## Title

Persistence of Naïve Statistical Reasoning Concerning Analysis of Variance

**Permalink**

https://escholarship.org/uc/item/5wr8p179

**Journal**

Proceedings of the Annual Meeting of the Cognitive Science Society, 31(31)

**ISSN**

1069-7977

**Authors**

Hachey, Krystal
Mewaldt, Steven
Trumpower, David

**Publication Date**

2009

Peer reviewed

# Persistence of Naïve Statistical Reasoning Concerning Analysis of Variance

**David L. Trumpower (david.trumpower@uottawa.ca)**
**Krystal Hachey (khach026@uottawa.ca)**
University of Ottawa, Faculty of Education, 145 Jean-Jacques-Lussier Street
Ottawa, ON K1N 6N5 Canada

**Steven Mewaldt (mewaldt@marshall.edu)**
Marshall University, Department of Psychology, One John Marshall Drive
Huntington, WV 25755 USA

## Abstract

Two experiments were conducted to assess the intuitive reasoning of students when examining data from an analysis of variance design. Participants were shown hypothetical datasets that differed with regards to within-group and/or between-group variability, and were asked to judge the amount of evidence that each provided in support of a particular theory. The first experiment (n=57) examined the influence of presentation format of the hypothetical datasets. Participants were randomly assigned to receive the hypothetical datasets in one of two formats: (1) group data stacked vertically in a single column, or (2) group data displayed side-by-side in two columns. In the second experiment (n=13), students' reasoning about the hypothetical datasets was assessed both before and after completing an introductory graduate level statistics course. Consistent with prior research, participants tended to place an inordinate amount of weight on the relative importance of between-group, as opposed to within-group, variability. The results indicate that neither presentation format (Experiment 1) nor statistics training (Experiment 2) is enough to overcome this aspect of naïve statistical reasoning.

**Keywords:** naïve statistics; statistical understanding; intuitive knowledge; conceptual knowledge; expert-novice differences

## Introduction

In recent years more students are being required to register for statistics courses as it is becoming more prevalent in various degree programs. As such, a greater number of students are experiencing statistics anxiety (Onwuegbuzie & Wilson, 2003). For educators, identification of statistics-naïve students' intuitive understandings and biases may help to alleviate their anxiety and focus their instruction, respectively. Thus, the purpose of this paper is to examine the persistence of intuitive understandings and biases concerning statistics.

Over forty years ago, Peterson and Beach (1967) reviewed the literature on the ability of individuals to be intuitive about, or estimate, statistics. They were impressed by the accuracy with which individuals with no formal training in statistics can estimate descriptive statistics, such as means and variability. However, there are some biases present in statistics-naïve individual's reasoning, as well. For example, Beach and Scopp (1967) and Kareev, Armon and Horwiz-Zeligar (2002) have noted that estimation of variability tends to be conservative. Peterson and Beach (1967) suggested that this may be due to an unwillingness to weight large deviations heavily.

Recently, there has been renewed interest in naïve statistics (e.g., Masnick & Morris, 2008; Trumpower & Fellus, 2008). This resurgence has focused on statistics-naïve students' ability to perform an intuitive analysis of variance (ANOVA) - that is, to detect between and within-group variability and integrate that information in order to make judgements about group differences. In the typical "intuitive ANOVA" study, participants are presented with datasets comprised of scores from two different conditions. They are then asked to make judgements concerning the evidence for a difference between the two conditions.

Using this experimental paradigm, Trumpower and Fellus (2008) showed some of the understandings and biases that individuals display when performing an intuitive ANOVA. Participants were given a cover story regarding a theory that frozen golf balls travel farther than unfrozen golf balls. They were told that in order to test the theory, frozen and unfrozen golf balls were hit with the same force by a robotic arm. The distances traveled by a group of unfrozen golf balls and a group of frozen golf balls were then shown to the participants, and they were asked to rate the amount of support that the hypothetical experiment provides for the claim that frozen balls travel farther (see Figure 1). Several such hypothetical datasets were presented, varying with respect to the magnitude of the between and/or within-group variability.

*Three normal golf balls and three frozen golf balls are each hit by the robot in random order:*

| *normal* | *frozen* |
|---|---|
| 300 | 450 |
| 250 | 400 |
| 350 | 350 |

Weak Strong
1 2 3 4 5 6 7 8 9 10

Why?

Figure 1. Example dataset used in Experiments 1 & 2.

Ratings indicated that statistics-naïve college students can detect and understand the importance of both between-group and within-group variability. For datasets that had the same within-group variability, participants reliably rated the dataset with larger between-group variability as providing stronger evidence for a difference between conditions. For datasets that had the same between-group variability, participants reliably rated the dataset with smaller within-group variability as providing stronger evidence for a difference between conditions. Thus far, participant's intuitive ANOVA reasoning is consistent with the strength of evidence that would be indicated by performing an actual ANOVA on the datasets. However, participants also displayed a bias of overweighing the importance of large between-group variability relative to small within-group variability. One of the hypothetical datasets (shown in Figure 1) depicted a 100 yard difference between the means of the frozen and unfrozen golf balls coupled with a 50 yard standard deviation within each group. Thus, this dataset had a 2:1 ratio of between to within-group variability. Another of the hypothetical datasets depicted a 4 yard difference between the means of the frozen and unfrozen golf balls. But, this was coupled with just a 1 yard standard deviation within each group, resulting in a 4:1 ratio of between to within-group variability. If one were to compute an actual t-test or ANOVA on these two datasets, the one with the greater ratio of between to within-group variability would generate the larger F-statistic. However, participants reliably rated the dataset with the smaller 2:1 ratio as providing stronger evidence for a difference between conditions. In fact, Trumpower and Fellus found this bias for large but less reliable group differences over small yet more reliable group differences in a group of students with no prior training in statistics as well as in a group of students who had just completed a university-level statistics course.

In a similar study, Masnick and Morris (2008) examined the intuitive ANOVA capabilities of 9 year olds, 12 year olds, and college students. They found that even the 9 year olds understand the relevance of between-group variability. The relevance of within-group variability appeared to be a more difficult concept to grasp, though. Only the college students' judgements about the data were influenced by within-group variability, albeit inconsistently. College students were sometimes more confident in a difference between two datasets when there was *larger* within-group variability (although Masnick and Morris point out that within-group variability was confounded with the magnitude of the data points in their study). In one other related study, Lubbock and Miller (1996) found that only about half of their 15 year old participants could identify within-group variability as being a source of "trustworthiness" of data.

Thus, it appears that that statistics-naïve individuals (as young as 9 years old) have an intuitive understanding of the relevance of between-group variability. But, intuitive understanding of the relevance of within-group variability is a bit less clear. As we described earlier, Trumpower and Fellus (2008) documented one particular bias that both statistics-naïve and experienced students have regarding the relative importance of within-group variability. The Trumpower and Fellus study had two limitations, however, which may temper any claims concerning the nature and stability of this bias in reasoning about within and between group variability. First, the naïve and experienced students were sampled from independent populations. The naïve participants were undergraduate-level teacher education students, whereas the experienced participants were graduate-level students, most of who were enrolled in a Master's in Nursing program. Any claims about the persistence of bias in ANOVA reasoning would be premature without demonstrating it in the same group of individuals before and after formal statistics instruction. Second, Trumpower and Fellus suggested that the bias may have been at least partially due to a misinterpretation of the data arising from the presentation format employed in the study. Because data from the frozen and unfrozen conditions were aligned, side-by-side, in two columns as depicted in Figure 1, participants (especially the experienced ones) may have misinterpreted the golf balls as being paired across conditions. That is, they may have believed that the data were from a correlated-groups design. Consistent with this interpretation, several of the participants actually computed difference scores by subtracting the distances of the frozen golf balls from the distances of the corresponding unfrozen golf balls when weighing the evidence provided by the data. Scanning across columns when performing this pairwise-type comparison might serve to highlight large between-condition differences while obscuring within-column variability. Consider again the data in Figure 1. This was the dataset mentioned earlier that had a 2:1 ratio of between to within-group variability. Computation of difference scores in this dataset would result in two instances where the frozen golf ball went 150 yards further than the corresponding unfrozen golf ball, and only one instance in which the frozen golf ball did not travel farther than the corresponding unfrozen ball. This may be contrasted with the dataset with a 4:1 ratio of between to within-group variability. In this dataset, computation of difference scores would have resulted in three instances in which the frozen ball went farther than the corresponding unfrozen ball, but none in which the frozen ball went farther by more than 5 yards. In the qualitative results provided by participants, several mentioned that the frozen golf balls went much farther than the unfrozen balls on 2 out of 3 instances in the former dataset and that all of the frozen balls went further than the unfrozen balls, but only by an "insignificant" amount, in the latter dataset. This sort of response was more prevalent in the statistics-experienced participants. Indeed, calculation of difference scores and thinking about whether differences are big enough to be considered "significant" are likely to be fresh in the minds of students after a semester studying statistics. The side-by-side presentation format of the data coupled with the fact that the experienced participants had recently performed correlated groups t-tests in their just completed statistics course may have primed this way of biased reasoning. If so, then the apparent persistence of the bias may be due to different factors in the naïve and experienced students.

The purpose of the current study was to build on the previous study completed by Trumpower and Fellus (2008). More specifically, we assess the persistence of the intuitive ANOVA bias identified by Trumpower and Fellus across different data presentation formats (Experiment 1), and within the same group of students before and after instruction (Experiment 2).

## Experiment 1

When teaching students to perform a correlated groups t-test or ANOVA by hand, paired data are almost exclusively presented in side-by-side columns. When performing the same type of analysis with most computer software programs, such as SPSS, data must be entered in separate columns, as well. Thus, the presentation format used by Trumpower and Fellus may have led statistics-experienced participants to misinterpret the data as a correlated-groups design, thereby leading them to calculate difference scores. This, in turn, may have focused their attention on large difference scores in favor of the frozen golf ball traveling farther and turned their attention away from within-column variability.

If so, then presenting all of the data in a single column, in which distances of the unfrozen golf balls are stacked above the distances of the frozen golf balls, should eliminate misinterpretation of the data as coming from a correlated groups design. Data to be analyzed with a correlated groups ANOVA are almost never presented in this stacked format in class or when entered into a data analysis software program. Because between-group designs are analyzed by performing calculations within each condition (i.e., computing group means and standard deviations) rather than computing difference scores across conditions, it was expected that this "stacked" format would reduce or eliminate the tendency to focus more so on large differences across conditions than on within-condition variability.

In Experiment 1, statistics-experienced participants were presented with hypothetical datasets to evaluate in an intuitive ANOVA paradigm as in Trumpower and Fellus (2008). Participants were randomly assigned to receive the datasets in a side-by-side, two column format or in a vertically stacked, single column format. Table 1 presents the descriptive statistics of the hypothetical datasets, as well as the F-statistic that would result if an actual ANOVA were performed on the datasets. The bias identified in Trumpower and Fellus (2008) was indicated by participants (both naïve and experienced) rating dataset 2 as providing stronger evidence than dataset 3, and dataset 1 as providing stronger evidence than dataset 3 (experienced participants only).

### Method

**Participants** Fifty-seven students, who had recently completed an introductory level university statistics course at Marshall University, served as participants in exchange for partial course credit. The participants were randomly assigned to one of two groups (n=28; n=29) as described below.

### Materials & Procedure

The students in both groups were given a test booklet, in which the first page was the same for all the students. The first page of the test booklet displayed the following scenario:

> *Suppose two scientists/entrepreneurs are considering whether or not to develop a golf ball freezer that can be attached to a regular golf bag. They have a theory that frozen golf balls travel farther than normal (i.e., unfrozen) golf balls. To test their theory, the scientists/entrepreneurs devise an experiment in which a robotic arm will be used to hit normal and frozen golf balls, all with the exact same force, after which the distance that each ball travels will be measured. In order to remain completely unbiased, the scientists/entrepreneurs will allow independent researchers (who are completely unaware of their theory) to conduct the experiment.*
>
> *Listed on the following page are hypothetical results from several such experiments. For each experiment, rate the amount of support (1=weak, 10=strong) that you think the test would provide for the claim that frozen golf balls go farther than normal golf balls, and briefly explain why.*

The second page of the test booklet displayed four hypothetical datasets. Each scenario provided the distances traveled by 3 frozen and 3 unfrozen golf balls. Participants were asked to rate, on a 10 point rating scale, the strength of support provided by the data for the claim that frozen golf balls go farther than normal golf balls. Participants were also provided space to explain the reason for each of their ratings.

The raw scores for both the unfrozen and frozen golf balls were displayed vertically in one column (one set stacked above the other) for one group and horizontally in two columns (side-by-side) for the other group.

Table 1: Means (and standard deviations) of conditions, and resulting F-statistics, of hypothetical datasets.

| Dataset | Unfrozen | Frozen | F |
|---|---|---|---|
| 1 | 300 (50) | 304 (50) | .01 |
| 2 | 300 (50) | 400 (50) | 6 |
| 3 | 300 (1) | 304 (1) | 24 |
| 4 | 300 (1) | 400 (1) | 15000 |

### Results

Ratings were analyzed using a 2 Format (side-by-side; stacked) x 4 Dataset split-plot ANOVA with repeated measures on the second factor. The main effect of Dataset was significant, F(3,162)=103.48, p<.001, but neither the Format main effect, F(1,54)=0.055, p=.816, nor the Dataset x Format interaction, F(3,162)= 0.638, p=0.591, was significant (see Figure 2).

The main effect of Dataset was followed up by all possible pairwise comparisons among the 4 datasets. The Tukey procedure was used to determine the critical values used for testing all pairwise contrasts (Maxwell & Delaney, 2003). Datasets 1 and 2 and datasets 3 and 4 differ only with respect to between-group variability (with datasets 2 and 4 having larger between-group mean differences). Participants rated dataset 2 as providing significantly stronger evidence than dataset 1, F(1,55)=71.60, p<.001 and rated dataset 4 as providing stronger evidence than dataset 3, F(1,54)=216.75, p<.001.

Datasets 2 and 4 and datasets 1 and 3 differ only with respect to within-group variability (with datasets 4 and 3 having smaller within-group standard deviations). Participants rated dataset 4 as providing stronger evidence than dataset 2, F(1,54)=22.23, p<.001, but they rated dataset 3 as providing *weaker* evidence than dataset 1, F(1,55)=11.61, p=.001.

Datasets 1 and 4 and datasets 2 and 3 differ with respect to both within and between-group variability. Dataset 4 has both a larger between-group mean difference and smaller within-group standard deviation than dataset 1. It was no surprise, then, that participants did rate dataset 4 as providing stronger evidence than dataset 1, F(1,54)=135.14, p<.001. Dataset 2 has a larger between-group mean difference but also a larger within-group standard deviation than dataset 3. Dataset 2 has a 2:1 ratio of between to within-group variability, whereas dataset 3 has a 4:1 ratio. Nonetheless, participants rated dataset 2 as providing stronger evidence than dataset 3, F(1,55)=102.67, p<.001.

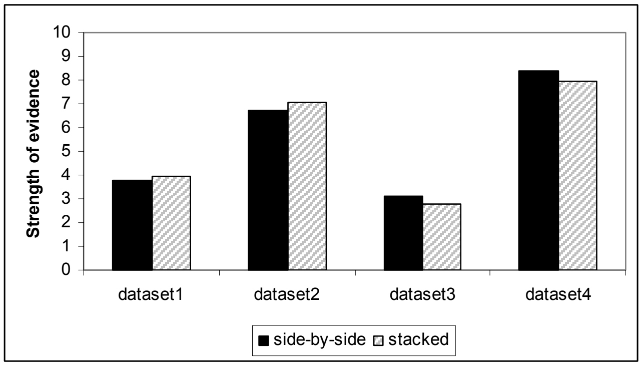

Figure 2. Mean ratings by Dataset and Format.

## Discussion

Format of the hypothetical datasets had no effect on student's ratings. Presenting data in a single, stacked column did not eliminate the bias of experienced students to consider larger, less reliable differences as stronger evidence than smaller, more reliable differences. Even when the data format discouraged computation of difference scores (i.e., one group's data stacked atop the other group's data), students still placed more emphasis on the magnitude of the difference in distances traveled by frozen and unfrozen balls than on the trustworthiness of data indicated by within-condition variability.

# Experiment 2

Thus far, the bias described by Trumpower and Fellus (2008) and in Experiment 1 has been demonstrated in students of different experience levels in statistics and has been shown to be robust with respect to data presentation formats. However, a stronger case for the persistence of the bias would be made if it could be demonstrated in the same students before and after instruction. In Experiment 2, the intuitive ANOVA abilities of a group of students was assessed during the first and last class of a graduate-level statistics course.

## Method

**Participants** Twenty-two students enrolled in an introductory, graduate level, statistics course at the University of Ottawa volunteered to participate in exchange for extra course credit. Alternatives were made available for those who chose not to participate. Several students were absent during the first class meeting and several other students dropped the course after attending the first week. A total of thirteen participants completed both the pre-and post-tests.

## Materials & Procedure

During the first day of the course, participants were asked to complete a pre-test that was exactly the same as the test booklet used in the stacked format condition of Experiment 1. On the last day of the course, participants were asked to complete a post-test that was identical to the pre-test except that it used a different cover story involving a hypothetical difference between baseball bats made of normal and genetically-engineered wood. Identical datasets were used on both the pre-test and post-test.

The likelihood that participants remembered the actual values of the datasets is minimal due to the length of time between pre and post-tests (3 months) and the frequency of other datasets encountered within the course during that time. The course itself covered both descriptive and inferential statistics, including the sign test, z-test for single samples, t-tests for independent and correlated groups, and correlation and regression.

## Results

A 2 Time (pre and post) x 4 Dataset repeated measures ANOVA was conducted on the participants' ratings. The main effect of Dataset was significant, F(3,33) =27.20, p<.001. Neither the main effect of Time nor the Time by

Dataset interaction was significant, F(1,11)=3.54, p=.087 and F(3,33)=1.95, p=.141, respectively (see Figure 3).

The main effect of Dataset was followed up by all possible pairwise comparisons among the 4 datasets. The Tukey procedure was again used to determine the critical values used for testing all pairwise contrasts. Recall that datasets 1 and 2 and datasets 3 and 4 differ only with respect to between-group variability (with datasets 2 and 4 having larger between-group mean differences). Participants rated dataset 2 as providing stronger evidence than dataset 1, F(1,11)=13.03, p=.004, and dataset 4 as providing stronger evidence than dataset 3, F(1,12)=27.31, p<.001.

Also, recall that datasets 2 and 4 and datasets 1 and 3 differ only with respect to within-group variability (with datasets 4 and 3 having smaller within-group standard deviations). Participants rated dataset 4 as providing stronger evidence than dataset 2, F(1,12)=26.76, p<.001, and dataset 3 as providing stronger evidence than dataset 1, F(1,11)=13.77, p=.003.

Finally, recall that datasets 1 and 4 and datasets 2 and 3 differ with respect to both within and between-group variability. Dataset 4 has both a larger between-group mean difference and smaller within-group standard deviation than dataset 1. Again, not surprisingly, participants rated dataset 4 as providing stronger evidence than dataset 1, F(1,11)=118.74, p<.001. Dataset 2 has a larger between-group mean difference combined with a larger within-group standard deviation than dataset 3 such that dataset 2 has a 2:1 ratio of between to within-group variability, whereas dataset 3 has a 4:1 ratio. Nonetheless, participants' ratings for datasets 2 and 3 were not statistically different, F(1,12)=0.15, p=.702.

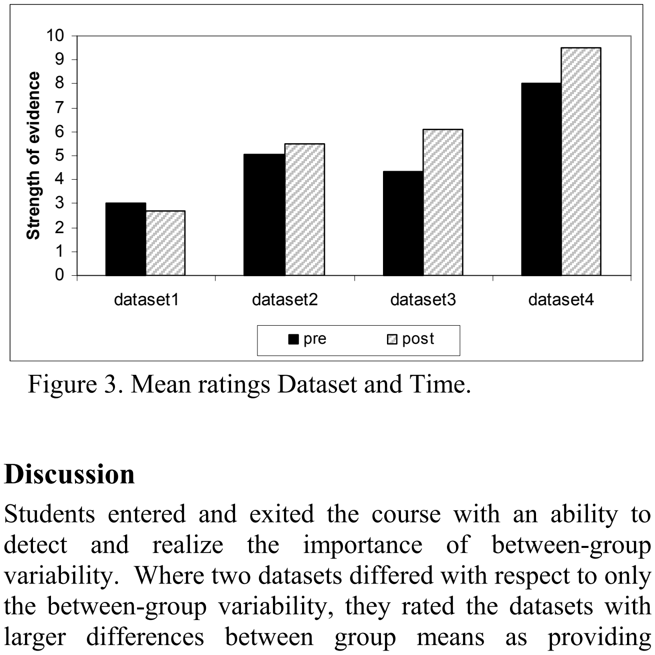

Figure 3. Mean ratings Dataset and Time.

## Discussion

Students entered and exited the course with an ability to detect and realize the importance of between-group variability. Where two datasets differed with respect to only the between-group variability, they rated the datasets with larger differences between group means as providing stronger evidence of differences between conditions.

They were also able to detect and to realize the importance of within-group variability to a certain extent. All other things being equal, students rated datasets with smaller within-group variability as providing stronger evidence of differences between the conditions. However, students again displayed an intuitive bias such that a small, yet reliable difference (dataset 3) was deemed no stronger than a larger, but less reliable difference (dataset 2). This bias was persistent, even after successfully completing a course covering inferential statistics. It might be noted, however, that the bias was somewhat weaker here than in Experiment 1. In Experiment 1, dataset 2 was rated as providing stronger evidence than dataset 3; here there was no significant difference.

## General discussion

In the present study we have shown that students have a robust intuitive bias to consider large between-group variability as more important than small within-group variability when detecting differences between conditions. The bias was shown to not be due to presentation format, and to be stubbornly resistant to change after instruction. At this point, the bias has been demonstrated in graduate and undergraduate students at different universities and in different academic concentrations.

The bias displayed in this study could be the result of several sources. One potential source is the contrast between statisticians' and laypersons' use of the word "difference". Statisticians may think in terms of statistically significant (i.e., reliable) differences, whereas laypersons are more likely to think in terms of practically important differences. While a four yard difference could be very reliable, and therefore significant to a statistician, it might not be construed as a meaningful difference to a layperson. Conversely, a fifty yard difference that is unreliable would not be considered significant to a statistician, but might still be seen as a potentially meaningful difference to a layperson. Put differently, laypersons may be more likely to spend money on a baseball bat that has the *potential* to hit the ball *50* yards further than they are to spend money on a baseball bat that is *guaranteed* to hit the ball only *4* yards further than normal bats. To use an analogy, the layperson may be like a gambler who is more likely to place their money on a bet with a potentially big, although uncertain, payoff than on a bet with a small, but more certain, payoff.

An additional source of the bias could be that students are using an intuitive confirmatory hypothesis testing procedure, along with the "bigger = more meaningful" layperson logic noted above. Preliminary analysis of students' verbal protocols indicates that many, in fact, did conduct a score-by-score analysis (as opposed to computing group means and variances), noting the number of scores from the experimental groups (i.e., the frozen and the genetically engineered bat conditions) that were larger than scores in the control groups. Large differences in favor of the experimental condition were noted as confirming the hypothesis, whereas small differences in favor of the experimental condition and differences in favor of the control condition were sometimes discounted as being due to error. For example, one participant in Experiment 1 explained their rating for dataset 2 by mentioning that 2 of 3 frozen golf balls traveled farther than the normal golf balls, but the distance of the third frozen golf ball (which did not go farther than any of the normal golf balls) may have been due to chance. This same participant, when explaining their rating for dataset 3 (in which all 3 frozen golf balls traveled farther than any of the normal golf balls, albeit by no more than 5 yards) simply mentioned that the differences were not significant.

The picture for statistics instructors is not completely bleak. Students do possess a good intuitive ability to detect and utilize between-group variability. The problem may be that this ability is too good. Students arrive in class with a strong tendency to focus on between-group differences and place too much emphasis on large differences. While it is easy for students to understand that between-group differences are (at least partially) due to the effect of the independent variable, it appears more difficult for them to understand that some of the observed between-group difference may be due to other random effects (i.e., error in measurement). As such, statistics educators may want to determine how to shift attention toward the relative importance of within-group variation. One strategy may be to encourage a more intensive focus on the actual sources of within group variation – i.e., help students understand why numbers within a given condition might vary and how these sources of variation signify unreliability in measurement that could lead to some of the observed between-group difference. Although we are sure that statistics instructors already make an effort to illustrate the difference between within- and between-group variability, our studies show that the standard approach is often not very effective.

Future studies will seek to pinpoint the source of the bias described in this study, with the goal of developing pedagogical strategies for overcoming it. It may be that students simply need to get a "feel" for analyzing variance without even thinking in a statistical sense.

## References

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General, 131(2),* 287-297.

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education,18,* 955-968.

Masnick, A.M. & Morris, B.J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79 (4),* 1032-1048.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison perspective (2nd ed.).* Mahwah, NJ: Lawrence Erlbaum Associates.

Onwuegbuzie, A.J. & Wilson, V.A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments—a comprehensive review of the literature. *Teaching in Higher Education, 8(2),* 195–209.

Peterson, C.R., & Beach, L.R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68 (1),* 29-46.

Trumpower, D.L., & Fellus, O. (2008). Naïve statistics: Intuitive analysis of variance. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Washington, DC: Cognitive Science Society, pp 499-503.