,317088
S02_45	46	4,934112	0,824431
S02_46	77	5,268922	0,722972
S02_47	73	2,68065	0,239205
S02_48	58	5,778881	0,224368
S02_49	78	3,174495	0,748909
S02_50	57	2,596288	1,410847
S02_51	40	4,877514	0,501347
S02_52	41	3,296344	0,282651
S02_53	44	5,115238	0,404574
S02_54	67	3,577764	0,796116
S02_55	22	4,714371	0,032175
S02_56	65	5,1033	1,19814
S02_57	61	2,751476	0,165503
S02_58	30	2,116343	0,809697
S02_59	36	2,542855	0,856851
S02_60	51	4,720657	0,928427
S02_61	61	5,739292	1,072281
S02_62	79	4,201976	0,184951
S02_63	66	4,407065	0,466574
S02_64	54	2,787978	1,418609
S02_65	71	4,140946	0,750038
S02_66	31	2,718223	0,202846
S02_67	36	3,807546	1,303887
S02_68	70	3,268213	0,307574
S02_69	62	2,464699	1,387569
S02_70	34	2,744409	1,33013
S02_71	23	4,91892	0,204444
S02_72	28	3,647488	1,178024
S02_73	33	3,656199	0,679955
S02_74	49	3,921241	0,203614
S02_75	32	3,709978	1,327832
S02_76	63	2,545961	0,50507
S02_77	20	5,298718	0,478911
S02_78	43	4,369217	0,605674

S02_79	51	5,177588	0,718616
S02_80	20	5,07613	0,551853
S02_81	55	5,672226	0,698536
S02_82	29	5,450519	0,074838
S02_83	42	3,267901	0,281035
S02_84	59	3,037042	1,473989
S02_85	67	4,969066	0,492411
S02_86	54	4,989444	0,256495
S02_87	34	5,671616	0,732382
S02_88	25	5,172765	0,028031
S02_89	25	2,533318	0,509228
S02_90	38	3,150999	0,044495
S02_91	60	2,778705	1,300843
S02_92	20	5,136438	1,097561
S02_93	33	2,515489	0,472891
S02_94	76	2,516357	0,579681
S02_95	76	2,289012	0,498669
S02_96	64	2,212518	0,134667
S02_97	40	4,127498	1,135583
S02_98	51	2,449233	0,904453
S02_99	65	4,972751	0,2181
S02_100	57	4,925262	0,048776

Appendix C

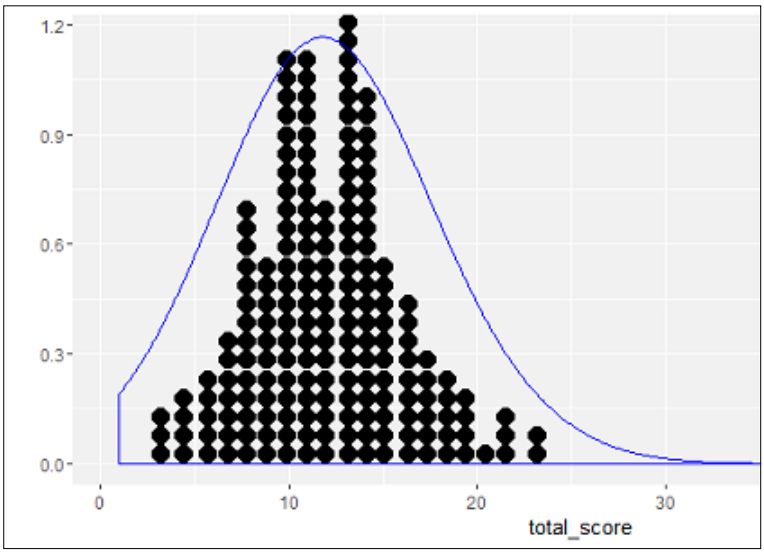

Figure 10. "Ideal" stimulus for the normality construct with inserted normal density curve

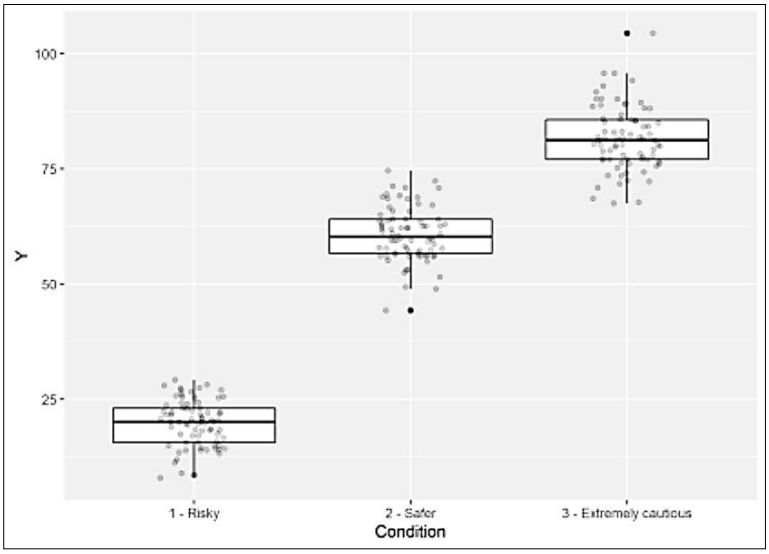

Figure 11. "Ideal" stimulus for the homoscedasticity construct

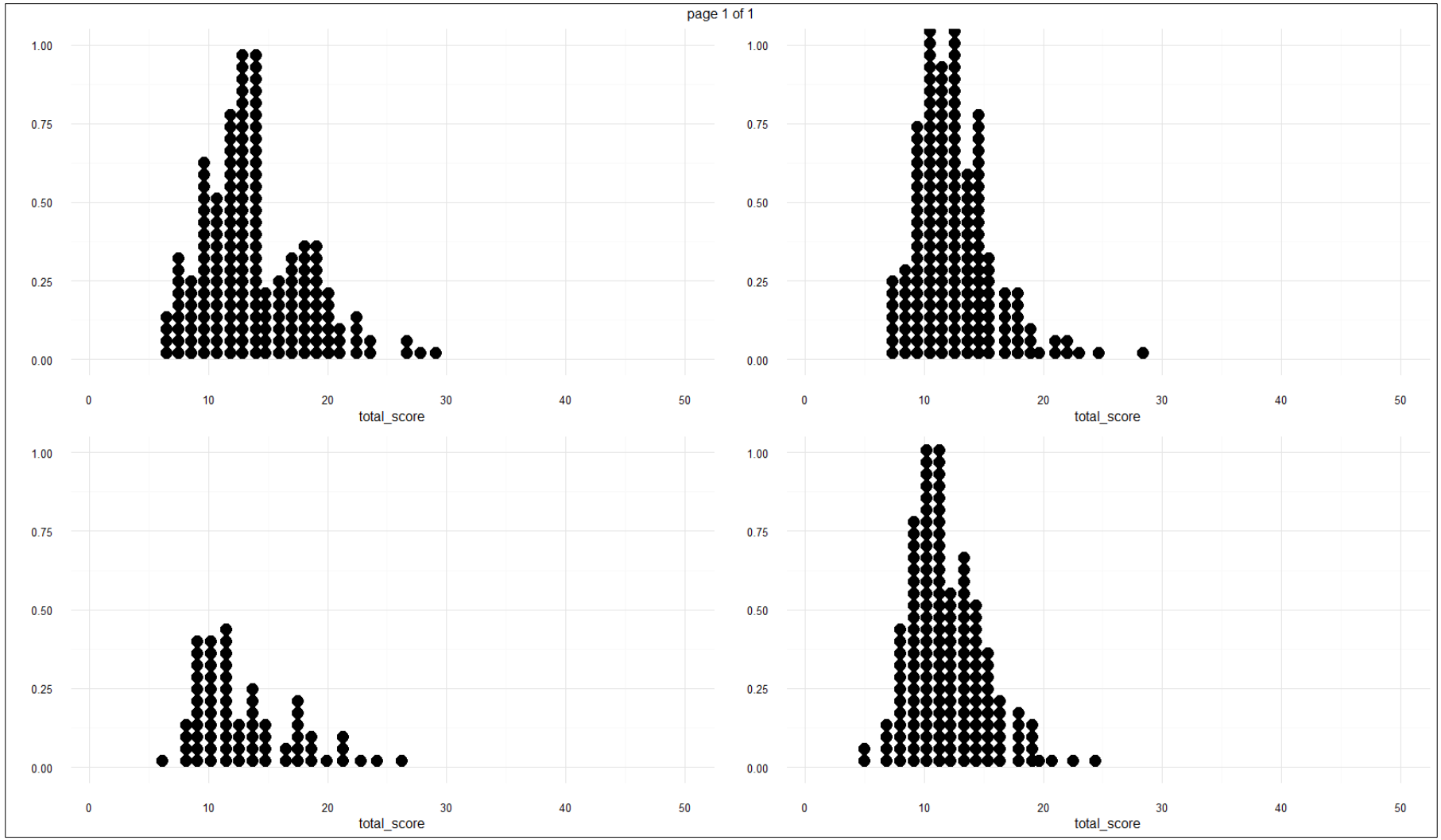

Figure 12. Example stimuli for the normality construct

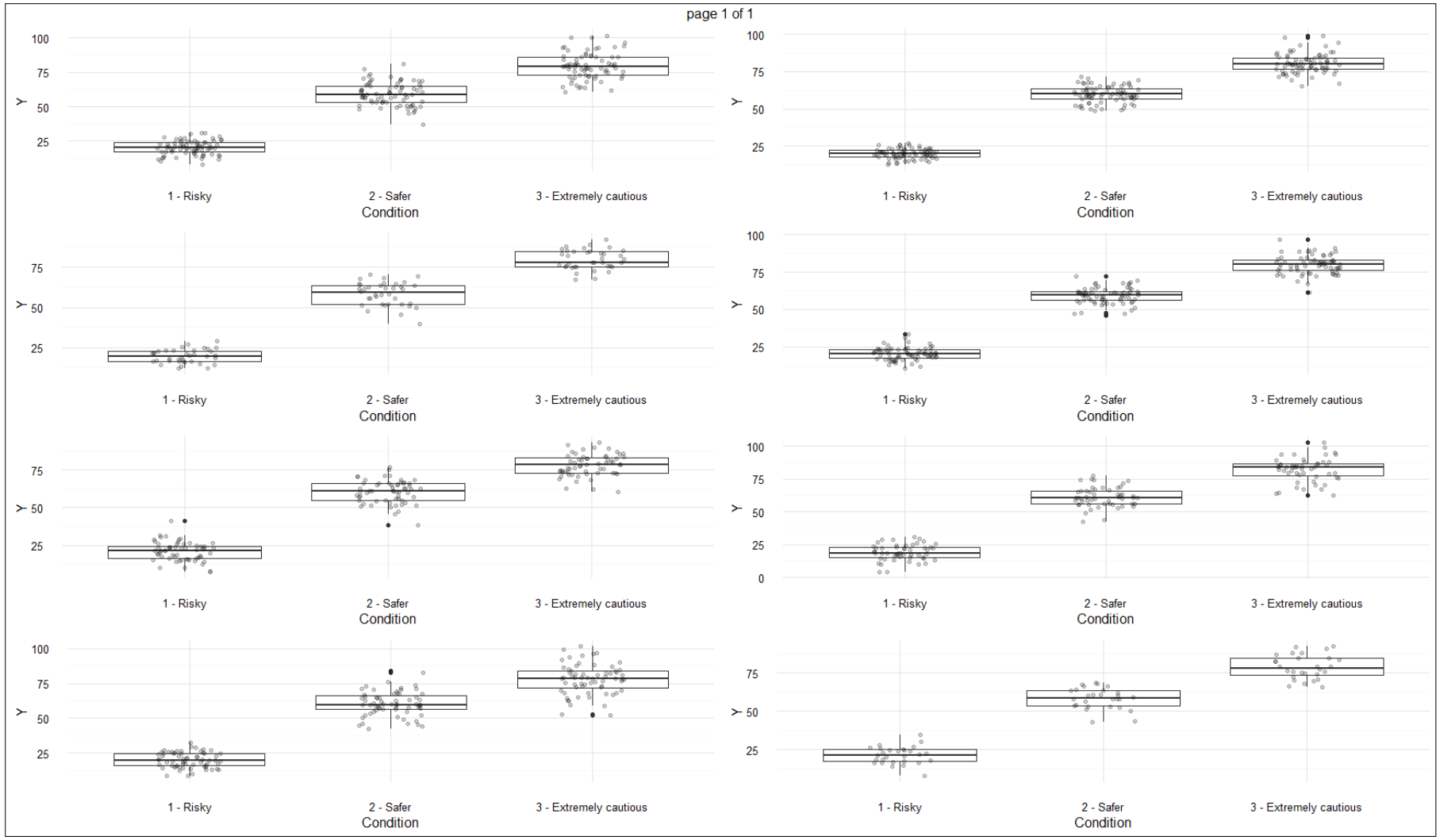

Figure 13. Example stimuli for the homoscedasticity construct

Appendix D

```
---
title: "Experimental evaluation of graphical exploratory data analysis"
author: "Martin Schmettow"
date: "`r format(Sys.time(), '%d %B, %Y')`"
output: html_document
---
```{r purpose, eval = T, echo = F}
purp.book = T
purp.tutorial = F
purp.debg = F
purp.gather = T
purp.mcmc = F #| purp.gather
purp.future = F
```
```{r libraries}
library(plyr)
library(pipeR)
library(dplyr)
library(tidyr)
library(pipeR)
library(readr)
library(haven)
library(stringr)
library(ggplot2)
library(openxlsx)
library(emg)
library(knitr)
library(moments)
library(car)
library(gridExtra)
library(lme4)
library(MCMCglmm)
library(brms)
library(rstanarm)
library(bayr)
rstan_options(auto_write = TRUE)
options(mc.cores = 3)
opts_knit$set(cache = T)
```
```{r profile, eval = T, echo = F, message = F}
## The following is  for running the script through knitr
# source("~/.cran/MYLIBDIR.R")
thisdir <- getwd()
# datadir <- paste0(thisdir,"/Daan/")
# figdir = paste0(thisdir, "/figures/")
## chunk control
opts_chunk$set(eval = purp.book,
               echo = purp.tutorial,
               message = purp.debg,
               cache = !(purp.gather | purp.mcmc))
options(digits=3)
opts_template$set(
  fig.full = list(fig.width = 8, fig.height = 12, anchor = 'Figure'),
  fig.large = list(fig.width = 8, fig.height = 8, anchor = 'Figure'),
  fig.small = list(fig.width = 4, fig.height = 4, anchor = 'Figure'),
  fig.wide = list(fig.width = 8, fig.height = 4, anchor = 'Figure'),
  fig.slide = list(fig.width = 8, fig.height = 4, dpi = 96),
  fig.half = list(fig.width = 4, fig.height = 4, dpi = 96),
  functionality = list(eval = purp.book, echo = purp.debg),
```
```

```
  invisible = list(eval = purp.book, echo = purp.debg),
  sim = list(eval = purp.book, echo = purp.tutorial),
  mcmc = list(eval = purp.mcmc, echo = purp.book, message=purp.debg),
  gather = list(eval = purp.gather, echo = purp.gather)
)
## ggplot
theme_set(theme_minimal())
```

```
# Simulation of stimuli for normality assessment
Data sets are created by drawing from the ex-gaussian distribution. The
below example shows the distribution with $\mu = 100, \sigma = 2, \lambda =
1/20$.
```{r}
data_frame(x = seq(0,200,1)) %>%
    mutate(total_score = demg(x, 100, 2, 1/20)) %>%
    ggplot(aes(x = x, y = total_score)) +
    geom_line()
```
Simulation
For the first part of the experiment, 100 stimuli are drawn that vary in how
much they are effected by the Gaussian component (large σ, little
skew) in relation to the exponential component (small λ).
```{r simulation_normal}
set.seed(42)
n_Stim = 100
S01 <-
    data_frame(Stimulus = str_c("S01_",1:n_Stim),
                          dist = "exgauss",
                          N    = round(runif(n_Stim, 20, 200),0),
                          mu   = 10,
                          sigma  = runif(n_Stim, 1, 4),
                          lambda = 1/runif(n_Stim, 1, 4))
# list of data frames
D01 <-
    S01 %>%
    alply(.margins = 1,
                .fun = function(s) data_frame(Stimulus = s$Stimulus,

Obs = 1: s$N,

total_score = remg(s$N, s$mu, s$sigma, s$lambda)))
# all values < 50
ldply(D01) %>%
  filter(total_score > 50) %>%
  print()
```
The following table shows the parameters of the `r n_Stim` data sets, the
plot shows the generated data sets (the stimuli). The parameters of the
simulated data sets were chosen as :
$\mu = 10$
$\sigma ~ uniform(1,4)$
$\lambda ~ uniform(1/4, 1)$
$N ~ uniform(20, 200)$
```{r simulation_normal_results}
kable(S01)
# plot(P01)
```
Objective criteria
Participants have to judge the data sets for normality. In the simpliest
case this is just a yes/no answer. The responses will then be compared to
objective criteria, possibly:
1. the amount of skewness in the population (as represented by the "true"
parameters)
```

```
2. the amount of skewness in the sample
3. result of a test for skew with Agostino test ($p < .05$)
4. result of a test for normality with shapiro test ($p < .05$)
```{r criteria_normal}
emg_skew <-
  function(mu, sigma, lambda) 2/(sigma^3 * lambda^3) * (1 + (1/(sigma^2 *
lambda^2)))^(-3/2) ## Wikipedia
C01 <-
  ldply(D01, function(d) skewness(d$total_score)) %>% ## sample skewness
  rename(skew_Sample = V1) %>%
  mutate(skew_Pop = emg_skew(mu, sigma, lambda)) %>%  ## population skewness
  full_join(select(ldply(D01,function(d)
agostino.test(d$total_score)$p.value),
                   Stimulus, agostino.p = V1)) %>%
  full_join(select(ldply(D01,function(d)
shapiro.test(d$total_score)$p.value),
                   Stimulus, shapiro.p = V1)) %>%
  mutate(agostino.nhst = ifelse(agostino.p < .05, "skew p<.05", "no skew"),
         shapiro.nhst = ifelse(shapiro.p < .05, "non-norm p<.05", "normal"))
%>%
  as_data_frame()
C01 %>%
  ggplot(aes(x = skew_Pop, y = skew_Sample, size = N))+
  geom_point(aes(color = agostino.nhst, shape = shapiro.nhst)) +
  geom_smooth(se = F, method = "lm")
## population skewness
head(C01) %>% kable()
C01 %>%
  mutate(agostino.rejected = agostino.p < .05,
         shapiro.rejected = shapiro.p < .05) %>%
  summarize(mean(shapiro.rejected),
            mean(agostino.rejected))
```
Example Stimuli
```{r sim_normal_create_plots}
# list of plots
P01 <-
  llply(D01[1:n_Stim],
        .fun = function(d)
          ggplot(d, aes(x = total_score)) +
          geom_dotplot(binwidth = 1) +
          xlim(1,50) +
          ylab("")
  )
marrangeGrob(P01[1:4], ncol = 2, nrow = 2)
```
Simulation of stimuli for homogeneity of variance assessment
Data sets are created by drawing from the a linear model with three groups with fixed means. Sample size varies, but the data is balanced. Residuals are normally distributed, but a scale parameter is applied to the standard deviation, letting it vary with the mean to a certain extent. The means (μ) of the three groups were fixed as $[1, 3, 4]$ Sample size, standard deviation of the first group and the scale parameter ϕ are varied across simulated data sets as follows:
The parameters of the simulated data sets were chosen as :
$\N_{grp} = uniform(20, 80)$
$\sigma ~ uniform(2,6)$
$\phi ~ uniform(0, 1.5)$
$\sigma_i = \sigma + \mu_i\phi$
In effect, when ϕ get larger, the variance in the groups more stringly increases with the mean, leading to more pronounced heteroscedasticity.
Simulation
```{r simulation_homo}
```
```

```
set.seed(42)
n_Stim = 100
## simulation parameters
S02 <-
  data_frame(Stimulus = str_c("S02_",1:n_Stim),
             N_grp    = round(runif(n_Stim, 20, 80),0),
             sigma  = runif(n_Stim, 2, 6),
             scale = runif(n_Stim, 0, 1.5))
## function to create one data frame
F02 <-
  function(P, mu = c(1,3,4)) {
  expand.grid(Condition = as.factor(c("1 - Risky",
                              "2 - Safer",
                              "3 - Extremely cautious")),
              Part = 1:P$N_grp) %>%
    full_join(data_frame(Condition = as.factor(c("1 - Risky",
                                        "2 - Safer",
                                        "3 - Extremely cautious")),
                         mu = mu),
              by = "Condition") %>%
    mutate(sigma = P$sigma + P$scale * mu,
           Y = rnorm(P$N_grp * 3, mu * 20, sigma))

}
# create data frames
D02 <-
    S02 %>%
    alply(.margins = 1,
                .fun = F02)
```

The following table shows the parameters of the `r n_Stim` data sets, the plot shows the generated data sets (the stimuli).

```{r sim_results_homo}
kable(S02)
# plot(P02)
```

Below are a few example plots:

```{r sim_homo_create_plots}
# list of plots
P02 <-
  llply(D02[1:n_Stim],
        .fun = function(d)
          ggplot(d, aes(x = Condition, y = Y)) +
          geom_boxplot() +
          geom_jitter(width = .4, alpha = .2)
  )
# examples
marrangeGrob(P02[1:8], nrow = 4, ncol = 2)
```

Objective criteria

Participants have to judge the data sets for homogeneity of variance. The responses will then be compared to objective criteria:

1. the amount of scale, relative to σ
2. result of the levene test ($p < .05$)

```{r criteria_homo}
fn.levene <- function(d) leveneTest(Y ~ Condition,
                                    data = d)$`Pr(>F)`[1]
# levene tests
C02 <-
  ldply(D02,fn.levene) %>%
  rename(levene.p = V1) %>%
  mutate(levene.nhst = ifelse(levene.p < .05, "heterosced p<.05",
"homosced")) %>%
```

```
  as_data_frame()


C02 %>%
  ggplot(aes(x = scale, y = N_grp))+
  geom_point(aes(color = levene.nhst))
head(C02) %>% kable()
C02 %>%
  mutate(levene.rejected = levene.p <= .05) %>%
  summarize(mean(levene.rejected))
```
```
```{r save_stimuli, message=FALSE, warning=FALSE, include=FALSE, eval = F}
for (i in 1:n_Stim) {
  ggsave(plot = P01[[i]],
         filename = paste0("S01_", i, ".png"),
         path = "stimuli")
}

for (i in 1:n_Stim) {
  ggsave(plot = P02[[i]],
         filename = paste0("S02_", i, ".png"),
         path = "stimuli")
}
```
```{r save_data, message=FALSE, warning=FALSE, include=FALSE, eval = T}
write.xlsx(D01, file = "S01.xlsx")
write.xlsx(C01, file = "Simuli_normal.xlsx")
write.xlsx(D02, file = "S02.xlsx")
write.xlsx(C02, file = "Simuli_homo.xlsx")
#save.image(file = "VEDA1.Rda")
```
Loading the data, the response variable is re-created. TRUE means: is normally distributed/has constant variance.
```{r load_data, opts.label = "gather"}
#load("VEDA1.Rda")
read_raw <- function(filename) {
    read_csv(filename) %>%
    select(2:8) %>%
    mutate(obs = row_number()) %>%
    mutate(TaskID = str_sub(StimID, 3,3)) %>%
    mutate(trial = obs %% (100 + 1))
  }
VEDA1_raw <-
  dir(pattern = "pp.*csv", recursive = T) %>>%
  ldply(read_raw) %>%
  as_data_frame() %>%
  rename(Part = participantID) %>%
  mutate(Task = ifelse(TaskID == "1", "Normality", "Constant Var"),
         grade = as.numeric(Grade),
         Stimulus = StimID,
         correct = Correctness) %>%
  select(-Grade, -TaskID)
VEDA1_Normal <-
  VEDA1_raw %>%
  filter(Task == "Normality") %>%
  left_join(C01) %>%
  mutate(reject.test = agostino.p < .05,
         correct = as.logical(correct),
         reject.part = (reject.test == correct))
VEDA1_ConstV <-
  VEDA1_raw %>%
  filter(Task == "Constant Var") %>%
  left_join(C02) %>%
```
```

```
  mutate(reject.test = (levene.p < .05),
         correct = as.logical(correct),
         reject.part = (reject.test == correct))
#write sav(VEDA1, "VEDA1.sav")
write_sav(VEDA1_Normal, "VEDA1_Normal.sav")
write_sav(VEDA1_ConstV, "VEDA1_ConstV.sav")
#save.image(file = "VEDA1.Rda")
```

Results on Normality

The following two plots show the association of the response (accept or reject normality) for the Shapiro test and the participants. We see an rather clear profile for the test: with increasing skew in the sample. The second plot shows the responses of participants, which generally is less clear cut and shows arge variation of the pattern across participants. It is immediatly clear that participants have severe difficulties in judging normality.

```{r eda_norm}
#load("VEDA1.Rda")
C01 %>%
  ggplot(aes(x = skew_Sample, y = N, col = shapiro.nhst)) +
  geom_point()
VEDA1_Normal %>%
  ggplot(aes(x = skew_Sample, y = N, col = reject.part)) +
  geom_point(alpha = .5) +
  facet_wrap(~Part)
```

We estimate a model for participant in dependence of sample skew and sample size.

```{r load_mcmc, eval = !purp.mcmc}
load("VEDA1_mcmc.Rda")
```

```{r mcmc:Norm, opts.label = "mcmc"}
#load("VEDA1.Rda")
rstan_options(auto_write = TRUE)
options(mc.cores = 3)
logit <- function(x) log(x/(1-x))
# M1_Norm <-
#   VEDA1_Normal %>%
#   mutate(min_sample = 20) %>%
#   brm(reject.part ~ skew_Sample + N + ((1 + skew_Sample + N) | Part),
#       family = bernoulli,
#       iter = 4000,
#       #prior = set_prior("normal(1,0.00001)", class ="sd", group =
"Stimulus", coef = "Intercept"),
#       data = .,
#       chains = 1)
#
# #save.image(file = "VEDA1.Rda")
M2_Norm <-
  VEDA1_Normal %>%
  mutate(min_sample = 20,
         skew_Sample = abs(skew_Sample)) %>%
  brm(reject.part ~  skew_Sample * N + ((1 + skew_Sample  * N)|Part ) +
(1|Stimulus),
      family = bernoulli,
      iter = 4000,
      #prior = set_prior("normal(1,0.00001)", class ="sd", group =
"Stimulus", coef = "Intercept"),
      data = .,
      chains = 1)
#save.image(file = "VEDA1.Rda")
#
# M3_Norm <-
```

```
#   VEDA1_Normal %>%
#   mutate(min_sample = 20) %>%
#   brm(reject.part ~  skew_Sample * N * trial + ((1 + skew_Sample  * N *
trial)||Part ) + (1|Stimulus),
#       family = bernoulli,
#       iter = 4000,
#       #prior = set_prior("normal(1,0.00001)", class ="sd", group =
"Stimulus", coef = "Intercept"),
#       data = .,
#       chains = 1)
#
# #save.image(file = "VEDA1.Rda")
```

Fixed effects

```{r tab:Norm_fixef}
#load("VEDA1.Rda")
M2_Norm %>% fixef() %>% kable()
```

Random effects

```{r tab:Norm_grpef}
M2_Norm %>%  grpef() %>% kable()
```

Results on Heteroscedasticity

The following two plots show the association of the response (accept or reject heteroscedasticity) for the Levene test and the participants.

```{r eda_constV}
#load("VEDA1.Rda")
C02 %>%
  distinct() %>%
  ggplot(aes(x = scale, y = N_grp, col = levene.nhst)) +
  geom_point()
VEDA1_ConstV %>%
  ggplot(aes(x = scale, y = N_grp, col = reject.part)) +
  geom_point(alpha = .5) +
  facet_wrap(~Part)
```

We estimate a model for participant in dependence of sample scale and sigma.

```{r mcmc:ConstV, opts.label = "mcmc"}
rstan_options(auto_write = TRUE)
options(mc.cores = 3)
# M1_ConstV <-
#   VEDA1_ConstV %>%
#   brm(reject.part ~ scale  * sigma + (1|Stimulus),
#             family = bernoulli,
#             data = .,
#       chains = 3)
# #save.image(file = "VEDA1.Rda")
M3_ConstV <-
  VEDA1_ConstV %>%
  brm(reject.part ~ scale  * N_grp * trial + (1 + scale  * N_grp *
trial|Part) + (1|Stimulus),
           family = bernoulli,
           data = .,
      chains = 1,
      iter = 4000)
#save.image(file = "VEDA1.Rda")
```

Fixed effects

```{r tab:ConstV_fixef}
M3_ConstV %>% fixef() %>% kable()
```

Random effects

```{r tab:ConstV_grpef}
```

```
M3_ConstV %>% grpef() %>% kable()
```

Further exploration of data

We have observed in both experiments that objective criteria (skew, scale, sample size) are being ignored by many participants. But the responses are not just random. The Stimuli intercept random effects show that stimuli systematically vary in how frequently they get rejected. Hence, there must be other criteria students use to judge the distributions. Maybe, participants had no clue about the objective criteria and used "fallback" heuristics, such as teh ruggedness of the distribution. Maybe, we can identify these heuristics by comparing plots of low and high rejection rates. For that purpose, we extract the stimulus-level random effects. They represent by how much a plot differs from the average rejection rate.

Normality

We start with the normality stimuli. The table below shows the Stimulus random intercepts.

```{r extract_stim_RE_Norm}
#load("VEDA1.Rda")
T_StimRE_Norm <-
  ranef(M2_Norm) %>%
  filter(str_detect(parameter, "Stimulus")) %>%
  mutate(parameter = str_replace(parameter, "Stimulus\\[S01_", ""),
         parameter = str_replace(parameter, ",Intercept\\]", ""),
         order = min_rank(center)) %>%
  rename(Stimulus = parameter) %>%
  arrange(order)
kable(T_StimRE_Norm)
```

The following plot shows the centers and 95% CIs for stimuli, ordered by center.

Although, the estimates are rather uncertain, there is considerable variance: stimuli vary by how frequently they are rejected.

```{r fig:caterpillar_Norm}
T_StimRE_Norm %>%
  ggplot(aes(x = order, y = center, ymin = lower, ymax = upper)) +
  geom_point() +
  geom_errorbar()
```

Now let's see, whether we can identify properties that are associated with high rejection:

We print every fifth stimulus, ordered by rejection rate

```{r fig:ConstV_ordered, opts.label = "fig.large"}
P02[T_StimRE_ConstV$Stimulus][seq.int(1, 100, 5)] %>>%
grid.arrange(grobs = ., nrow = 5, ncol = 4)
```

Appendix E

How to make learning statistics fun

Recent research on the effects of game-based learning (Kebritchi, Hirumi, & Bai, 2010; Wouters et al., 2009) for learning performance has established that game-based learning does have benefits for the learner. However, results are ambivalent. Research could not confirm whether game-based learning results in higher cognitive gains than traditional learning methods. However, it has been shown to improve the learner's attitude towards learning (Vogel et. al., 2006). Attitude, in turn, has been established to have a major impact on one's learning achievements.

The statistics education of psychologists at the University of Twente has received major criticism in recent years. For example, students criticize the utility of statistical workshops given as part of their B1 and B2 statistical education. According to them, the workshops mainly teach the students "what buttons to press in SPSS" instead of thematizing the when? and why? of statistical analyses, often resulting in major frustration on the part of the students.

We want to assess whether learning statistics in a game-based way can help to adjust students' negative attitude towards learning statistics. In turn, an enhanced attitude towards statistics is expected to benefit students' learning achievements.

Sources:

Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & education*, *55*(2), 427-443.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, G. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, *34*(3), 229-243.

Wouters, P., van der Spek, E. D., & van Oostendorp, H. (2009). Current practices in serious game research: A review from a learning outcomes perspective. *Games-based learning advancements for multi-sensory human computer interfaces: Techniques and effective practices* (pp. 232-250)

The experiment

During the experiment, you are randomly assigned to one of two groups. One group gets to play a statistical game with lots of game elements, the other group will play a statistical game with just a few game elements. This way, we want to examine the degree of game elements needed to make the statistics learning experience more fun for the student. The whole game should take about *30-45min* to complete.

After completion of the game, you will be asked to give your opinion on the fun-factor of the statistical game and we will further examine your attitude towards learning statistics.

Appendix F

Table 5

Random intercepts and 95% CIs for normality stimuli, ordered by point estimate

Stimulus	Point Estimate	Lower	Upper	Order
35	-2.190	-3.183	-1.300	1
92	-1.936	-3.015	-1.102	2
33	-1.875	-2.973	-0.972	3
54	-1.370	-2.154	-0.608	4
89	-1.352	-2.236	-0.547	5
96	-1.280	-2.206	-0.620	6
14	-1.098	-2.016	-0.349	7
93	-1.075	-1.891	-0.386	8
15	-0.993	-1.745	-0.255	9
13	-0.948	-1.916	-0.002	10
40	-0.937	-1.712	-0.127	11
7	-0.816	-1.707	-0.115	12
23	-0.810	-1.625	0.054	13
36	-0.803	-1.675	-0.071	14
66	-0.703	-1.574	0.117	15
67	-0.699	-1.498	-0.016	16
61	-0.660	-1.560	0.181	17
9	-0.636	-1.420	0.116	18
82	-0.633	-1.332	0.230	19
95	-0.615	-1.398	0.165	20
22	-0.573	-1.303	0.142	21
4	-0.572	-1.373	0.119	22
53	-0.569	-1.416	0.196	23
21	-0.560	-1.357	0.103	24
79	-0.551	-1.384	0.151	25
52	-0.551	-1.229	0.199	26
65	-0.542	-1.448	0.290	27
56	-0.516	-1.254	0.209	28
72	-0.504	-1.247	0.295	29
44	-0.428	-1.337	0.360	30
26	-0.423	-1.235	0.295	31
27	-0.408	-1.175	0.308	32
19	-0.357	-1.052	0.368	33
5	-0.343	-1.156	0.324	34
91	-0.321	-1.057	0.480	35
63	-0.314	-1.143	0.420	36
2	-0.294	-1.091	0.646	37
10	-0.278	-1.000	0.574	38
55	-0.277	-1.022	0.639	39
81	-0.270	-0.979	0.324	40
48	-0.205	-0.882	0.477	41

83	-0.192	-0.883	0.509	42
49	-0.166	-0.899	0.598	43
29	-0.142	-0.763	0.543	44
45	-0.115	-0.841	0.514	45
38	-0.092	-0.905	0.719	46
100	-0.088	-0.888	0.570	47
97	-0.086	-0.862	0.521	48
12	-0.059	-0.812	0.611	49
90	-0.018	-0.660	0.731	50
69	-0.005	-0.916	1.043	51
59	0.017	-0.629	0.749	52
57	0.027	-0.750	0.811	53
94	0.056	-0.983	0.865	54
86	0.113	-0.658	0.750	55
25	0.123	-0.584	0.910	56
28	0.138	-0.754	0.862	57
60	0.178	-0.475	0.864	58
41	0.180	-0.493	0.851	59
6	0.187	-0.485	0.936	60
24	0.203	-0.539	0.870	61
88	0.213	-0.543	0.978	62
11	0.240	-0.462	0.969	63
32	0.298	-0.464	1.076	64
77	0.319	-0.468	1.233	65
87	0.337	-0.356	1.149	66
68	0.420	-0.349	1.052	67
8	0.452	-0.285	1.253	68
17	0.466	-0.526	1.542	69
76	0.494	-0.118	1.321	70
99	0.516	-0.229	1.313	71
37	0.547	-0.186	1.396	72
3	0.604	-0.196	1.329	73
34	0.616	-0.096	1.313	74
85	0.618	-0.156	1.287	75
80	0.644	-0.454	1.505	76
31	0.651	-0.039	1.357	77
51	0.660	0.006	1.406	78
16	0.669	-0.049	1.451	79
42	0.673	-0.099	1.571	80
84	0.697	-0.303	1.679	81
98	0.709	-0.125	1.613	82
30	0.785	-0.028	1.505	83
75	0.820	0.065	1.567	84
47	0.843	0.152	1.700	85
1	0.848	0.100	1.677	86
62	0.860	0.134	1.601	87
43	0.889	0.121	1.786	88

46	0.897	0.080	1.603	89
39	0.931	0.151	1.650	90
20	0.952	0.316	1.682	91
64	0.984	0.218	1.790	92
70	1.129	0.381	2.222	93
74	1.216	0.537	2.014	94
73	1.228	0.171	2.298	95
78	1.236	0.494	2.033	96
58	1.341	0.627	2.359	97
50	1.443	0.597	2.470	98
71	1.528	0.627	2.392	99
18	1.571	0.795	2.543	100

Table 6

Random intercepts and 95% CIs for homoscedasticity stimuli, ordered by estimate value

Stimulus	Point Estimate	Lower	Upper	Order
52	-1.132	-2.064	-0.441	1
13	-1.062	-2.018	-0.270	2
93	-0.957	-1.658	-0.278	3
73	-0.868	-1.577	-0.194	4
66	-0.866	-1.668	-0.194	5
83	-0.863	-1.653	-0.200	6
69	-0.826	-1.533	-0.094	7
15	-0.793	-1.546	-0.195	8
42	-0.791	-1.471	-0.187	9
86	-0.755	-1.603	-0.037	10
53	-0.743	-1.463	-0.061	11
40	-0.724	-1.436	-0.023	12
72	-0.714	-1.423	0.048	13
8	-0.711	-1.382	-0.054	14
49	-0.699	-1.566	-0.072	15
4	-0.696	-1.466	-0.038	16
45	-0.678	-1.372	-0.025	17
79	-0.650	-1.366	-0.043	18
21	-0.647	-1.387	0.040	19
26	-0.632	-1.341	0.005	20
89	-0.631	-1.327	0.025	21
57	-0.630	-1.506	0.152	22
41	-0.620	-1.466	0.069	23
11	-0.488	-1.148	0.176	24
68	-0.479	-1.348	0.268	25
12	-0.464	-1.136	0.195	26
27	-0.441	-1.153	0.169	27
81	-0.365	-1.079	0.274	28
39	-0.359	-1.203	0.304	29
71	-0.339	-0.997	0.401	30
16	-0.326	-1.085	0.386	31

47	-0.313	-1.201	0.497	32
61	-0.264	-0.902	0.364	33
30	-0.232	-1.014	0.479	34
70	-0.223	-0.902	0.552	35
29	-0.213	-0.925	0.488	36
24	-0.193	-1.010	0.562	37
59	-0.180	-0.868	0.442	38
38	-0.158	-0.957	0.605	39
90	-0.156	-0.896	0.516	40
36	-0.126	-0.791	0.591	41
63	-0.107	-0.823	0.564	42
77	-0.105	-0.780	0.621	43
91	-0.074	-0.822	0.537	44
75	-0.038	-0.817	0.700	45
7	-0.024	-0.789	0.706	46
2	-0.011	-0.699	0.657	47
98	0.006	-0.621	0.653	48
64	0.011	-0.676	0.816	49
85	0.025	-0.622	0.730	50
94	0.064	-0.731	0.716	51
99	0.088	-0.643	0.770	52
6	0.091	-0.607	0.680	53
32	0.126	-0.668	0.832	54
96	0.139	-0.593	0.938	55
9	0.172	-0.471	0.869	56
37	0.173	-0.611	0.839	57
88	0.173	-0.492	0.955	58
14	0.174	-0.467	0.897	59
97	0.175	-0.527	0.896	60
20	0.214	-0.455	0.990	61
55	0.232	-0.496	0.979	62
62	0.241	-0.579	1.053	63
23	0.283	-0.476	0.941	64
48	0.305	-0.390	0.968	65
17	0.329	-0.361	0.997	66
67	0.344	-0.398	1.194	67
95	0.350	-0.333	1.006	68
3	0.375	-0.417	1.162	69
18	0.381	-0.224	1.093	70
58	0.410	-0.243	1.166	71
10	0.438	-0.238	1.057	72
33	0.440	-0.151	1.169	73
28	0.447	-0.337	1.142	74
22	0.463	-0.218	1.182	75
76	0.470	-0.172	1.180	76
5	0.480	-0.237	1.156	77
87	0.486	-0.176	1.148	78

74	0.491	-0.195	1.181	79
35	0.493	-0.399	1.399	80
50	0.502	-0.232	1.250	81
92	0.512	-0.298	1.330	82
44	0.573	-0.156	1.300	83
25	0.634	0.005	1.394	84
46	0.644	-0.008	1.320	85
43	0.650	-0.105	1.389	86
19	0.654	0.058	1.373	87
60	0.699	0.037	1.470	88
34	0.706	0.042	1.430	89
1	0.733	-0.062	1.554	90
100	0.740	0.028	1.407	91
78	0.743	0.092	1.387	92
56	0.763	0.037	1.569	93
54	0.782	0.152	1.518	94
82	0.812	0.165	1.615	95
51	0.824	0.206	1.615	96
31	0.831	0.180	1.501	97
84	0.860	0.091	1.710	98
80	0.893	0.137	1.674	99
65	1.087	0.396	1.784	100

Appendix G

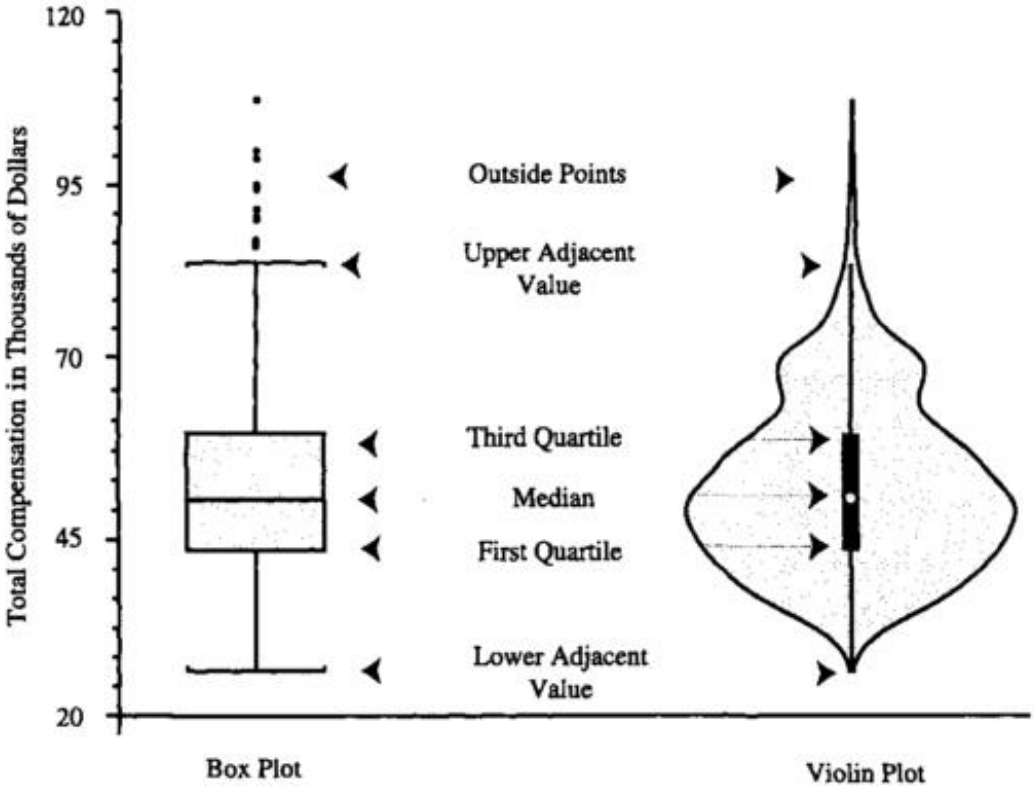

Figure 14. The components of a violin plot in comparison to the components of a box plot. Adapted from Hintze and Nelson (1998).