A COMPARISON OF PARAMETRIC, NONPARAMETRIC, AND OBSERVATION ORIENTED MODELING TECHNIQUES

A Masters Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Psychology

By

Kathrene Diane Valentine

May 2013

A COMPARISON OF PARAMETRIC, NONPARAMETRIC, AND

OBSERVATION ORIENTED MODELING TECHNIQUES

Psychology

Missouri State University, May 2013

Master of Science

Kathrene Diane Valentine

ABSTRACT

Many individuals have shown concern with the current statistical procedure employed by most social scientists, namely that of null hypothesis significance testing (NHST). Alternative techniques do exist, such as nonparametric testing, which focuses less on assumptions and uses simpler analyses, and observation oriented modeling (OOM), a new modeling paradigm created by Dr. James Grice as a way to focus on data at the individual level. Using these three techniques (a parametric ANOVA, a nonparametric Quade test, and OOM's ordinal pattern analysis) a dataset collected on the QWERTY effect was analyzed to look into the effect of typability of keypress combinations on pleasantness ratings for words. The analyses revealed that the nonparametric test was more powerful and capable of distinguishing more statistical differences and larger effect sizes than the parametric test. These tests are also compared to the OOM test which provided information about how individuals matched on this expected pleasantness pattern. These incongruent findings indicate a need for a greater understanding of all available statistical techniques and more investigation into the proper use of these techniques.

KEYWORDS: statistical technique, parametric, nonparametric, observation oriented modeling, QWERTY effect

This abstract is approved as to form and content

_____

Erin Buchanan, PhD
Chairperson, Advisory Committee
Missouri State University

A COMPARISON OF PARAMETRIC, NONPARAMETRIC, AND OBSERVATION ORIENTED MODELING TECHNIQUES

By

Kathrene Diane Valentine

A Masters Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Psychology

May 2013

Approved:

_____
Erin Buchanan, PhD

_____
Wayne Mitchell, PhD

_____
Melissa Fallone, PhD

_____
Thomas Tomasi, PhD, Associate Dean Graduate College

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In current statistical thinking, we are quick to stamp out any ideas students or researchers have about stating, for instance, "I have *proven* that the Earth is round." Instead we shuffle them towards the conventional caveated conclusion, "The Earth is round, ($p < 0.05$)" (Cohen, 1994). Many such issues plague statisticians today; the bulk of which have come from a lack of understanding of the plethora of tools in the statistical toolbox. Within this manuscript, this downfall, frequently referred to as Null Hypothesis Significance Testing (NHST), will be discussed and propose alternative options by comparing and contrasting traditional parametric NHST with nonparametric statistics, and a new form of statistical analysis, Observation Oriented Modeling (OOM). Three methods will be introduced and then applying them to a dataset collected on the QWERTY effect (described below).

Null Hypothesis Significance Testing (NHST)

To fully understand the NHST procedure, the procedure's origins will be introduced, followed by how the procedure is used today. Many believe that Ronald A. Fisher is responsible for the patchwork procedure of NHST. In reality, Fisher's ideas looked very little like NHST as it is taught and applied currently in classrooms. Fisher believed in creating one "null" hypothesis, which he described as a hypothesis to be "nullified", not as zero change or a zero correlation (Lehmann, 2011). For example, in traditional Fisherian NHST, a null hypothesis could be $\rho = 0.2$, instead of no population correlation. Second, Fisher believed that the use of an omnibus 5% level of significance

showed a "lack of statistical thinking" (Gigerenzer, Krass, & Vitouch, 2004), and instead believed we should report the exact significance value, which is in line with current APA standards (Wilkinson & APA Taskforce on Statistical Inference, 1999). Since Fisher is not wholly to blame for this procedure as used today, we turn to one man who helped to inspire both Fisher, and the two men who would challenge Fisher's theory.

The second voice that contributed greatly to the formation of NHST as it is used today was simply called Student to the public. To a select few at the time, he was known as William Gosset. Gosset created what is now called Student's *t*-test, and contributed to work on the correlation coefficient (Lehmann, 2011). Gosset conversed with both Fisher and Pearson throughout the years, encouraging Fisher with his development of hypothesis testing. Through conversations, Gosset also inspired Egon Pearson with the idea of the alternative hypothesis (Lehmann, 2011). This idea of an alternative hypothesis, not included in Fisher's theory, was used by the two men who altered Fisher's hypothesis testing procedure to create their own decision theory.

Jerzy Neyman and Pearson created Neyman-Pearson decision theory, which instead of consisting of one hypothesis to be nullified, consists of two hypotheses (a null and an alternative hypothesis) and a binary decision criteria (Lehmann, 2011). The key to decision theory is being able to look at the outcome and make the most cost-effective decision (Gigerenzer, 2004). For example, let us say that you are trying to compare two methods of producing pencils. With this decision theory, your null hypothesis may be that two methods produce the same number of usable pencils each day. Your alternative hypothesis may be that one method works better than another, but you are not sure which. Based on the analysis used and the consequent *p*-value, you would decide which method

you believed was significantly better at producing pencils, and thus would be the most effective method of production.

First, due to the production of two hypotheses, the researcher now must investigate possible decision errors (Lehmann, 2011). A researcher may falsely reject the null (Type I error) or falsely fail to reject the null (Type II error). In order to assess the probability of committing one of these errors, alpha and beta levels must be defined. An alpha level is the probability that you will commit a Type I error, which is equal to the cutoff *p*-value that you set for your hypothesis test. A beta level is the probability that you will commit a Type II error, and is found by subtracting the power of the study (the probability that the null will be rejected if, in fact, the null is true) from one (Aron, Aron, & Coups, 2009). Second, Neyman and Pearson clearly state that a hypothesis should not be blindly supported based solely on the support of one statistical test, but that replication and reproduction of results are imperative. Third, this analysis is clearly against setting omnibus alphas and betas and instead, is geared toward being able to adjust the analysis to the needs of the particular task at hand (Gigerenzer, 2004).

Current NHST procedures consist of three main steps: 1. Create only one hypothesis that states that there are no differences between populations. 2. Use the conventional rejection region of 5% to reject the null hypothesis. 3. Repeat for all analyses (Gigerenzer, 2004). This use of NHST should sound familiar. This procedure comes with many underlying assumptions. For instance, let us assume that an individual has decided to run a repeated measures analysis of variance (ANOVA). Now, before the analysis can be examined there are many assumptions that must be met for the analysis to be accurate. The assumptions are as follows: 1. data screening has been performed for

accuracy, missing, multicollinearity, and outlier analysis, 2. data is normally distributed, 3. data combinations are linear, 4. variances are homogeneous, 5. errors are independent, and 6. variables pass the test of sphericity. Once we are sure that these assumptions are met we can analyze the repeated measures ANOVA knowing our results are accurate.

Through all these steps it is easy to get lost, nevertheless a quick comparison of these three processes shows that Fisher, Gosset, Neyman, and Pearson are not entirely responsible for the null ritual. The blame can only be placed on ambition. During the time these procedures were arising, the field of psychology was known as a "soft" science. Many experimentalists took up these procedures in order to show evidence of the researched phenomena, and to gain acceptance for the field as a science. In the rush to demonstrate our capabilities, the steps of the processes were muddled, the rational for the procedures were misplaced, and the role of the researcher as a sentient part of the analysis process was let slip away. This odd procedure has been allowed to continue on and is being taught, even by this writer, to introductory statistics classes everywhere. Pleas from highly regarded individuals have not been heard (Cohen, 1990; LeBel & Peters, 2011; Rosnow & Rosenthal, 1989; Skinner, 1956; Tukey, 1969). NHST has continued to persist in Psychology classrooms, research, and journals. Herein, I propose that we examine the ongoing statistical process, and outline applicable alternatives in the statistician's toolbox: nonparametric statistics and Observation Oriented Modeling.

Nonparametric Analysis

In contrast to traditional parametric NHST, nonparametric statistics place less emphasis on assumptions. They also use simpler models that require less calculation, and

are based on simpler theories that allow a researcher to more easily assess their data and choose the correct analysis. Additionally, nonparametric statistics are frequently more powerful than parametric statistics (in this case NHST).

For the purpose of this study, a nonparametric analysis for several related samples needed to be chosen. Two well-known dependent samples nonparametric tests are the Friedman test and Quade test. The Quade test was chosen over Friedman's test because with only three experimental groups (as in this study), Friedman's test suffers from a lack of power (Conover, 1999). Quade's test depends on only three assumptions: 1. variables are mutually independent, 2. observations within each block (for each participant) may be ranked according to some criterion, and 3. a sample range is capable of being determined within each block so they may be ranked. The test focuses on the ranked observations within each block and the average participant rank for the different word types in this study (Conover, 1999).

The repeated measures ANOVA and Quade test results are comparable by viewing the asymptotic relative efficiency (A.R.E.) of the tests. A.R.E. refers to test comparisons based on sample size: A.R.E is the ratio of required sample sizes if both tests have comparable levels of power and significance. The test that requires the smallest sample size is preferred with a higher A.R.E. (Conover, 1999). When the test is better, the A.R.E. value is close to one, though the value can approach infinity. For the purposes of this experiment, G* Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) was used to compute the sample sizes necessary to complete the Quade test and the *F*-test (repeated measures ANOVA).

Observation Oriented Modeling

Many researchers have claimed that changing the way we use null hypothesis significance testing could help psychology (Cohen, 1994; Fidler & Loftus, 2009; Gigerenzer et al., 2004; Mulaik, Raju, & Harshman, 1997; Rosnow & Rosenthal, 1989; Schmidt & Hunter, 1997). Conversely, Grice (2012) argues that our problems are "much deeper and wider" than just the NHST. He argues that with a different philosophical idea (realism), the absence of estimating population parameters, the use of integrated models, a renewed appreciation of "eye-tests," and the formation of deep structures of data, that the social sciences can improve the way they grow and amass knowledge.

A New Paradigm. Grice argues that we first need to shift our philosophical position to be in line with St. Thomas Aquinas and Aristotle. By viewing psychology through the lens of realism, instead of positivism, we should be able to properly and effectively conduct research and analyze data. In contrast to positivism—which is solely concerned with finding an effect, not with how the effect occurred—realism is the belief that effects conform to their cause and that all theories have an underlying truth. By viewing science as knowing nature through its causes, we can use Aristotle's four causes (material, efficient, formal, and final) to think in terms of forming structures and processes for phenomena. Switching to this philosophy allows for techniques that match the daily activities of social scientists in their endeavors to unravel the story of how humans operate.

When using OOM, researchers can stop focusing on samples as they relate to the population. As Grice stated in a presentation at Missouri State University (2012), "Who is the population anyway?" If a researcher actually attempted to count the population, it

would be impossible. For instance, consider research concerning individuals with cancer. Are we considering everyone with any type of cancer? Do we count all individuals with cancer just in the United States? Even as we are counting the population, attrition and remission must be considered. Trying to estimate and understand the abstract idea of a population is difficult and probably not useful. OOM does not use population parameters and their various underlying assumptions; instead the researcher looks at observations at the level of the individual. Skinner (1956) stated that methods allowing direct observation of the individual would be the most important in behavioral sciences, and OOM is one way in which this idea can be put into practice.

As a replacement to procedures currently used in psychological research, Grice addressed the need for integrated models. In contrast to simpler variable-based modeling, where X causes Y, an integrated model accounts for the structures and processes that go into the actual phenomena studied. When a detailed framework is constructed for an experiment, it is easy to ask specific questions about the model and to add to or change said model as each experiment reveals something new.

Modeling Paradigm. These integrated models can be analyzed on any type of data social scientists could collect. From ordinal rankings to frequency counts, all analyses are calculated in the same general fashion. This simplicity occurs because observation oriented modeling works on the deep structure of the data. Through observation definition, the program then breaks these units into binary code. This technique creates a common language, which allows analysis through various techniques. Deep structures can be arranged to form a matrix, which can then be manipulated via matrix algebra, binary Procrustes rotation, and other operations to investigate the data.

The most important values from any OOM analysis are the PC and PCC (percent match and percent complete match). These values represent how well your observations matched the stated or expected pattern or, in the case of causal modeling, how many of your observations conformed to a given cause. Matches are the proportion of observations that align with our designated pattern on at least one dimension. Alternatively, complete matches are the proportion of observations that match the designated pattern on all dimensions. The PCC value replaces all of the conventional values for effect size used in statistical analyses.

In OOM *p*-values are no longer used. As a secondary form of reference value, a chance value or c-value, is obtained by randomizing observations anywhere from 100 to 5,000 times. This randomization procedure is much like bootstrapping without replacement where the original data is shuffled a number of times to create a number of comparable data sets. These randomized data sets are then compared to the designated pattern. If the randomized data sets fit the pattern more often than the actual data does, the c-value will be high (close to 1). Low c-values are indicative of distinct observations that are not likely due to chance. Although low c-values, like low *p*-values, are desirable, c-values do not adhere to a strict cut-off and should be considered a secondary form of confirmation for the researcher that their results are distinct.

In order to analyze the repeated measures data from the QWERTY study, I will be using OOM's newest analysis: Ordinal Pattern Analysis. This analysis allows the researcher to designate the expected ranked pattern: each variable as being higher, lower, or equal to the other variables. See Figure 1 for an example of a defined pattern (Note that plus signs represent hypothesized squares for the given pattern and "O"s represent non-

hypothesized squares). Once this pattern is defined, the program then analyzes the data to see if each individual's set of observations match this expected ordinal pattern. A PC, PCC, and c-value is generated based on the number of individuals that completely match the expected ordinal pattern. This analysis does not form any type of linear or nonlinear equation or regression, but simply looks for those individuals who match the expected ordinal pattern.

QWERTY

The QWERTY effect is the phenomenon whereby people rate words as more pleasant if the words are composed of more letters that lie on the right hand side of a QWERTY keyboard than on the left hand side. Most recently, Jasmin and Casasanto (2012) wrote about how the QWERTY effect influenced our perceptions of old words (prior to the invention of the QWERTY keyboard), new words (after the invention of the QWERTY keyboard), and made-up words (pseudowords that imitate real words). However, their study did not fully consider the implications of embodied cognition, which is the notion that our procedural actions influence other psychological phenomena (Beilock & Holt, 2007). Therefore, we might expect that actions that can be performed more fluently are perceived as more pleasant (Oppenheimer, 2008). Previous studies on embodied cognition and typing have shown unconscious activation of procedural skills, even though no typing was performed (Beilock & Holt, 2007; Ping, Dhillon, & Beilock 2009; Yang, Gallo, & Beilock, 2009). To expand upon this research, the study presented here has examined more than simple left-right preference. Beilock and Holt (2007) demonstrated that expert typists preferred typing easier combination of letters; those with

combinations on opposite hands over single hand combinations. This finding leads us to believe that "typability" needs to be taken into account when considering the perceived pleasantness of these words. This experiment sought to mimic the Jasmin and Casasanto study and also to look into how "typability" affected pleasantness (i.e. are the letters right-left-right-left more pleasant than right-left-left-right?).

For the purpose of this study the words were separated into three categories: repeated keystrokes not paired together (RN), repeated keystrokes together (RY), and not repeated with different fingers (DN). See Table 1 for examples. We would expect that more typable words would be preferred over those that are less typable.

Hypotheses

By analyzing the data from the QWERTY study with a repeated measures ANOVA, a Quade test, and an Ordinal Pattern Analysis we can see the differences between the three methods and examine how well each method fits and explains the data at hand. The author predicts the following:

- Hypothesis 1: The ratings of the three different types of keypresses will differ significantly from one another.
- Hypothesis 2: The nonparametric Quade test will be more powerful than the parametric ANOVA. This hypothesis will be examined by analyzing A.R.E. of the statistics.
- Hypothesis 3: OOM will provide more meaningful information than both the parametric and nonparametric tests by being able to compare how each individual responded to the stimuli and how likely these results are to come about from simple randomized data.

METHOD

Participants

Participants consisted of 157 undergraduate students, who were recruited from General Psychology classes through the human subject's pool and were given class credit for their participation. Participants were tested in groups that ranged from one to four participants. The Institutional Review Board approved this experiment on April 4th, 2012 (Project #12375).

Materials

A list of 240 words was compiled using the English ANEW (Bradley & Lang, 1999) norms to offer a variety of real words and pseudowords that were typed with repeated keystrokes not paired together (RN), repeated keystrokes together (RY), and not repeated with different fingers (DN). The ANEW database did not contain a sufficient number of words in each category so 76 (31.7%) words were added. Each participant was only asked to rate 120 of the 240 words to control for the effect of fatigue on their judgments. Of the 120 words that participants rated, they were randomized and split between real words (60) and pseudowords (60). For clarification, a table of word examples is listed in Table 1.

Procedure

Participants were asked to read and sign a consent form upon entry into the experiment. All participants were tested on desktop IBM clone computers with Windows

7 operating system. Participants completed a one minute typing test (Typing Master, Inc., 2013). The participants were asked to complete as much of the Aesop's fables test as they could in one minute. After the typing test, participants were asked to alert the researcher who recorded the typing speed and errors that appear on the screen after completion of the test.

Participants were then asked to record which hand they write with most often. Lastly, participants completed a pleasantness rating task. They were shown a word in the middle of the screen in 18 point Arial font. Participants were asked to use a self-assessment manikin (SAM; see Figure 2, Jasmin & Casasanto, 2012) to rate their judgments. A SAM scale is a nine point visual emotional scale, which will allow the participants to rate their perceived pleasantness of a word. They were asked to rate this word as unpleasant (1) to pleasant (9) by using the number pad or clicking on the appropriate number on the screen. The scale was always shown at the top on the screen. At no point in the experiment were participants asked to type the word they saw on the screen. Participants rated 120 words and were then given course credit for their participation.

Assessments

In order to assess how these various procedures work, all three methods of analysis will be compared through a dataset testing for the QWERTY effect. For NHST, we would start by creating our null and alternative hypotheses, which would be for the null, that there would be no difference between the three categories of words, and for the alternative, that there would be some significant difference between the three groups at

the $p < 0.05$ level. We would assess this by examining a repeated measures ANOVA as participants rated all word types. If the omnibus *F*-test from this analysis is significant, we are then able to analyze post-hoc tests to investigate group differences. Finally, we will calculate effect sizes for the analyses and can speak to the general results of the study (Cohen's *d* using the average standard deviation as the denominator).

For the nonparametric Quade test, summary scores were created for each of the three word types. Each participant was left with three observations, which were then ranked 1, 2, or 3, where the smallest number (least pleasant perception) was given a 1, and the largest number (most pleasant perception) was given a 3. Rankings are completed within participants. In the case of ties the average rank was used (which occurred once across participants). Then, the range for each participant was created based on the original observations. For example, if the original observations were 2, 5, and 6, the range would be 4. These ranges for all participants are then ranked across participants, from the smallest number, which receives a 1, to the largest number. Again, in the case of ties the average rank was used (which occurred sixty-six times). This number (the rank of the block) was then used to create the Quade scores for each participant. These scores were created by multiplying the individual's rank by the difference between that individuals rank and their average rank (in our experiment the average rank was equal to 2). These ranks represent the overall ranking each category so that an average rating of 40 is perceived as more pleasant than an average rating of 1, which is perceived as more pleasant than an average rating of -40.

As Conover (1999) states, if the *F* statistic, which is usually performed on the raw data, is instead computed on the ranks created by the Quade procedure, the statistic is the

same as the test statistic for the Quade test. Therefore, a repeated measures ANOVA would be examined. If the *F*-test from this analysis is significant, we are then able to analyze post-hoc tests to investigate group differences. Finally, we will calculate effect sizes for the analyses and can speak to the general results of the study (Cohen's *d* using the average standard deviation as the denominator).

The OOM Ordinal Pattern Analysis procedure is much simpler. The averaged data are simply entered into the program, an ordinal analysis is selected, and the expected pattern is defined. The expected pattern for this analysis was that RN would be rated higher (more pleasant) than DN, which would both be rated higher than RY. This defined pattern can be seen in Figure 1. Use of the randomization tests when analyzing the data will show not only the percent of individuals that the pattern fit on at least one factor, and the percent of individuals that completely matched the pattern, but also how many times this percentage of individuals (or more) was found by randomizing the data 1,000 times. The randomization trials give the researcher another facet to examine if data conform to model fit better than a random arrangement of the data.

RESULTS

Hypothesis 1

Parametric (NHST) Repeated Measures ANOVA. As there were several words rated for each typing combination, the three different typing conditions were averaged for each of the participants' judgments. A repeated measures ANOVA was performed on these summary scores. The main effect of word typability was significant, $F(2, 312) = 29.46$, $p < 0.001$, $\eta^2 = 0.16$.

Dependent t-tests were used for post hoc comparisons. Repeated letters not paired together (RN: $M = 5.17$, $SD = 0.69$) words were not significantly preferred over the non-repeating words (DN: $M = 5.09$, $SD = 0.68$), $t(156) = 1.95$, $p = 0.053$, $d = 0.12$. The RN words ($t(156) = 6.04$, $p < 0.001$, $d = 0.42$) and DN words ($t(156) = 7.55$, $p < 0.001$, $d = 0.31$) were significantly preferred over the repeating words with double keystrokes (RY: $M = 4.87$, $SD = 0.75$). This finding illustrated that participants preferred words whose letters are typed with repeated keystrokes not paired together or different keystrokes over repeated keystrokes paired together.

Nonparametric (NHST) Quade Test. The Quade test was used to analyze the differences in perceived pleasantness between the three different typing combinations. The Quade test is an extension of the Wilcoxon signed rank test (Conover, 1999). The test critically measures ranked observations within each block, which refers the average participant rank for the different word types in this study.

Therefore, a repeated measures ANOVA was analyzed on the Quade ranked data resulting in a significant main effect, $F(2, 312) = 29.45$, $p < 0.001$, $\eta^2 = 0.16$. Dependent t-

tests were computed as post hoc analyses on these ranked scores. The RN words (M = 29.78, SD = 80.09) were significantly preferred over the DN words (M = 10.64, SD = 52.51, $t(156) = 2.07$, $p = 0.041$ $d = 0.28$), and the RY words (M = -40.42, SD = 69.80, $t(156) = 6.25$, $p < 0.001$, $d = 0.93$). Also, the DN words were significantly preferred over the RY words, $t(156) = 6.81$, $p < 0.001$, $d = 0.83$. This result shows that individuals preferred words whose letters were typed with repeated keystrokes not paired together over both repeated keystrokes paired together and not repeated keystrokes, and words that repeated keystrokes with the same fingers were preferred less than those with no repeated keystrokes using different fingers.

Hypothesis 2

These statistics were then compared by viewing the asymptotic relative efficiency (A.R.E.) of the tests. Using G*Power 3, the F test would need 8 participants to find significant effects, and the Quade test would need only 6. Thus, when comparing the F test to the Quade test an A. R. E. of 0.75 results, and when comparing the Quade test to the F test an A. R. E. of 1.33 results. From these findings, it is clear that the Quade test was a more powerful and efficient choice for this data. Also, the *post hoc* analyses showed that the Quade test was able to detect the difference between the RN and DN word ratings that the F test did not show. Interestingly, the effect sizes for the nonparametric Quade tests also show that this test was capable of identifying larger effects than the parametric ANOVA.

Hypothesis 3

OOM (NON-NHST). The hypothesized pattern of ordinal observations defined for participant observations consisted of the RN words begin rated the highest, followed by the DN words, and finally the RY words (See Figure 1). Observations for individuals over the three typing combinations show that part of this pattern matched for 311 of the possible 471 matches, creating a proportion of 0.66. When a randomization test was analyzed, the c-value was found to be less than 0.001 indicating that results were unique. The entire pattern matched on 65 out of the possible 157 individuals, creating a proportion of 0.15. Again, when a randomization test was analyzed, the c-value was < 0.001 portraying unique data patterns. These results are consistent with those found from the Quade test, also suggesting that RN words are perceived as more pleasant than DN words, and RY and DN words are preferred over RY words.

DISCUSSION

Hypothesis 1

Our results support this hypothesis by showing that both the parametric ANOVA and the nonparametric Quade test have shown significant differences between types of keypresses. The ANOVA indicated that repeating not together and no repeating keystrokes were significantly preferred over repeating together keystrokes, but were unable to distinguish a difference between repeating not together and no repeating keystrokes. The Quade test, on the other hand, was capable not only of distinguishing that repeating not together and no repeating keystrokes were significantly preferred over repeating together keystrokes, but also that repeating not together keystrokes were significantly preferred over not repeating keystrokes.

Hypothesis 2

Our results also supported this hypothesis by showing that the A.R.E. for the Quade test is significantly higher when compared to the parametric ANOVA. The Quade test was also able to not only distinguish that repeating not together and no repeating were rated significantly higher than repeating together words, but also that repeating not together was rated significantly higher than no repeating keypress words. This nonparametric test was also capable of revealing much larger effect sizes between these comparisons than the parametric ANOVA.

Hypothesis 3

This hypothesis was supported as using OOM's Ordinal Analysis allowed us to designate that this difference not only appears on some level (as in parametric and nonparametric tests) but also that it appears on a case-by-case basis on the individual level a high percentage of the time. This analysis also revealed that our results were unique, and not likely due to random chance by examining a randomization test and c-values.

Conclusions

It is clear to see that, as so many before have stated, there are obvious flaws in the current methodology employed in the social science today (Cohen, 1990; Gigerenzer, 2011; LeBel & Peters, 2011; Rosnow & Rosenthal, 1989; Skinner, 1956; Tukey, 1969). Parametric and nonparametric statistics comprise the majority of published analyses in social science literature today, but should we let this continue? Is there an alternative avenue that can be used to present evidence of phenomena? As Tukey (1969, pg 722) once stated "Exploration relies greatly on looking around. Indeed, unless practical psychologists produce new ways to receive the message from data, there will continue to be no substitute for visual techniques in exploring data…There really seems to be no substitute for 'looking at the data'." His statement clearly agrees that we should be focusing more on what our data can *show* us, as opposed to whether or not they surpass an arbitrary level of significance. Within this manuscript we can see that OOM is capable of helping the social sciences transition from the current dogmatic practices of NHST into the more simplistic and fundamental resource of OOM.

As can be seen in this paper, there are valuable alternatives to classic parametric NHST. Depending on the data to be analyzed and the hypothesis in question, different methods may be best suited to the study at hand. In this scenario, the nonparametric test was better suited to determining the differences in the data than the parametric method, and the OOM analysis allowed the researcher a wealth of insight into the individuals studied as compared to the other two analyses. It is important to note that the given techniques did not describe the data in congruent ways. This problem is concerning as these differing outcomes may allow researchers the ability to pick and choose their analyses based on what conclusions they wish their data to support. While both the parametric and nonparametric techniques used here give differing pictures of the same data, OOM presents an unbiased view of the information at hand and allows the researcher to make an informed decision about their data.

Overall, the use of NHST in any form needs to be re-evaluated in the social sciences. With the plethora of alternatives to this procedure it is a wonder that we are still rooted in this old and misleading tradition. Although remedies such as effect sizes, confidence intervals, and a priori power analysis may help alleviate some concerns regarding NHST (Fidler & Loftus, 2009; Gigerenzer, 2004; Wilkinson & APA Taskforce on Statistical Inference, 1999), only by discarding the procedure all together and reevaluating our underlying philosophies can we move forward. The only way to assure that our results and conclusions are valid is to be sure that the methodology and analyses that we use are the optimal choice for the data at hand.

Table 1. Examples of word types.

Word Type	Real words	Pseudowords
Repeated keystrokes not together (RN)	cute, wasp, bath	tutk, yame, tove
Repeated keystrokes together (RY)	hide, army, frog	neeb, zafe, celm
Not repeated with different fingers (DN)	blue, pest, slap	pobe, voke, plin

	RN	DN	RY
Highest Score	+	O	O
	O	+	O
Lowest Score	O	O	+

Figure 1. Example of set OOM pattern.

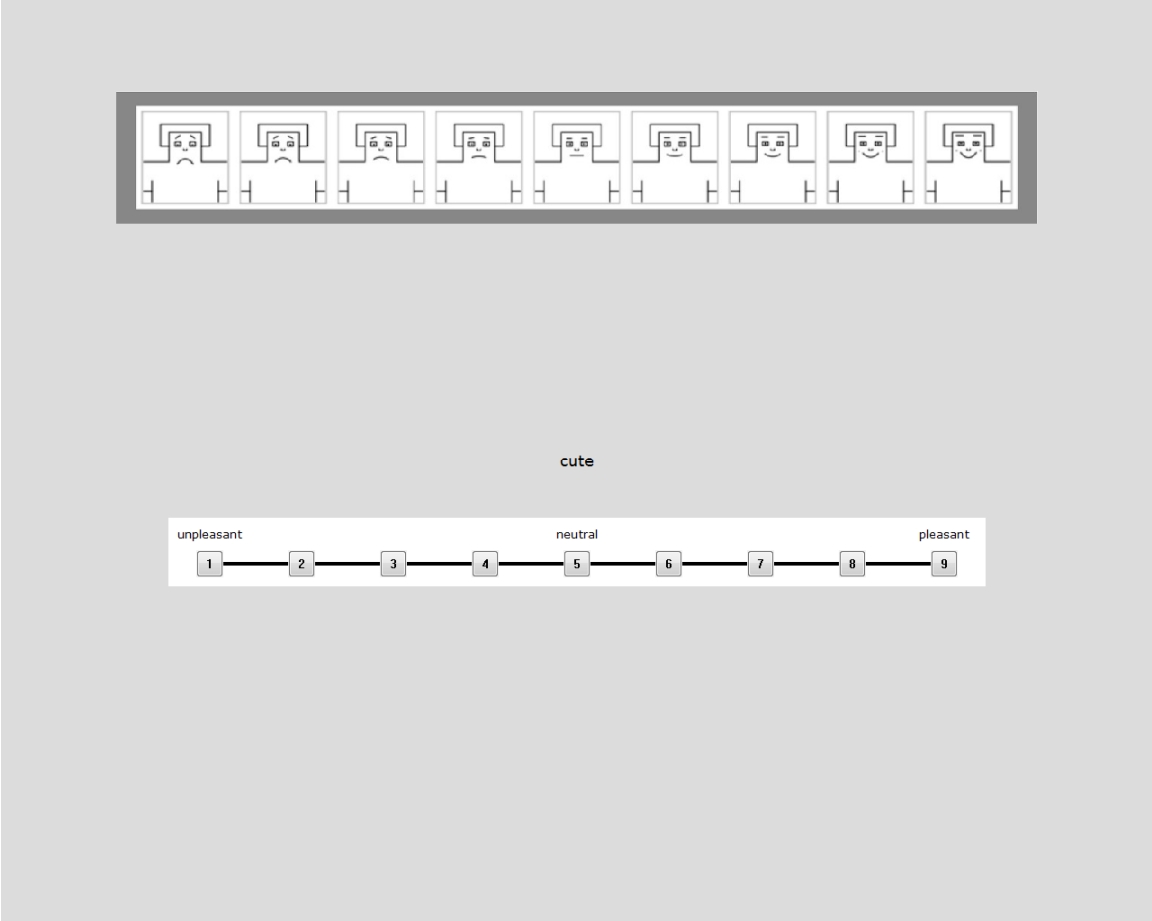

Figure 2. Screenshot of experimental rating procedure.

REFERENCES

Aron, A., Aron, E. N., & Coups, E. J. (2009). *Statistics for Psychology* (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Beilock, S. L. & Holt, L. E. (2007). Embodied preference judgments: Can likeability be driven by the motor system? *Psychological Science, 18,* 51-57. doi:10.1111/j.1467-9280.2007.01848.x

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (pp. 1-45). Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45* (12), 1304-1312.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49* (12), 997-1003.

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). Danvers, MA: John Wiley & Sons, Inc.

Faul, F., Erdfelder, E., Lang, A, & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods, 39* (2), 175-191.

Fidler, F. & Loftus, G. R. (2009). Why figures with error bars should replace *p* values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology, 217*, 27-37.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. *The Sage Handbook of Quantitative Methodology for the Social Sciences*, 391-408.

Grice, J. W. (2012, February 24). Observation oriented modeling: A common sense approach to data conceptualization and analysis. Recorded at Missouri State University. Podcast retrieved from iTunes RStats podcasts.

Jasmin, K., & Casasanto, D. (2012). The QWERTY effect: How typing shapes the meanings of words. *Psychonomic Bulletin & Review*, *19*, 499-504. doi:10.3758/s13423-012-0229-7

LeBel, E. P. & Peters, K. R. (2011). Fearing the future of empirical Psychology: Bem's

(2011) evidence of Psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*(4), 371-379.

Lehmann, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics.* New York, NY: Springer.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. *What if There Were No Significance Tests?*

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends In Cognitive Sciences*, *12*, 237-241. doi:10.1016/j.tics.2008.02.014

Ping, R., Dhillon, S., & Beilock, S. L. (2009). Reach for what you like: The body's role in shaping preferences. *Emotion Review, 1*, 140-150. doi:10.1177/1754073908100439

Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedure and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276-1284.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. *What if There Were No Significance Tests?*

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11*(5), 221-233.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83-91.

TypingMaster, Inc., (March Aesop's Fables. TypingMaster, Inc. Web. Retrieved from http://www.typingtest.com/

Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. doi: 10.1037/0003-066X.54.8.594

Yang, S., Gallo, D., & Beilock, S. L. (2009). Embodied memory judgments: A case of motor fluency. *Journal of Experiment Psychology: Learning, Memory, & Cognition, 35,* 1359-1365. doi:10.1037/a0016547