Effect of Formative Feedback via Interactive Concept Maps on Informal Inferential Reasoning and Conceptual Understanding of ANOVA

SAIT ATAS

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for
for the Ph.D. degree in Education

ABSTRACT

This study assessed the knowledge structure of undergraduate participants related to previously determined critical concepts of Analysis of Variance (ANOVA) by using Pathfinder networks. Three domain experts' knowledge structures regarding the same concepts were also elicited and averaged to create a referent knowledge structure. The referent knowledge structure served as a basis for formative feedback. Then, each participant's knowledge structure was compared with the referent structure to identify common, missing, and extraneous links between the two networks. Each participant was provided with individualized written and visual, and multi-media feedback through an online Concept Mapping tool based on the principals of formative assessment and feedback in an attempt to increase their conceptual knowledge of ANOVA.

The study was conducted with 67 undergraduate participants from a mid-size university in the United States. Participants completed two data collection tools related to the critical concepts of ANOVA. Later, three different types of feedback around the critical concepts were given to participants in three stages. First, each participant was given visual feedback as a result of the comparison between their own knowledge structures and the referent knowledge structure to highlight similarities and differences between the two. Then, participants were provided with individualized written and multi-media feedback to emphasize conceptual understanding behind ANOVA procedures. This procedure was followed by the re-assessment of participants' reasoning ability related to ANOVA and knowledge structures related to critical concepts to measure the effect of the intervention.

Results suggest that participants both in control and intervention groups had the same level of statistics experience and anxiety before this study indicating that randomization of participants into two different groups was successful. Moreover, women participants reported a statistically significant higher level of statistics anxiety than men, however, it seems that this small difference did not limit their ability to perform required statistical tasks. Further, findings revealed that participants' conceptual knowledge related to critical concepts of ANOVA increased significantly after the individualized feedback. However, the increase in the conceptual understanding did not help participants to transform this knowledge into more formal understanding related to procedures underlying ANOVA. Moreover, even though, previous similar studies suggest that participants are consistent in using a single strategy for making inferential reasoning across datasets, in the present study, qualitative data analysis revealed that statistics learners demonstrate diverse patterns of inferential reasoning strategies when they were provided with different size of datasets each with varying amount of variability. As a result, findings support the use of an extended framework for describing and measuring the development of participants' reasoning ability regarding consideration of variation in statistics education.

Keywords: Informal Inferential Reasoning, Conceptual Knowledge, Formative Assessment, and Feedback

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisor Prof. David Trumpower for the continuous support for my Ph.D. studies and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. studies.

Also, I would like to thank the members of my thesis committee: Prof. Bethany White, Prof. Christine Suurtamm, Prof. Timothy Goldsmith, and Prof. Tracy Vaillancourt, for their encouragement, insightful comments, and recommendations.

Finally, I am grateful to my parents who encouraged me and prayed for me throughout the time of my studies. Most importantly, I wish to thank my loving and supportive wife, Ecenur who provides unending inspiration, love, and guidance with me in whatever I pursue. I would not imagine successful completion of this thesis without her continuous support and encouragement.

This thesis is dedicated to my parents, my wife, Ecenur, and my two wonderful daughters, Bahar and Berranur.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

If you are an instructor of statistics or a student in a statistics classroom at the undergraduate level, it is likely that you have heard many stories from students about their fear of learning about core concepts in statistics such as standard deviation and variance. The term “variance” has been considered as one of the fundamental ideas in statistics education. Supporting this idea, MacGillivray (2004) has described statistics as “science of variation”. Indeed, Analysis of Variance (ANOVA) is one of the most commonly used hypothesis tests in statistics. As a hypothesis testing method in statistics, ANOVA helps people to determine the strength of evidence in support of some hypothesis based on sample data, thus allowing one to make judgements about general populations (e.g., Brand B shoes are perceived as being better than Brand A shoes). However, recent research in statistics education indicates that most undergraduate students studying statistics have difficulty in understanding core concepts (e.g., Delmas, Garfield, Ooms, & Chance, 2007). There are several reasons for this difficulty including instructors being more focused on the computational aspects of statistics instead of addressing the rationale behind statistical procedures (Garfield, 1995) and students’ difficulty in conceptualizing some core ideas in statistics, such as probability and variability (Garfield & Ben-Zvi, 2007; Trumpower, 2013).

Informal Inferential Reasoning activities have been emphasized as a means to increase students’ conceptual knowledge related to core concepts in statistics, such as sampling, confidence interval, and probability, without following formal calculations. However, recent research found that students’ reasoning ability in Informal Inferential Reasoning activities related to more advanced topics like ANOVA is still faulty (Obrecht, Chapman, & Suarez 2010;

Trumpower, 2013, 2015), even after successfully completing courses in statistics (Kelly, Sloane, & Whittaker, 1997).

Recently, it has been shown that formative use of Structural Assessment of Knowledge, that is, assessing and providing feedback on students' structural knowledge, is useful in increasing students' conceptual understanding in other domains, such as science (Sarwar, 2011; Sarwar & Trumpower, 2015). In this study, the effectiveness of feedback on students' Informal Inferential Reasoning ability for improving conceptual understanding related to ANOVA will be explored. It can be argued that the use of Informal Inferential Reasoning activities related to ANOVA and providing feedback on students' Informal Inferential Reasoning ability through Structural Assessment of Knowledge is a viable means of enhancing students' conceptual understanding of ANOVA, one of the central ideas in statistics education.

Conceptual Framework

The framework of this dissertation utilizes three different conceptual perspectives: the distinction between procedural and conceptual knowledge, Informal Inferential Reasoning in statistics, and Structural Assessment of Knowledge. The purpose of this particular section is to provide a glimpse of how these three conceptual perspectives guide the overall study. However, these three ideas are discussed in more detail in later sub-sections within the focused literature review.

First, discussion around procedural and conceptual knowledge provides insights into the origins of these two terms and utilizes a comprehensive understanding of the distinction between the terms. In brief, procedural and conceptual knowledge refer to two different, yet related, types of knowledge. Although procedural knowledge is the focal point of most students, and many instructors, of statistics, conceptual knowledge may be equally or even more important. This distinction between the two types of knowledge guides and impacts the overall design as well as the focused literature review, data collection and analysis, and results sections of the present study. For instance, the comprehensive understanding of conceptual and procedural knowledge leads the researcher to the development of assessment tools to ensure that both procedural and conceptual knowledge of students are assessed in this study. As such, the Structural Assessment Tool used in the study not only assesses students' knowledge structures related to conceptual understanding using previously determined critical concepts of ANOVA but also provides immediate multi-media feedback in the same nature to enhance conceptual understanding of ANOVA. As a second theoretical perspective, Informal Inferential Reasoning is taken into consideration as a potential method for supporting the development of conceptual knowledge related to ANOVA in an engaging and informal way to prepare students for more formal learnings related to the procedures of ANOVA. In this dissertation, Informal Inferential Reasoning is used as an intervention to support underlying understanding of the ANOVA procedures without requiring students to go through any *formal* procedural calculations. That being said, the Informal Inferential Reasoning tool used in the present study will be a precursor to a more formal learning of ANOVA procedures by providing students with opportunities to practice underlying critical concepts of ANOVA without following any formal statistical procedures. Subsequently, this will help students build a solid conceptual understanding of crucial concepts with feedback provided as a result of structural assessment which, in return, helps to utilize a smooth transaction to the procedural knowledge of the same concepts using Informal Inferential Reasoning practices.

As another example of how the conceptual framework of this study guides the study, the multi-media feedback prepared by the researcher as an intervention features the examples of both conceptual and procedural knowledge to highlight the importance of both knowledge types for a complete understanding of ANOVA. Moreover, as mentioned, the use of two different assessment tools in the study leads to the generation of various types of data (i.e., quantitative and qualitative) for each student and, thus the use of various data analysis techniques for each type of data. Overall, it can be said that the distinction between conceptual and procedural knowledge serves as an umbrella theoretical perspective that informs the reaming parts of the study. That being said, the literature on Structural Assessment of Knowledge provides guidance on how to effectively assess knowledge organization of undergraduate students to identify their existing conceptual knowledge related to ANOVA and on how to provide students with different types of feedback to help them improve both conceptual and procedural knowledge of ANOVA. Further, Informal Inferential Reasoning activities helps learners to consolidate their conceptual understanding as well as to assess students' formal reasoning abilities regarding ANOVA.

Focused Literature Review

People often make inferences in everyday life. For instance, we may make decisions about what time to leave home based on our past experiences of traffic at various times of the day or we may make decisions on which computer to buy based on customer ratings. We typically use sample data on hand (in these examples, data on traffic and customer ratings) to make inferences about the future (as in the first situation) or to generalize about larger groups of things (as in the second). When referring to statistics, this type of reasoning is called *inferential reasoning*. Inferential reasoning allows one to go beyond summarizing the sample data to make inferences about unknown populations (Coladarci, Cobb, Minium, & Clarke, 2011). It has become a central concept in statistics education and is strongly suggested both for K-12 and undergraduate education to promote students' more general reasoning skills (Coladarci et al., 2011; National Council of Teachers of Mathematics, 2009; Weinberg, Weisner, & Pfaff, 2010). Further, as an influential report in the field of statistics education, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) 2016 report also recommends focusing on teaching statistical thinking in introductory post-secondary statistics courses. As this dissertation focuses on students' difficulty with statistical thinking related to ANOVA as an important inferential reasoning method, it is pertinent to clearly explain the logic behind ANOVA in a statistical sense and to present the actual formal calculations required to conduct ANOVA. Thus, we may have a comprehensive understanding of ANOVA before delving into the problems faced by students.

Analysis of Variance (ANOVA)

The rationale behind ANOVA. As an important method of inferential reasoning, ANOVA helps people to decide "whether or not an observed difference between different groups can plausibly be attributed to chance" (Moore & Notz, 2014, p. 515). In other words, ANOVA helps to determine the likelihood that an observed difference between the means of two or more groups would have occurred if due to only random factors. To illustrate this point, suppose that two scientists conducted an experiment to test a baseball bat made with genetically engineered wood that they believe can make baseballs travel further than regular wooden bats. In their experiment, they use a robotic arm to hit baseballs with bats made of the two types of wood, all with the exact same force, after which the distance that each ball travels is measured. Now, suppose that the distances (in meters) of the balls hit with the two different types of baseball bats were as follows:

Distances of baseballs by type of bat

Type of wooden bat	
Normal	**Genetically engineered**
230	420
370	280
300	350

Note. Measurements are given in meters.

In order to make a dependable and reliable decision following the logic of ANOVA, first the scientists would need to calculate the average distance traveled by each group of balls then calculate the difference between the average distances traveled by each group. In this example, the balls hit by the genetically engineered wood bat went 350m on average, whereas the balls hit by the normal wood bat went only 300m on average. Thus, the mean distance of the group of balls hit by the genetically engineered bat was 50m longer than the mean of the group of balls hit by normal wood bat, on average. However, before concluding that genetically engineered bats are actually better than normal wood ones, we need to recognize that not all of the balls hit with the same force and with the same type of baseball bat travelled the same distance. For instance, the three balls hit by the normal wood bat deviate by 70m from one to another. This might have been due to *uncontrolled factors*, including wind, air temperature, humidity, etc. These types of factors are called *random factors*. The distances of the balls hit by the genetically engineered also varied by 70m from one another, on average. Considering that the 70m average deviation within the groups of balls was actually larger than the the 50m difference between the means of the groups, it seems plausible that the observed difference between the sample of balls hit by genetically engineered and normal wood bats in our example could have been the result of random factors alone. Thus, determining the size of effect due to random factors on the distances travelled by the baseballs is important in determining if one type of bat is better than the other. Specifically, the ratio of between-group difference to within-group difference helps decide whether or not a significant difference exists between the distances of the two types of bats. Bigger ratios would provide stronger evidence to support a decision that one type of bat is better than the other. The baseball example has been used as a model to illustrate the underlying logic behind ANOVA. However, this logic could be applied to any situation where one needs to make a comparison between the sample means of two or more groups in order to make inferences about the unknown populations from which the sample data were obtained. Given the logic behind ANOVA, let us now have a closer look at the formal calculations needed when conducting an ANOVA.

Formal procedures in conducting ANOVA. In the previous section, I have been informally estimating the observed between-group variation (i.e., the difference that exists between groups) and observed within-group variation (i.e., the differences that exists within each group), and comparing the between-group variation to the within-group variation of two groups of data. In a formal sense, this procedure is called an ANOVA. ANOVA is based on the ratio of between-group variation to within-group variation, and this ratio is referred to as the F-ratio. F-ratio gives an idea about evidence against the null hypothesis that there is no effect of an independent variable on a dependent variable. In our baseball example, the F ratio provides evidence against the claim that the type of bat has no effect on distances travelled by baseballs. The bigger the F-ratio, the stronger is the evidence to reject the null hypothesis and conclude that the type of bat actually has an effect on distances travelled by baseballs. Below are the equations to determine the F-ratio.

$$F = \frac{variability\ due\ to\ independent\ variables\ and/or\ random\ factors}{variability\ due\ to\ random\ factors\ alone}$$

$$F = \frac{between\ group\ variation}{within\ group\ variation}$$

$$F = \frac{MS_B}{MS_w}$$

$$F = \frac{SS_B/df_B}{SS_W/df_W}$$

$$F = \frac{\sum n_k\,(\bar{x}_k - \bar{\bar{x}})^2/a-1}{\Sigma(n_k-1)s_k^2/(N-a)}$$

Where;

k = specific group (i.e., specific level of the independent variable)

n_k = number of observations or groups in the kth group

N = total number of scores across all groups

$\bar{x}_k$ = sample mean of the kth group
$\bar{\bar{x}}$ = overall mean (i.e., mean of the group means)
a = total number of groups
s_k = sample standard deviation of the kth group
df_B = between groups degrees of freedom = a − 1
df_W = within group degrees of freedom = N − a
SS_B = between groups sum of squares = $\sum n_k\,(\bar{x}_k - \bar{\bar{x}})^2$
SS_W = within group sum of squares = $\Sigma(n_k - 1)s_k^2$
MS_B = between groups mean square = SS_B/df_B
MS_w = within groups mean square = SS_W/df_W

Now if we apply the equations to the data from the baseball example, it will look as follows:

Type of wooden bat	
Normal	**Genetically Engineered**
230	420
370	280
300	350
Mean= 300	Mean=350
Standard deviation=70	Standard deviation=70

Note. Measurements given in meters.

Overall mean = $\bar{\bar{x}}$=325

First, between groups mean square will be calculated:

$$MS_B = \frac{SS_B}{df_B}$$

$$MS_B = \frac{\sum n_k(\bar{x}_k - \bar{\bar{x}})^2}{a-1}$$

$$MS_B = n_{normal}(\bar{x}_{normal} - \bar{\bar{x}})^2 + n_{genetically\ engineered}\left(\bar{x}_{\text{genetically engineered}} - \bar{\bar{x}}\right)^2 \Big/ a-1$$

$$MS_B = (3(300-325)^2 + 3(350-325)^2) \Big/ 2-1$$

$$MS_B = 3750$$

Second, within group mean square will be calculated:

$$MS_w = SS_W \Big/ df_W$$

$$MS_w = (n_k - 1)s_k^2 \Big/ (N-a)$$

$$MS_w = (n_{normal} - 1)s_{normal}^2 + \left(n_{genetically\ engineered} - 1\right)s_{genetically\ engineered}^2 \Big/ (N-a)$$

$$MS_w = (3-1)70^2 + (3-1)70^2 \Big/ (6-2)$$

$$MS_w = 4900$$

Then, the ratio of between groups mean square, MS_B, to within group mean square, MS_w will be calculated:

$$F = \frac{MS_B}{MS_w}$$

$$F = \frac{3750}{4900}$$

$$F = 0.765$$

In brief, ANOVA examines the difference that exists between groups, which could be due to both random factors *and* any effect of the independent variable, relative to the difference that exists within groups, which can only be due to random factors. For instance, while the between group mean square, MS_B , in the third formula analyzes how much, on average, the group mean varies from the overall mean of all groups in the experiment, the within group mean square, MS_w, analyzes how much each individual score within a group, on average, differs from the group mean to which that score belongs. As mentioned, the variability between groups refers to the combined effect of both the independent variable (i.e., type of bat) and random factors (i.e., wind, air temperature, etc.). In contrast, the variability within groups refers only to the effect of random factors. Collectively, the ratio of between-groups variability to within-group variability shows how much greater the combined effect is than the effect of random factors alone.

$$F = \frac{variability\ due\ to\ independent\ variables\ and/or\ random\ factors}{variability\ due\ to\ random\ factors\ alone}$$

To the extent that the influence of the combined effects is larger than the influence of random factors alone, the ratio will be larger, and there is stronger evidence to suggest that the independent variable had an effect. Thus, the result of an ANOVA allows us to make an inference about whether the independent variable (i.e., type of bat) in an experiment has a real effect on the dependent variable (i.e., distance travelled by baseballs) or if it is likely that only random factors could have caused the difference between group means.

As seen in the calculations above for the baseball example, the *F-ratio* is smaller than 1, indicating that the combined effect of the independent variable and random factors (i.e., as indicated by the mean square between groups, MS_B) is not larger than the effect of random factors alone (i.e., as indicated by the mean square within group, MS_w). In other words, 0.765 does not provide strong evidence against the hypothesis that the type of ball has no effect on the distances travelled by the baseballs. Now, we can recall that I arrived at the same conclusion informally when explaining the logic behind ANOVA by using the baseball example at the beginning of the section. While analyzing the data sets informally, I noted that the between-group difference in group means of 50m is smaller than the average within-group differences of 70m. Informally, I used the difference in group means as an easy to estimate proxy for MS_B and the average of the group standard deviations as an easy to estimate proxy for MS_W. Thus, the ratio of between-group variation to within-group variation was smaller than 1 and, again, did not provide strong evidence for an effect of the type of bat.

As seen above, a valid decision-making process in ANOVA typically not only involves making calculations and applying equations, but also requires understanding the rationale behind these procedures. In particular, turning back to my ANOVA example on baseballs, while *procedural understanding* requires making calculations and following steps, *conceptual understanding* of the process requires reasoning, such as what is the logic behind these calculations or following steps, and relates to questions like: why is the ratio of difference between groups to difference within groups important? Or why not consider only the difference between groups and make a decision based on the difference between groups? (i.e., in the normative logic of ANOVA, both the between-group and within-group differences are equally important simply because the difference within groups tells us about the consistency of the observed scores). Following the steps for ANOVA without considering the logic behind it may prevent the use of inferential reasoning in an effective manner. Therefore, the possession of both conceptual understanding and procedural knowledge at the same time is considered to be the best condition with respect to the efficient utilization of inferential reasoning. Similarly, the GAISE report (2016) also encourages statistics educators to focus on both types of knowledge by highlighting underlying conceptual understanding behind statistical procedures in introductory statistics classrooms.

However, there is ample evidence in the literature to demonstrate that although students are good at following the formal procedures of inferential reasoning, they have difficulties in grasping the underlying logic of these procedures and in applying these procedures to new contexts (Ben-Zvi & Garfield, 2005; Trumpower, 2013). For instance, Garfield and Ahlgren (1988) describe the experiences of statistics educators in education and social sciences, wherein most of their students in introductory statistics courses do not understand the concepts being studied. Likewise, Weinberg et al. (2010) point out that students at all levels have difficulties with inference-related ideas in statistics. To illustrate, Delmas, Garfield, Ooms, and Chance's (2007) study demonstrates the difficulty that students face pertaining to the conceptual understanding of the fundamental concepts in statistics. They developed a measurement tool, the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) and administered the test before and after an introductory statistics course. The authors concluded that students demonstrated limited conceptual understanding related to inferential reasoning, including probability and variability.

In another study, in an attempt to understand graduate students' informal reasoning about ANOVA, Trumpower (2013) investigated the reasoning that students use when performing an informal ANOVA task. Consistent with the nature of Informal Inferential Reasoning activities, students were not asked to perform any calculations in this task, therefore it is called informal ANOVA (i.e., iANOVA). Specifically, in the iANOVA task used in the study, a cover story was shown to participants in which two researchers were conducting an experiment to test their hypothesis that frozen golf balls can go farther than normal balls. Different hypothetical results, each listing the distances traveled by 3 frozen and 3 unfrozen golf balls, were presented on a worksheet, much like the baseball example provided earlier. Then, participants were asked to rate their confidence on a 10-point scale in the claim that frozen golf balls go farther than normal golf balls. The author concluded that students tended to perceive the average difference between normal and frozen balls as being more important than the consistency of those differences, even though both are equally important in a formal statistical sense of ANOVA. In another study, Trumpower (2015) reported similar results, highlighting that the intuitive reasoning used by participants was not consistent with the normative logic of ANOVA. Participants, even those with a statistics background, tended to place more emphasis on the difference between groups than consistency within groups. Moreover, in their attempt to understand graduate students' reasoning processes with different types of data presentation in an intuitive ANOVA task, Atas, Trumpower, and Filiz (2013) also concluded that students' reasoning about ANOVA in a given task is not consistent with the underlying rationale behind the formal ways of doing ANOVA.

The results from these studies highlight that informal reasoning used by students while participating in informal ANOVA tasks is contrary to the logic of ANOVA. Students perceive inconsistent larger differences as being more important than consistent smaller differences between groups even if the small but consistent differences would generate more significant results in a formal statistical sense. However, as Trumpower (2015) stated, the reason behind students' difficulties with the task is not fully known. In the present study, a revised iANOVA task including a different cover story as well as larger and smaller datasets was used to determine the reason behind these difficulties. A detailed description of the task will be presented in the methodology section of this study (see Appendix A for three different iANOVA tasks).

Altogether, across time and researchers, it has been shown that difficulties pertaining to increasing the conceptual understanding behind formal procedures related to inferential reasoning have been around for a long time and still exist in the statistics literature. However, why is it difficult for students to develop conceptual understanding in statistics courses? Indeed, researchers have suggested reasons for the various challenges identified in the literature. The most common reasons include an excessive focus on the formal calculations in statistics courses (Garfield, 1995), the use of technical language in statistics courses (Rumsey, 2009), and negative attitudes towards conceptualizing some ideas in statistics, such as probability and variability (Gal & Ginsburg, 1994; Lancaster, 2012).

The first challenge is related to focusing too heavily on the mechanical parts of statistics in classrooms instead of paying attention to increasing conceptual understanding behind the mechanical parts. Garfield and Ben-Zvi (2008) highlight that "traditional approaches to teaching statistics focus on skills, procedures, and computations, which do not lead students to reason or think statistically" (p. 7). In 2009, Makar and Rubin demonstrated this challenge when they worked with primary school teachers to investigate the development of statistical thinking and reasoning at the primary school level. The researchers found that the students, whose teachers focus too much on procedural knowledge such as graphing skills and calculations during teaching practices instead of paying attention to the logic behind the calculations, were not able to go beyond summarizing the data to make inferences about populations.

The second challenge revolves around the terminology/language used in statistics classrooms. While investigating the importance of language used in statistic classrooms, Kaplan, Fisher, and Rogness (2009) indicated that the domain specific language used in classrooms could make a domain appear more difficult for students than it is. Then, the authors drew attention to the lexical ambiguity in statistics whereby terms were used differently than in everyday usage. Lexical ambiguity may lead students to make false associations between words they are familiar with from daily language and the same ones used in a specific domain with a different meaning. For instance, while the word "average" is used to describe the mean of a dataset in statistics, it is mostly used to mean "typical" or "normal" in the daily usage. Similarly, while the term "confidence" is used to refer to the level of assurance in the daily life; however, in contrast, the term involves a probabilistic sense which does not refer to being certain in statistics. Similarly, as highlighted by Rumsey (2009), the complicated nature of formal terminology in statistics classrooms, such as sample, population, variation, spread, and sampling distribution, decreases the ability to process information easily. Therefore, use of non-technical language at the beginning of the courses may be beneficial to increase conceptual knowledge (D. L. Trumpower, personal communication, October 2, 2013). Supporting this argument, in a study of the use of language by pre-service teachers, Makar and Confrey (2005) reported that the utilization of informal language in statistics classrooms helped pre-service teachers to articulate their understanding of variation in a way that the technical language, such as interquartile range or standard deviation, did not. Furthermore, they claimed, when using non-standard language, the prospective teachers could integrate ideas of distribution and variation.

The final challenge relates to students' impressions of statistics as being a challenging subject matter. As noted by Trumpower (2015), many students perceive statistics as difficult, regardless of their demographic background. Lancaster (2012), for instance, reported that “most students at university level develop antipathy towards statistics” (p.4). The negative attitude toward learning statistics might be result of a larger focus on formal calculations and the use of technical language in statistics classrooms, as described in the first two challenges. While reviewing the role of affect and attitudes in the learning of statistics, Gal and Ginsburg (1994) noted that non-cognitive factors such as motivation and attitudes can impede learning of statistics and limit the extent to which students develop meaningful intuitions in statistics classrooms.

These are interconnected explanations for the difficulties pertaining to conceptual understanding of inferential reasoning identified in the literature. Moving away from formal procedures to a less formal conceptual approach with use of non-technical terminology might lead students to improve their conceptual understanding of complex concepts in statistics and to develop more positive attitudes towards learning statistics. Drawing on the difficulty identified in the literature, as well as the experience of students who have successfully completed an introductory statistics course, it is possible to conclude that students continue to have problems with the underlying logic behind statistical procedures. Therefore, there is a need for going beyond the *procedural knowledge* of statistical ideas and to focus on increasing the *conceptual understanding* behind the formal procedures used in statistics.

In order to encourage a deeper conceptual understanding behind statistical procedures, there has been an attempt to change the focus of statistics education in the last decade (Rossman, 2008; Trumpower & Fellus, 2008; Makar & Rubin, 2009). This movement has been described as moving away from computational “recipes” and statistical tools to statistical processes in statistics (Allen, Folkard, Lancaster & Abram, 2010; Makar & Rubin, 2009). As a result of this movement, Informal Inferential Reasoning (IIR) has recently been proposed as a potential means of developing the conceptual knowledge that underlies formal procedures in statistics education. IIR activities also align very well with the development of informal statistical inference (ISI), which is a broader term to describe the developmental stages of reasoning ability related to statistical inference (e.g., see Makar et al., 2011, Ben-Zvi et al., 2012, Baker & Derry, 2011). First, the origins of the terms "procedural understanding" and "conceptual understanding" as well as the distinction between these terms will be presented. This will be followed by an attempt to discuss the potential of Informal Inferential Reasoning activities to foster conceptual understanding of statistical procedures.

Procedural and Conceptual Knowledge

The distinction between the terms "procedural" and "conceptual" knowledge was initially made by different philosophers. For instance, Michael Polanyi's seminal books titled "Personal Knowledge: Towards a Post-Critical Philosophy" (1958) and "The Tacit Dimension" (1966) are early contributors to the concepts of procedural and conceptual knowledge. In the introduction of one of Polanyi's books, "Personal Knowledge", he challenges the conventional view of the completely "objective scientists" by stating that "Into every act of knowing there enters a passionate contribution of the person knowing what is being known and that this coefficient is no mere imperfection but a vital component of his knowledge" (pg., IV). Similarly, Polanyi's most famous quote "We can know more than we can tell" indicates that all types of knowledge involves a degree of personal knowledge which is later called "tacit knowledge". To Polanyi, tacit knowledge is gained from experiences and cannot be easily articulated to others. He further explains that even though people know how to do things they may not always be able to articulate to others why they do what they do. Therefore, he claims that all types of knowledge have a tacit dimension and the degree of this tacitness varies (Polanyi, 1958).

The question here is then "Is it possible to transfer tacit knowledge between individuals?". Polanyi (1958) answers this question by noting that "tacit knowledge" can be made more "explicit knowledge" with the use of appropriate language in order to transfer knowledge between individuals. Polanyi explains the knowledge transition process between tacit and explicit knowledge based on the interaction between the two as he says, "we define the process by which the tacit co-operates with the explicit, the personal with the formal" (pg., 90). In a similar vein, while examining the current relevance of Polanyi's work to organizational context, Grant (2007) emphasizes that tacit knowledge is usually introduced as "implicit knowledge" and this knowledge is usually compared with "explicit knowledge". Sternberg (1999) also relates implicit knowledge to tacit knowledge due to associated difficulty with verbalizing this type of knowledge. Drawing on Polanyi's work, Grand (2007) situates tacit and explicit knowledge on a continuum where it is not easy to disconnect these two knowledge types; however, the two ends of the continuum symbolize two different types of knowledge. Based on this continuum, knowledge begins at an implicit level, or procedural level, and over time becomes increasingly explicit or conceptual with the use of appropriate means or vice versa. As indicated by Neuweg and Fothe (2011) making tacit knowledge explicit is important in order to illuminate the knowledge base behind personal practices.

From these considerations, it seems the origins of "conceptual and procedural knowledge" go back to the discussion around tacit knowledge, implicit knowledge, and explicit knowledge. That being said, every type of knowledge involves a degree that is implicit which is usually defined as intuitive knowledge about the fundamental structure of a complex phenomenon (Reber, 1993).

To Polanyi, this implicit part of knowledge is associated with “tacit knowledge” which is gained through personal experiences and is mostly difficult to articulate to others. However, implicit knowledge can be transferred through use of appropriate language and instruction and in return the knowledge becomes “explicit knowledge” to the person himself/herself and to others as well. Further, the process of verbalizing implicit knowledge requires making the logic behind underlying concepts clear to others. Thus, this process is more likely to support conceptual understanding of procedures or "conceptual knowledge” as Baroody, Feil, and Johnson (2007) highlight that explicitly justifying connections between different types of knowledge makes knowledge more effective.

From the mathematics research point of view, Hiebert and Grouws (2007) also support articulating implicit knowledge as a means to promote students’ understanding of the conceptual structure of a given topic. According to Hiebert and Grouws (2007), this process could include “discussing the mathematical meaning underlying procedures, asking questions about how different solution strategies are similar to and different from each other” (p. 384). It is also worth noting that the distinction between these knowledge types does not imply that these knowledge types are totally different from each other. From this point of view, it could be argued that the more educators/instructors highlight the rationale behind mathematical/statistical equations/formulae taught in classrooms, the more solid understanding would be obtained by learners. Conversely, the more educators/instructors focus on only presenting the steps required to make calculations, the more automatic and rote performance would dominate the learning environment. Thus, one would expect that such a knowledge transferred from one individual to another one would be weak and limited to certain contexts.

Further to discussions around conceptual and procedural knowledge from physiological perspectives, these terms were introduced to educators by Hiebert and Lefevre (1986). Although Hiebert and Lefevre described conceptual knowledge as a network of knowledge in which all the pieces of information are strongly linked to each other, they define the term procedural knowledge as being knowledge of rules or procedures for solving problems (Hiebert & Lefevre, 1986). Similarly, procedural knowledge was also described by Schneider and colleagues as "the ability of executing action sequences to solve problems" (Schneider, Rittle-Johnson, & Star, 2011, p. 1525). Based on these definitions, it can be concluded that while procedural knowledge represents the knowledge of procedures, conceptual knowledge requires understanding the rationale behind these procedures.

Recent research on statistics and mathematics education supports the idea that learners should develop both procedural and conceptual understanding of a particular topic to fully comprehend the relevance and logic of statistical procedures and be able to transfer knowledge to new contexts. However, the flexibility side of conceptual understanding which allows learners to transfer the knowledge to different situations easily makes conceptual understanding more critical in a given context than procedural understanding which is mostly tied to a specific context and is difficult to transfer. Moreover, Hiebert and Lefevre (1986) highlight that procedural knowledge involves knowledge that is weak in relation and is tied closely to the context in which it is learned. Because procedural knowledge is not linked to other knowledge, it can only be applied in very similar contexts. Thus, procedural knowledge cannot be easily generalized to other situations without conceptual understanding. Conversely, conceptual understanding can be considered as a network of knowledge in which all the pieces of information are strongly linked and easy to recall when needed. Therefore, this type of knowledge is more flexible and transferable to other contexts (Hiebert & Lefevre, 1986). Thus, it is important for students to have conceptual understanding in addition to procedural knowledge as Baroody et al. (2007) argued that “linking procedural to conceptual knowledge can make learning facts and procedures easier, provide computational shortcuts, ensure fewer errors, and reduce forgetting (i.e., promote efficiency)” (pg. 127).

Evidence from research on teaching mathematics also supports that conceptual understanding helps to make procedural knowledge stronger and easier to transfer compared to procedural knowledge alone. For example, Rittle-Johnson and Alibali (1999) investigated the effect of different types of instructions (procedural versus conceptual) on students’ conceptual understanding of the concept of equivalence in math. Results showed that conceptual instruction focusing on the deep understanding of the related concepts helped grade 4 and 5 students to generalize their new knowledge more broadly and transfer the knowledge easily into new equation problems. In another study, Rittle-Johnson, Siegler, and Alibali in 2011 conducted two experiments to measure 5th- and 6th-grade students’ conceptual and procedural knowledge of decimal fractions before and after an instructional intervention. In experiment 1, the researchers provided feedback-type support in order to increase students’ conceptual and procedural knowledge while they manipulated the amount of support in experiment 2 in order to improve and assess students’ procedural knowledge only. Results indicated that conceptual and procedural knowledge did not develop in a linear fashion where improvement in one domain of knowledge supports the improvement in the other, but instead they developed iteratively. Findings also highlighted that conceptual knowledge leads to more flexibility which allows students to easily transfer this type of knowledge to similar situations.

Derived from the discussions around the origins and the description of the terms procedural and conceptual knowledge, one may tend to acknowledge that while there are different kinds of knowledge such as “procedural knowledge” and “conceptual knowledge”, there are also different depths of knowledge as anyone can possess a deep understanding of any procedure or a superficial understanding of concepts being studied. As one may not be able to define the first as being a weak understanding, the second also would not be considered as a strong understanding of the phenomenon just because it entails knowledge of concepts. Supporting this argument, Star (2005) also underlines that not all procedural knowledge is necessarily superficial or rote while criticizing the current perception of procedural knowledge by recent researchers. In a similar vein, de Jong and Ferguson-Hessler (1996) refer to the “quality of knowledge” and attempt to define it by using the term “depth of knowledge" where the term depth refers to the extent that knowledge is “firmly anchored in a person’s knowledge base and external information has been translated to basic concepts, principles, or procedures from the domain in question” (de Jong & Ferguson-Hessler, 1996, p. 107).

That being said, in this dissertation, I conceptualize *procedural knowledge/understanding* as the ability to routinely apply equations of ANOVA without necessarily considering the relation between concepts existing in the equations, whereas *conceptual knowledge/understanding* refers to understanding of the underlying logic behind statistical procedures such as ANOVA, and is comprised of knowledge of the relationships among relevant concepts. Furthermore, I consider the relation between conceptual and procedural knowledge as being iterative and dynamic where each type of knowledge supports each other during the learning process. From the statistics education point of view, the possession of both conceptual understanding and procedural knowledge at the same time is considered to be the best condition in terms of building solid statistical inference ability. Now, Informal Inferential Reasoning will be presented as a potential means of developing the conceptual knowledge that underlies formal procedures in statistics education.

Informal Inferential Reasoning

Informal Inferential Reasoning refers to the making of inferences or judgements about populations based on evidence drawn from samples, without following any formal statistical procedures (Zieffler, Garfield delMas, & Reading, 2008). Although the definition provided by Zieffler et al. refers to moving beyond data and drawing evidence-based conclusions, Pfannkuch (2011) makes further explication of the Informal Inferential Reasoning process by highlighting that Informal Inferential Reasoning needs to involve consideration of variation and incorporating some degree of uncertainty into conclusions since one cannot be certain of inferences about populations that are made based on sample data.

Drawing on the definition of Zieffler et al. (2008) and the additional perspectives on Informal Inferential Reasoning provided by Pfannkuch (2011), three important components of Informal Inferential Reasoning become apparent: (1) the making of inferences that go beyond sample data (Zieffler et al., 2008); (2) use of sample data as evidence to articulate inferences (Zieffler et al., 2008); and (3) recognition of variation and thus incorporation of uncertainty into inferences about unknown populations (Pfannkuch, 2011). Designing Informal Inferential Reasoning activities/intervention based on these three critical components described above may support the development of Informal Statistical Inference (ISI) which is also suggested to improve reasoning ability of students in introductory statistics classrooms for K-12 and higher education institutions (see Makar et al., 2011, Ben-Zvi et al., 2012, Baker & Derry, 2011). The steps involved in the development of Informal Statistical Inference such as moving beyond sample data to make inferences about populations, using the sample data as evidence for inferences, and recognizing the probability and variability in the sample data align quite well with the informal ANOVA intervention used in this study. Similar to formal inferential reasoning, Informal Inferential Reasoning allows students to make generalizations that go beyond sample data, too. However, contrary to formal inferential reasoning, Informal Inferential Reasoning helps students make inferences in an informal manner, such as comparing two box plots, graphs, or two samples of data sets without doing any formal calculations. Thus, whereas formal inferential reasoning utilizes procedural knowledge, Informal Inferential Reasoning does not necessarily.

An example of an Informal Inferential Reasoning task may illustrate the essential components of informal inferential reasoning. Suppose that grade nine students are asked to collect heights of their own class of 26 students. Afterwards, students investigate the tallest, shortest, and the most common height measure for their own class. Then, based on their own class's data, students are asked to predict the tallest, shortest, and the most common height measure for the grade nine class next door. In this phase, the students' estimation might be identical to their own class's data because they might think that other students have the same height. In particular, they might provide the same values for their estimates. After being shown the data pertaining to the class next door, students may begin to describe the differences between the two data sets and then compare these data sets of the students' heights. Further, going through comparisons between the data for the two classes, students might notice variation in the data sets. As a result, their estimation of the three-other grade nine classes in the school might be different than that of the estimation about the class next door. The students might provide a range of values for each estimate instead of mentioning only one value for each (i.e., the tallest measure could be between 1.53 m and 1.60 m).

As seen in the example, students may describe the measures of the heights of their own class and make inferences based on the data of their own class. This relates to the making of inferences that go beyond the sample data. Afterward, when the students compare their own data with the data of the other class, they may begin to notice that the values are not identical for the two classes and thus provide a range of values for estimating values for a third class instead of focusing on individual data points for their estimations. This relates to the recognition of variation and incorporation of some degree of uncertainty into conclusions. Afterward, being able to make inferences about other grade nine classes and the school in general based on the data of two classes relates to the use of sample data as evidence to articulate inferences. This Informal Inferential Reasoning task might help students draw conclusions about an unknown population (in this case, a school) based on the data at hand without using any formal procedures.

Indeed, Informal Inferential Reasoning is a way of utilizing inferential reasoning informally. It can be argued that introducing students to Informal Inferential Reasoning tasks before formal procedures of statistics can help students build up the conceptual understanding behind the procedural knowledge. Moreover, Informal Inferential Reasoning might allow students to think informally about the logic underlying formal inferential reasoning procedures, such as ANOVA and hypothesis testing. Several studies (e.g., Pfannkuch, 2006; Makar & Rubin, 2009; Trumpower, 2013) indicate that Informal Inferential Reasoning has the potential to increase conceptual understanding related to inferential reasoning. To make this argument more concrete, an example of an experimental Informal Inferential Reasoning study will be used to demonstrate how researchers used Informal Inferential Reasoning to help students enhance their conceptual understanding and thus develop their inferential reasoning skills.

In such a representative study, Makar and Rubin (2009) studied the development of Informal Inferential Reasoning at primary schools with four different teachers. Essentially, different teachers presented varying Informal Inferential Reasoning tasks to develop the different components of Informal Inferential Reasoning identified by the authors (i.e., use of probabilistic language and making of generalizations). Each teacher taught an inquiry-based unit on statistics to introduce students to Informal Inferential Reasoning. One of the teachers, for example, prepared an inquiry unit on healthy eating to examine how students used informal ideas to make generalizations. Students were asked to work in groups of 3-4 students to answer the question 'Do students in grade 6 eat healthier than students in grade 2?' Then, students collected their own data in the school through a survey and analyzed their findings; the students then presented their conclusions to the class.

After collecting the data, when students were working in their groups, the teacher prompted the students with questions, like 'What kinds of conclusions can you make regarding eating habits of grade 6 students?' and 'What if you had collected extra data regarding female and male students? What kinds of conclusions would you be able to make?' Such questioning techniques help students link their data to the research question. According to the authors, classroom discussion was therefore critical in moving students beyond data to make generalizations such as 'Boys ate more junk food in the upper grades' and/or 'there were more girls than boys who had a healthy lunch'. After getting inferences about grade 6 students from students, the teacher then asked additional questions, like 'Where did you get that information from?' to help the students use the sample data at hand as evidence for their conclusions. According to the researchers, use of Informal Inferential Reasoning as an investigative process in classrooms led most of the students to make generalizations that went beyond the data at hand and to use data as evidence for the basis of their arguments. Informal Inferential Reasoning thus increased the conceptual understanding of the participants.

As mentioned earlier in this section, moving beyond the data and use of data as evidence for arguments without doing any formal calculations are two essential components of Informal Inferential Reasoning. When the experimental study was examined, it was shown that students were able to demonstrate these two important aspects of Informal Inferential Reasoning through the activity. In the study, for instance, students made inferences about eating habits of upper grade students when some of them said, ‘Boys eat more junk food in the upper grades’.

Indeed, there are other aspects of the study that need further examination. First, the teachers encouraged students to collect their own data and make conclusions based on the data at hand. This may have allowed students to be active members of the investigative process and thus, the Informal Inferential Reasoning activity may have developed a need for the students to go beyond the data and make inferences about grade 6 students. Second, the organized classroom discussions helped students both to connect the data with the research questions under investigation and to use the data as evidence for their conclusions. As mentioned in the study, some of the students had difficulty connecting the data with the research questions when they completed their data collection and were working in their groups. However, the teacher’s questions, like “What can you say about male and female students’ eating habits?” and “Where did you get that information from?” helped the students to connect the data with the conclusions and use the data as evidence for their conclusions. Overall, it is shown through the study that

Informal Inferential Reasoning tasks can help students develop their conceptual understanding related to inferential reasoning.

In general, Informal Inferential Reasoning has become a growing phenomenon recently and has the potential for future applications. However, it is not the intention of this study to claim that Informal Inferential Reasoning itself is entirely a solution for the challenges identified in the literature, unless some other factors that have arisen from empirical studies and literature reviews are considered to help in the process of increasing conceptual knowledge. As one of the prominent suggestions, it has recently been argued that formative use of Structural Assessment of Knowledge (i.e., assessing and providing feedback on students' structural knowledge) may be useful in increasing students' conceptual knowledge (Goldsmith, Johnson, & Acton, 1991; Sarwar, 2011; Sarwar & Trumpower, 2015). Further, the GAISE (2016) report also encourages statistics educators to use assessments as a means to improve and evaluate student learning in the field of statistics education. Thus, the utility of assessing students' knowledge organization and providing them with feedback for increasing their conceptual understanding related to ANOVA was considered.

Structural Approach to Knowledge

In this dissertation, I borrow the definition of the term "knowledge structures/networks" from Schvaneveldt, Durso, and Dearholt (1989) to refer to psychological models which represent knowledge as a structure involving nodes (concepts) and links (i.e., semantic relations) between nodes. As the term "concept" will be used frequently in this dissertation it is worthwhile to give a clear definition of the term. Novak (1998) defines "concept" as a perceived regularity in events or objects, or records of events or objects, designated by a label" (p. 21).

However, the structural approach to knowledge entails considering knowledge as cognitive representations consisting of organized interrelated constructs (Jonassen, Beissner, Yacci, 2013). In addition, the structural approach to knowledge was a result of perceiving these structures as "semantic networks". For example, while working on the semantic processing of knowledge, Quillian (1969) described knowledge structures as semantic networks which again involve concepts/constructs and relational links between them. This perspective allowed Quillian (1969) to develop the spreading-activation theory of human semantic processing that was implemented in computer simulations of memory search and comprehension. To Quillian, memory search begins with calling relevant properties of a concept and continues to bring less relevant information as the search continues. Concepts and relations between these concepts are the two critical elements which form the basis of semantic networks in human semantic knowledge structures and memory studies. Similarly, Vygotsky (1997) argued that learners develop integrated knowledge schemas by linking each concept to another one. Further, he mentioned that successful learning depends on the organization of previous schemas that learners already have because these previous schemas are used as building blocks for forthcoming learnings. In other words, students' existing schemas essentially affect their future learning. As a result, significant learning occurs if a strong connection is made between what is known and what is to be known (Ausubel, Novak, & Hanesian, 1978). Supporting this argument, Ausubel (1968) argues that "the most important single factor influencing learning is what the learner already knows" (Ausubel, 1968, p. VI). Focusing on the construction of the concept of related schemas, Eggen and Kauchak (2007) indicate that learners construct these schemas by organizing meaningful relations between concepts.

Along the same line of reasoning, Goldsmith, Johnson, and Acton (1991) note that to be knowledgeable in a domain requires an understanding of interrelations between the critical concepts in the domain. Therefore, identifying the knowledge related to critical concept relations in a domain (i.e., knowledge networks or knowledge structures) is important because they represent learners' existing knowledge in a domain. In parallel, the U.S. National Research Council (2001) suggested assessing knowledge structure of a domain as a crucial predictor of meaningful learning.

Assessment of Knowledge Structures and Pathfinder Software

As outlined by Piaget (1977), knowledge structures are not directly observable but stay at the center of cognitive actions. Adopting this argument, various methods have been utilized to assess knowledge structures of learners. This procedure is often called Structural Assessment of Knowledge and it has been utilized both to externalize and assess knowledge structures in various domains. According to Trumpower and Sarwar (2015), Structural Assessment of Knowledge simply refers to a general process of assessing students' knowledge organizations. Direct measure of Structural Assessment of Knowledge such as the Concept mapping technique and indirect measure of Structural Assessment of Knowledge such as the Pathfinder Network (PFnet) Algorithm are the two most common methods suggested to facilitate the Structural Assessment of Knowledge process.

Goldsmith et al., (1991) proposed that the Structural Assessment of Knowledge process to visualize and evaluate knowledge structures involves three main steps, starting with (1) elicitation of student's understanding of concept relations in a domain, followed by (2) conversion of this conceptual understanding into a graphical representation (i.e., knowledge map), and concluding with (3) comparison of this knowledge map with a referent knowledge map derived from domain experts.

In addition to these three steps proposed by Goldsmith et al., (1991), an additional step derived from the formative use of Structural Assessment of Knowledge that was used in this dissertation is to (4) provide students with feedback upon elicitation and assessment of their knowledge structures based on the Structural Assessment of Knowledge principles described in the first three steps. The fourth step will benefit from the principles that govern formative assessment to utilize formative use of the Structural Assessment of Knowledge process. Thus, the first three steps within the Structural Assessment of Knowledge process will be used to create a basis for formative feedback as a last step. A detailed description of these four steps, incorporating some statistical concepts related to ANOVA, are presented below.

1. *Elicitation of knowledge structures*. As a first step of Structural Assessment of Knowledge, elicitation of students' understanding of concept relations can be conducted in two main ways— directly or indirectly. First, students can generate their own concept maps directly by sorting pre-determined critical concepts with concept mapping (Novak, 1990). According to Novak and Cañas (2008), "Concept maps are tools for organizing and representing knowledge. They include concepts, usually enclosed in circles or boxes of some type, and relationships between concepts or propositions, indicated by a connecting line between two concepts" (p. 1). Then, these links between the nodes are labeled depending on the nature of relation between the concepts. In general, concept maps provide information about why students perceive two concepts to be related. However, it has been noted that as students create concept maps, they may tend to create maps that are visually appealing (e.g., symmetrical or hierarchical) rather than accurately representing the complexity of conceptual relations (Trumpower, Sharara, &

Goldsmith, 2010). Also, the quality of a concept map depends not only on students' domain knowledge but also on their skills with constructing concept maps (Trumpower & Sarwar, 2010). Further, there are some concerns about scoring concept maps drawn by students. Although, there are different scoring methods such as scoring based on overall perception of concepts visualized by the map, the number of nodes and links, or based on the correct links (McClure, Sonek, & Suen, 1999), there are still concerns noted related to the time required to utilize these methods and the reliability of these scoring methods (Ruiz-Primo & Shavelson, 1996). It can be argued that shortfalls of concept mapping as a method of elicitation of knowledge networks can be addressed by indirect measures of Structural Assessment of Knowledge such as the PF algorithm as described below. Schvaneveldt and Durso (1981) proposed the PF algorithm to generate network models to represent knowledge structures based on proximity data.

As an alternative method, PF is an indirect measure of knowledge organization and starts with rating the degree of relatedness between concepts on a Likert type scale. To illustrate, Figure 1 presents a sample task on how students rate the degree of relatedness of pairs of concepts from the domain of experimental design (e.g., independent variable, dependent variable, random factors, and variation) on a 5-point scale with 1 being least related and 5 being most related. Subsequently, the PF scaling algorithm generates a network representation of the structure of the relatedness ratings submitted for analysis (Schvaneveldt, 1990). In other words, the PF scaling algorithm converts a matrix of relatedness judgments submitted by raters into a network structure, also known as a knowledge structure, or PFnet, which is very similar to a concept map. The section below will describe how ratings are converted through the PF scaling algorithm into a PFnet.

		less related ←				→ more related
Concept Pairs		1	2	3	4	5
Independent variable	Dependent variable					
Random factors	Independent variable					
Dependent variable	Random factors					
Variation	Random factors					
Dependent variable	Variation					
Variation	Independent variable					

Figure 1. Relatedness rating task for experimental design concepts

2. Representation of knowledge organization. As a second step in the Structural Assessment of Knowledge process, proximity data are converted through a scaling algorithm into a PFnet. Schvaneveldt, Durso, and Dearholt (1989) contextualize the term *proximity* as measurement units such as similarity or relatedness which indicate the extent to which entities belong together psychologically. The proximity in PFnets are defined by the two important parameters, *r* and *q*. Thus, the networks created based on these two parameters are denoted as *PFnet* (*r*, *q*). The *r* parameter determines the distance between two nodes/concepts based on the semantic distance between each pair of concepts, measured by relatedness ratings between concept pairs in the example above. In particular, the *r* parameter is used to identify indirect path distances between two concepts, and its value ranges from 1 to infinity (Cooke 1992; Schvaneveldt, 1990). The *q* parameter determines the maximum number of links in paths based on the *triangle inequalities* principle. The triangle inequality principle removes redundant links in a PFnet by requiring that the sum of the distances of two sides of a triangle must be greater than or equal to the third side. Specifically, two concepts are considered to be linked in a PFnet if the length of any indirect path between these two concepts is equal to or longer than the length of the direct link between these two concepts (Nash & Nash, 2003). The following example helps illustrate the triangle inequalities principle in PFnets. Assume that the distance[1] between statistical concepts, *independent variable* and *random factors*, is denoted by *d (independent variable, random factors),* and the triangle inequality requires that a third concept, say *variation* be restricted as follows*:*

d (*independent variable, variation*) ≤ *d* (*independent variable, random factors*) + *d (random factors*, *variation)*

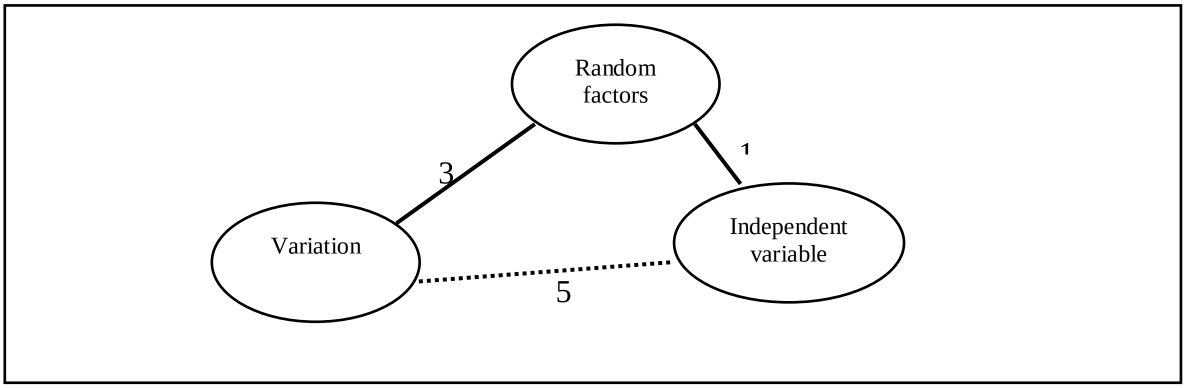


Figure 2. PFnet with three concepts based on triangle inequality

As seen in Figure 2, the dotted line between independent variable and variation does not meet the criteria for triangle inequality, because the direct path (distance = 5) is longer than the indirect path (distance = 3 + 1 = 4), thus this direct link should be removed from the PFnet. The complexity of a PFnet depends on the variations of previously indicated two parameters, *r* and *q*.

The PF scaling algorithm is also useful in creating PFnets with more than three concepts. In this case, the PF scaling algorithm moves beyond examining the length of indirect paths for only three links (i.e., traditional triangle inequality) to a maximum number of *n* – 1 links, where *n* is the total number of nodes (concepts) in a network. The PFnet takes its simplest form when

---

[1] The distance between concepts is typically determined by reverse-scoring the relatedness/similarity rating for a concept pair. For example, using the 5-point rating scale depicted in figure 1, suppose that the concepts *independent variable* and *dependent variable* are rated as a *4*. The distance between these two concepts would then be defined as: $distance = (total\ number\ of\ scale\ values + 1) - rating\ of\ relatedness$

= (5+1)-4

=6-4

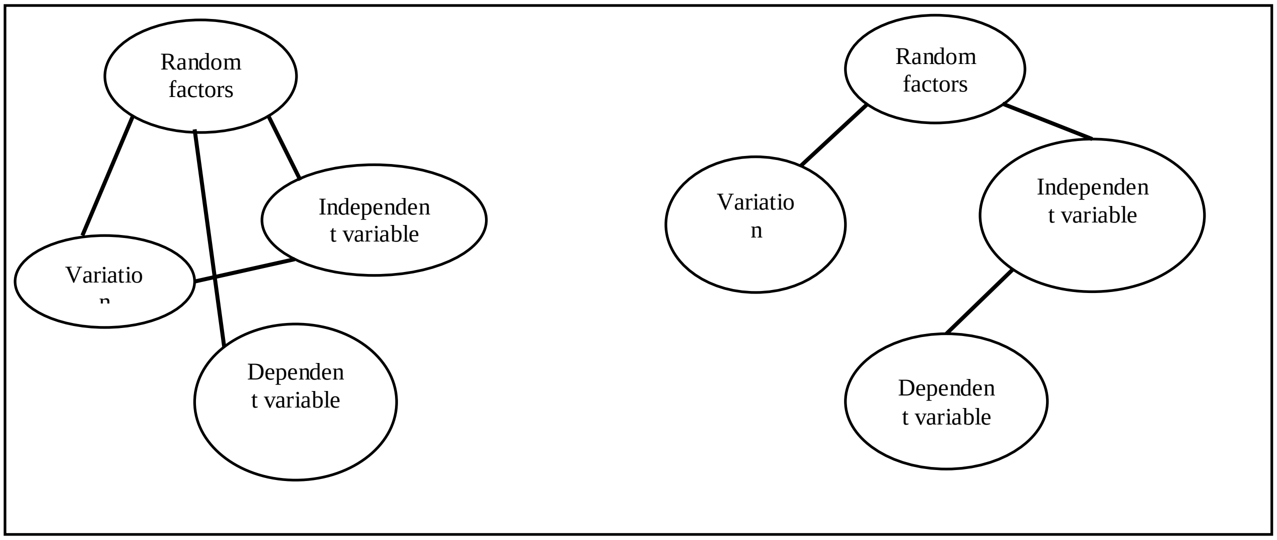
=2.

the parameters r and q have their maximum values ∞ and $n - 1$ respectively. Under these parameters, interpretation of the knowledge structure is easier (Cooke 1992).

As stated by Azzarello (2007) the final PFnet provides information about relatively how strongly students perceive pairs of concepts to be related (Azzarello, 2007). Unlike in the concept mapping technique, links between concepts are not labeled in the PFnet method. However, with assessment using PFnets, knowledge structures of large numbers of students can be evaluated in a relatively short amount of time (automatically via computer) compared to the concept mapping technique (which is typically evaluated by a teacher). Remember that whatever technique is used, either concept mapping or PFnets, the result is a graphical representation of one's structural knowledge (i.e., student map) in a domain. The resulting student map involves a cluster of concepts. These concepts are usually illustrated in nodes and a line appears between two nodes if a strong relation is indicated. Subsequently, the resulting student map is assessed in comparison with a referent knowledge map obtained from experts in the field (see, e.g., Trumpower & Sarwar, 2010). Altogether, the Structural Assessment of Knowledge process along with PFnets described above has been shown to be a valid measure of higher order knowledge in different subject matters (e.g., Boldt, 2001; Goldsmith et al., 1991; Trumpower & Sarwar, 2010; Sarwar & Trumpower, 2015). A more detailed description of the validity of Structural Assessment of Knowledge measures will follow.

3. Evaluation of knowledge structures. As a final step in Structural Assessment of Knowledge, students' individual knowledge structures are usually assessed based on the comparison between students' knowledge organizations and a referent knowledge organization (i.e., knowledge organization created based on one or more domain experts' relatedness ratings for pre-determined concepts). The notion behind this approach relies on two assumptions: (a) expert and novice individuals organize their knowledge differently which is a key factor in differentiating their performance in a certain domain, and (b) there is some model knowledge organization that best reflects the related knowledge in a domain (Acton, Johnson, & Goldsmith, 1994).

After generating students' knowledge structures using the PF scaling algorithm as described, the next step is to generate a referent network as a criterion. This is usually accomplished by averaging ratings of relatedness of two or more domain experts and then submitting the averaged ratings to the PF scaling algorithm to generate a single referent network (more will be said on this process below). The comparison of student and referent networks are mostly based on the s*imilarity* analysis between the networks. The similarity index (PFsim) is the number of links in common divided by the total number of unique links that are in both the student and referent networks combined (i.e., PFsim = common links / (total links − common links)). The PFsim index ranges from 0 to 1 where '0' indicates two completely unrelated structures and '1' indicates two identical structures (Goldsmith & Davenport, 1990). For example, suppose that Figure 3 presents a student and a referent knowledge structure obtained both from a student's and an expert's relatedness ratings for the same set of concepts using the PF scaling algorithm method as described in the previous section on how to elicit knowledge structures.



2.1.a Student knowledge structure **2.1.b** Referent knowledge structure

Figure 3. Hypothetical pathfinder networks

As seen in Figure 3, there are two common links between the two knowledge structures (i.e., the link between “random factors” and “variation” and the link between “random factors” and independent variable”). Further, compared to the referent knowledge structure, the student knowledge structure has two extra links (i.e., the link between “random factors” and “dependent variable” and the link between “variation” and “independent variable”) and one missing link between “independent variable” and “dependent variable”. In this case, PFsim score for the student knowledge structure would be equal to 0.4 (i.e., common links / (total links − common links) = 2/(7-2) = 2/5 = 0.4). This PFsim score would indicate a moderate “similarity” between the student and referent knowledge structures.

As illustrated above, knowledge structures are compared to referent networks to identify misconceptions or incomplete domain knowledge (Azzarello, 2007; Goldsmith & Johnson, 1990). The literature on how to create effective referent knowledge structures suggests using more than one expert’s relatedness ratings to create a referent network because this method can help to overcome personal biases through aggregation of relatedness ratings for the same set of concepts (Sarwar, 2011). Creating referent networks based on more than one domain experts' ratings on the same pairs of concepts produces better results in assessing students' knowledge organizations as this method reduces the possibility of having various idiosyncrasies of individual experts (e.g., Goldsmith & Johnson, 1990; Taricani & Clariana, 2006). Further, as highlighted by Acton, et al. (2014), averaged referent networks were found highly predictive, consistent, and valid. Therefore, given the evidence existing in the literature, the current study embraced this referent-based approach to assessing structural knowledge of participants related to conceptual understanding of a commonly used statistical procedure.

Structural Assessment of Knowledge and Validity

Hypothetically speaking, if Structural Assessment of Knowledge is to be considered a valid measure of knowledge structures, one would expect PFsim scores obtained through the comparison of referent and student networks to be related to some other valid conventional measures of knowledge. Empirical studies demonstrate that PFsim scores have repeatedly been shown to be a valid measure of structural knowledge (Azzarello, 2007; d'Appolonia, Charles, & Boyd, 2004; Goldsmith, Johnson, & Acton, 1991). For example, Goldsmith et al., (1991) reported that PFnet similarity scores of undergraduate and graduate students in statistics and computer programming courses increased significantly following instruction. In a similar vein, Kraiger, Ford and Salas (1993) observed differences in pre-and post PFnet similarity scores among Navy pilots receiving air crew coordination training. Likewise, Curtis and Davis (2003) also noted significant improvements in PFnet similarity scores among business students following instruction in a human resource management course. In another study by McGaghie, McCrimmon, Mitchell, Thompson, and Ravitch (2000), the authors investigated the similarity between students and instructor's knowledge maps using the PFsim scores generated by the PF scaling algorithm. They found that students' PFnets become increasingly similar to instructor's PFnet from pre- to post-intervention. Focusing on the validity and utility of structural assessment using Pathfinder networks, Davis, Curtis, and Tschetter (2003) examined the relation between PFsim scores obtained through the PF scaling algorithm and two other traditional measures such as final exam and in-class cumulative points for an accounting course. The correlations between the scores from the two conventional outcome measures and the PFsim scores of structural knowledge were significant. Further, a line of support for the validity of Structural Assessment of Knowledge measures can be found in Trumpower and Sarwar's (2010) study where they used 11 physics related concepts to measure students' conceptual understanding before and after an intervention. Participants' PFsim scores increased significantly from pretest (i.e., 0.45) to posttest (i.e., 0.56) after a remedial activity. Similarly, Casas-García and Luengo-González (2013) focused on the comparison of students' knowledge structures on the topic of "angles" from various school levels (i.e., primary, secondary, olympiad, and undergraduate levels). The authors found that the similarity between the knowledge structures increased with age and experience of the students.

In sum, empirical studies that focus on the validation of Structural Assessment of Knowledge measures followed two different methodologies in providing supporting evidence of validity. The first group of these studies reported an improvement in the similarity of a pathfinder network to a referent network after instruction while the second group of studies evidenced that the similarity between students' and expert's knowledge structures is positively related to other conventional measures of learning outcomes, such as course grades or final examination scores. As evidenced by empirical studies mentioned above, Structural Assessment of Knowledge not only provides a unique and potentially valuable method for assessing knowledge structures but also has the potential to measure higher order knowledge organizations in different domains accurately.

Formative use of Structural Assessment of Knowledge

Davis, Curtis, and Tschetter (2003) noted that “research that identifies or investigates remedial strategies that minimize the differences in the quality of knowledge revealed by structural assessment would be beneficial” (pg. 204). Recently, the Structural Assessment of Knowledge process has been proposed as a potential basis for such formative feedback to enhance student learning and to reduce differences between referent and student knowledge structures. Subsequently, formative use of Structural Assessment of Knowledge (i.e., assessing and providing feedback on students’ structural knowledge) has been shown to be critical not only for identifying students’ existing knowledge organization in a domain but also for improving students’ knowledge organization by providing feedback on students’ current learnings (e.g., Filiz, Trumpower, & Vanapalli, 2014; Goldsmith et al, 1991; Hay, Tan, & Whaites, 2010; Sarwar, 2011; Sarwar & Trumpower, 2015). According to Cohen (1985) feedback “is one of the more instructionally powerful and least understood features in instructional design” (p. 33). From a traditional point of view, Hattie and Timperley (2007) describe feedback as provided information related to anyone’s performance. Although this type of feedback provides students with an overall idea about their performance, the level of engagement with feedback is very limited as Trumpower and Sarwar (2010) highlight the importance of acting on received feedback to enhance learning in formative assessment process. Therefore, this type of feedback is called evaluative feedback and/or feedback *of* learning. Conversely, formative feedback guides students through the necessary actions in order to achieve the desired learning outcomes (Sadler, 1989). Thus, formative feedback enables learners to reinforce their strengths as well as to determine their weaknesses (Brown, 1997). However, in order to achieve higher level of cognitive skills, formative feedback should follow certain principles as suggested by Brookhart (2007), Carless (2006), Gibbs and Simson (2004), Hattie and Timperley (2007), and Nicol and Dick (2006). According to these researchers, the principles of formative feedback as follows;

- Timely: Learners have a chance to use the feedback and apply it for future learnings when the feedback is provided on a timely manner when learners can still recall what happened on a given task (Gibbs & Simpson, 2004).

- Motivational: Feedback on learners' work should encourage positive motivational beliefs and self-esteem. Negative impacts on motivation and self-esteem are unintended results of high-stakes evaluative feedback where usually only marks or grades are given. Therefore, using low-stakes formative assessment tools within the learning process and giving marks on the final version of the work can encourage positive motivational beliefs and self-esteem (Gibbs & Simpson, 2004; Nicol & Dick, 2006).

- Individual/personal: Individualized feedback can address each learner's unique strengths and weaknesses better and help them to learn at their own pace.

- Manageable: Quality formative feedback delivers relevant and useful information and allows students to act on it. Brookhart (2007) mentions that feedback needs to be specific enough that a student can understand what to do next, but not to be too specific that teacher does the work.

Given the principals of effective formative feedback, the multi-media feedback provided to students in this study conformed to the principals of formative feedback described. First, the feedback in this study was provided on a timely manner right after the assessment. Second, it helped learners to identify areas of misunderstanding related to ANOVA and provided students with a chance to use the feedback and apply it for future learnings where students were required to engage in visual, written, and multi-media feedback instead of simply skipping the feedback. Third, the feedback was tailored to each student's weaknesses and strengths as it was generated as a result of comparison between each student's concept map and a referent map. As a last note, the feedback in this study went beyond evaluative statements and presented descriptions of possible areas where students might have missed crucial information related to ANOVA. A detailed discussion of how the feedback was provided and used in the present study will be presented in the procedure sub-section of the methodology section.

To explore the effectiveness of feedback generated by Structural Assessment of Knowledge, Trumpower and Sarwar (2010) provided physics students with feedback to enhance their understanding of any missing concept relations after each student's misconceptions were identified using PFnets. Students were first asked to rate the degree of relation between concept pairs and these ratings were converted into knowledge structure/concept map-like networks. Afterward, students were shown both their knowledge structure and a referent knowledge structure, and then asked to reflect on similarities and differences. Additionally, students were required to solve problems and review examples which were specifically focused on the missing concept relations in their knowledge maps. Each student's knowledge structure was evaluated by the researchers in order to determine the individualized feedback (i.e., problems to solve and examples to review) corresponding with that student's missing links and misconceptions. The individualized feedback was found to be effective in increasing students' understanding related to missing concept relations. Likewise, Sarwar (2011) used PFnets to formatively assess and to create individualized interventions for high school physics students. PFsim scores were obtained through the comparison between students' and referent knowledge structures. The interventions involved giving written feedback to the students on the differences between theirs and the referent networks. Following the intervention, the study found a significant improvement in the students' knowledge structures by comparing pre- and post-intervention similarity indices.

Drawing on the results from the studies described above, it can be argued that providing formative feedback that focuses on the relation between core concepts related to ANOVA can increase students' conceptual understanding. In this dissertation, formative use of Structural Assessment of Knowledge was used to provide detailed feedback on students' Informal Inferential Reasoning ability. The feedback provided in this research incorporated Informal Inferential Reasoning activities which were previously described as holding potential for increasing students' conceptual understanding of statistics. Thus, this dissertation not only focused on partially filling the gap in understanding the issues of students' difficulty with ANOVA by using an intuitive ANOVA task but also helped to understand the link between increasing Informal Inferential Reasoning ability and conceptual understanding of ANOVA as measured by Structural Assessment of Knowledge. While the iANOVA task helped to identify difficulties with Informal Inferential Reasoning and to form the context for the feedback, the Structural Assessment of Knowledge was used to assess conceptual understanding and to generate individualized feedback related to logic of ANOVA.

Overview of the Present Study

Beside an attempt to better understand students' difficulty with the inferential reasoning related to ANOVA, I also examined a strategy for increasing students' conceptual understanding of ANOVA. Although recent research that used feedback as a pathway to develop students' understanding related to various topics showed that it was effective, the required amount of time and work to evaluate student's knowledge structures and to provide feedback based on the evaluation were intensive (see Sarwar, 2011). Therefore, it may not always be possible to evaluate learners' work in a short amount of time and provide feedback for a large number of students in a timely manner which might negatively impact the effectiveness of the process. To overcome the limitations related to time and workload, a web-based application called the Concept Maps for Learning (CMfL) tool was used. The CMfL tool was developed based on the PFnet generation algorithm and has been used previously in other research to demonstrate that it follows the principals of effective formative assessment and contributes to assessment for learning practices (e.g., Filiz, Trumpower & Atas, 2012; Filiz, Trumpower & Atas, 2013).

The aims of this study were, first, to map out undergraduate students' current knowledge of ANOVA which is a commonly taught formal statistical test in tertiary statistics courses when introducing students to hypothesis testing, and second, to provide individualized visual, written and multi-media feedback to increase conceptual understanding of ANOVA via the CMfL website. Thus, in return, an increase in conceptual understanding of ANOVA may help students to demonstrate better Informal Inferential Reasoning. iANOVA tests were used to measure students' Informal Inferential Reasoning related to the normative logic of ANOVA. Immediate individualized feedback and associated instructional material (e.g., visual, written and multi-media feedback) were then provided to increase conceptual knowledge related to ANOVA based on the comparison of their knowledge map and an averaged referent structure. Through the CMfL website, students were able to create knowledge structures on previously determined critical concepts of ANOVA. In this sense, the website not only incorporated all Structural Assessment of Knowledge steps described before but also provided learners with different types of immediate feedback as a formative addition to the Structural Assessment of Knowledge process. As suggested by the relevant literature, the feedback given to participants utilized informal inferential reasoning activities to help participants focus on the rationale behind ANOVA and to avoid formal computations of ANOVA as consistent with the nature of these activities. The research questions below were addressed in the study:

1) What are some of the reasons behind students' difficulty in Informal Inferential Reasoning as measured by the iANOVA task?
2) Is there a relation between students' performance on the Informal Inferential Reasoning task and conceptual understanding related to ANOVA measured through Structural Assessment of Knowledge?
3) Can formative feedback that focuses on conceptual understanding of ANOVA improve students' Informal Inferential Reasoning as measured by the intuitive ANOVA tasks

CHAPTER 2

METHODOLOGY AND PROCEDURE

Research design

This study utilized a mixed-methods design that involves collecting, analyzing, and combining both quantitative and qualitative data in the research process to understand the research problem better (Creswell & Clark, 2017). In mixed methods research, researchers often adopt the pragmatist world view and utilize both quantitative and qualitative approaches (Creswell, 2013). An embedded mixed methods design was used in the present study. In an embedded mixed methods design, the researcher collects and analyzes both quantitative and qualitative data within either traditional quantitative or qualitative research; thus, an embedded design is based primarily on one type of data while the secondary data type is included for support and enhancement (Creswell & Clark, 2017). I embedded a qualitative strand into the quantitative data collection tool in order to be able to better answer the first research question (i.e., What are some reasons behind students' difficulty in Informal Inferential Reasoning as measured by the iANOVA task?) and to provide additional insight into any potential changes in informal reasoning on the iANOVA task following feedback (i.e., the third research question).

Participants

A power analysis with one between-group factor and two within-group factors for a split-plot factorial ANOVA revealed that 70 students in total is sufficient to detect 75% power to reject the null hypothesis of no difference between participants in the control and intervention group. Once the research proposal was approved by the committee members, I had to go through two different ethics approval process because the data collection site was a different university than where I actually conducted the study. First, an ethics approval was obtained from the

University of Ottawa. Then, an application for another ethics approval was made to a mid-size university in the United States (see Appendix B for the complete consent forms from both universities). After obtaining the approval from the Institutional Review Board (IRB) of the university in the United States, the research proposal was advertised in the online research pool of the Department of Psychology where all available approved research proposals along with the contact information of the primary researcher were posted to recruit volunteer participants from the Psychology classes. Then, interested potential participants contacted the researchers by email indicating their willingness to participate in the study.

Initially 67 undergraduate students from the data collection site contacted the researcher indicating their willingness to join the study. However, 13 students were excluded from the data analysis since these students left the study at various stages of the research and did not come back to complete the study. Those participants who completed the study received research credits for their participation in the study. Demographic data related to self-identified gender, age, and major were also collected. As shown in Figure 4 below, 41 female and 13 male participants completed the study. Randomization of self-identified gender differences into the control and intervention groups was also successful where there were seven male and 20 female participants in the intervention group whereas there were six male and 21 female participants in the control group.

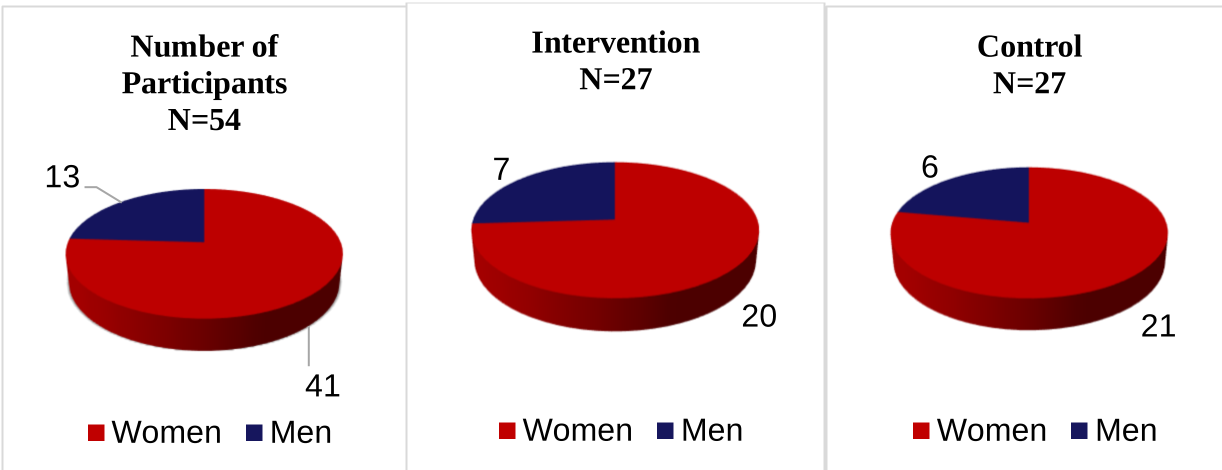


Figure 4. Distributions of participants into the control and intervention groups by self-declared gender

In addition, 46 of the total students were aged between 18 and 22 years old while two students were aged between 23 and 27 years, three students were between 28 and 32 years old, and three students were 38+ years old (see Figure 5).

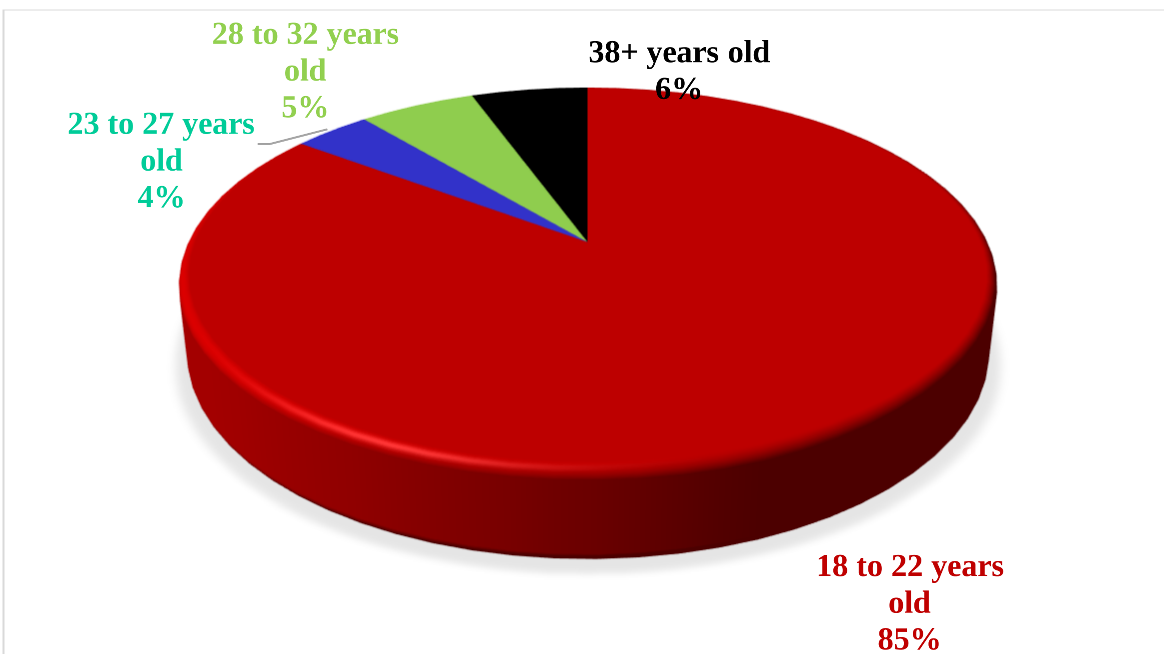


Figure 5. Participants by age groups

Participants who completed the study were enrolled in diverse programs at the university. Twenty-six of them were from Psychology, six of them from Medical Sciences, five of them were from Business, four from Engineering, three from English Literature, three from Criminology, two were from Arts Studio Design, two were from Business Management, two were from Teacher Education, and one was from Communications. Most of the students were in their first and second year of the program.

Participants also had a diverse background in statistics as measured by the number of statistics courses taken by the participants. Although 27 participants had taken one statistics course, ten participants had two courses, two participants had three courses, and 15 students have never taken a statistics course before the present study. Students' actual year of study in their program and background in statistics were not relevant to the inclusion criteria in the present study as previous studies highlighted that there are often little to no significant difference in conceptual understanding of ANOVA between those students who have a statistics background and those who do not, even between undergraduate and graduate students after first year statistics courses (Ben-Zvi & Garfield, 2005; Trumpower, 2013; 2015). Moreover, the information regarding participants' perceived level of anxiety towards statistics was also collected. Participants rated their level of anxiety on a scale of 1 to 5, with 1 being least and 5 being most anxiety. Six students reported their level of anxiety as 1, fourteen reported as 2, twenty-two students reported as 3, eight students reported as 4, and four participants reported as 5. A detailed discussion of any statistical differences by demographic background on key measures in the study such as perceived level of anxiety and statistics experiences will be presented in the beginning of the quantitative results.

Data collection instruments

Two different data collection tools were used in the study: an Informal Analysis of Variance (iANOVA) task and the Concept Map for Learning Website (CMfL). While the iANOVA task aimed to assess participants' ability to perform informal reasoning with respect to ANOVA, CMfL intended to measure participants' conceptual understanding of ANOVA as indicated by their knowledge structure for critical pre-determined concepts related ANOVA. Descriptions of both instruments are as follows:

iANOVA Task. The iANOVA task involves four components: a cover story, hypothetical data sets, a rating scale, and a "why" question aiming to understand students' reasoning strategies. The iANOVA task was used three times during the research: before intervention, right after intervention, and a week after the intervention. A different yet similar cover story was used every time to eliminate the effect of context on students' reasoning process. For example, while the first cover story described two scientists/entrepreneurs conducting an experiment to test their hypothesis that soccer balls made of 32 panels are better than soccer balls made of 14 panels, the second cover story described an experiment where a scientist compares two types of diet. Then, each cover story was followed by twelve hypothetical datasets (i.e., six smaller data sets and six larger data sets) showing scores obtained as a result of each experiment presented in the cover story. Although different data points were used with each iANOVA task, the size of data sets and the ratio of the between group variation to within group variation remained the same across iANOVA tasks (A detailed presentation of all three cover stories and hypothetical datasets are presented in Appendix A). After viewing each dataset, participants indicated the amount of their confidence on a 10-point scale from 1 to 10, 1 being weak and 10 being strong, in the claim that one type of soccer balls or diets is better than the other one based on the raw data presented.

Subsequently, participants were required to briefly answer a “why” question to provide a rationale for their rating for each experiment. The answers to each why question helped the researcher to better understand learners’ reasoning process and to identify students’ difficulty in different ANOVA scenarios. It has been proven by prior studies that the use of iANOVA task is a valid tool as a measure of understanding related to ANOVA (see; Trumpower, 2013, 2015).

Concept Maps for Learning (CMfL) Website. The CMfL website is a computer-based formative assessment tool that provides specific feedback on one’s knowledge structure by comparison to a referent knowledge structure. In this study, CMfL was used as a structural assessment of knowledge tool. The CMfL website was developed by a group of doctoral students, including the researcher of the present study, and their supervisor at the University of Ottawa’s Faculty of Education over a two-year period. The CMfL was used to measure participants’ structure of knowledge organization on ANOVA and to provide immediate feedback on their iANOVA performance before and after the intervention. The process for generating participants’ concept maps and for presenting different types of feedback are described below.

First, a group of experienced statistics educators created a list of critical concepts related to ANOVA. The concepts identified by the experienced educators on ANOVA are as follows: (1) *random factors*, (2) *independent variable*, (3) *dependent variable*, (4) *differences between groups* (5) *differences within groups*, (6) *sample size*, (7) *strength of evidence* that an independent variable has a real effect on a dependent variable. Then, all three subject matter experts rated the degree of relatedness of all pairwise combinations of these concepts on a 5-point scale. Subsequently, the rating for each concept pair from three experts were averaged to obtain an average rating for each concept pair rated. Then, the CMfL tool converted these average ratings through the Pathfinder scaling algorithm into a referent knowledge map. As a result, the referent knowledge map showed the link between related critical concepts based on the pairwise ratings of the subject matter experts. Figure 6 below shows the referent knowledge map for the concepts of ANOVA used in this study.

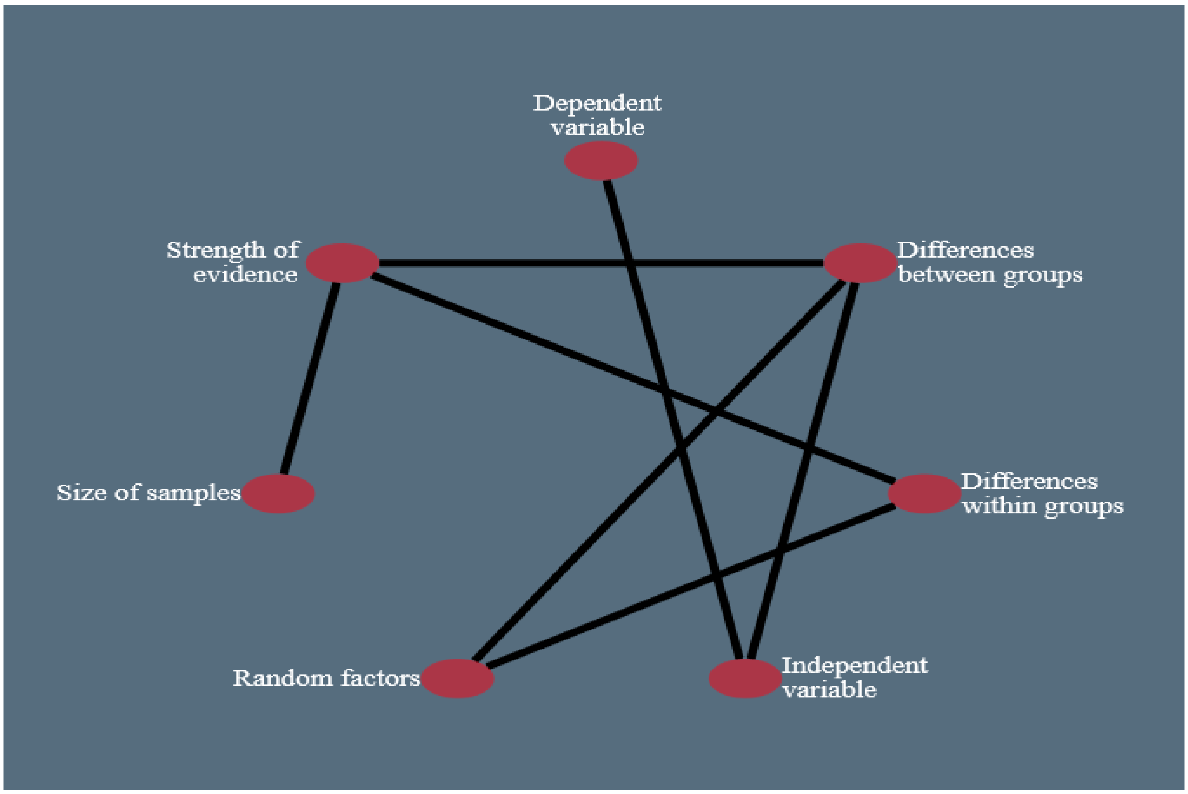


Figure 6. The referent knowledge map

As seen in Figure 6, there are seven links between the pre-determined concepts. Once the referent map was created, the researcher of this study prepared multimedia materials such as written and visual feedback to demonstrate how any linked concepts are related in the referent map. Later on, these multimedia materials were then uploaded into the CMfL website and linked to the referent knowledge map. Later, participants started using CMfL to create their own concept maps related to ANOVA and to receive individualized feedback. The description of how participants completed the CMfL task will be presented below.

Procedure

An experimental design with one intervention and one control group was used in the study. As seen in Figure 7, students were randomly assigned to either the intervention or control group. Also, the order of the tasks within each data collection tool was randomized for each participant to avoid any possible effect of order as a confounding variable. In other words, participants had randomized order of Size and Dataset conditions and randomized order of concepts pairs for ratings. Both data collection instruments were web-based and accessible through the CMfL website (www.conceptmapsforlearning.com). Further, the URL, along with a user name and password, were sent to those who were willing to participate in the study. Once the participants logged in using the user name and password, they first saw a consent form explaining the overall procedure posted on the web as an opening page (see Appendix B for the complete consent form). Then, they were requested to click on the button, saying “I agree to complete the study,” thus expressing their compliance to participate in the study. After expressing their agreement to participate, participants in the intervention and control group followed similar yet slightly different procedures.

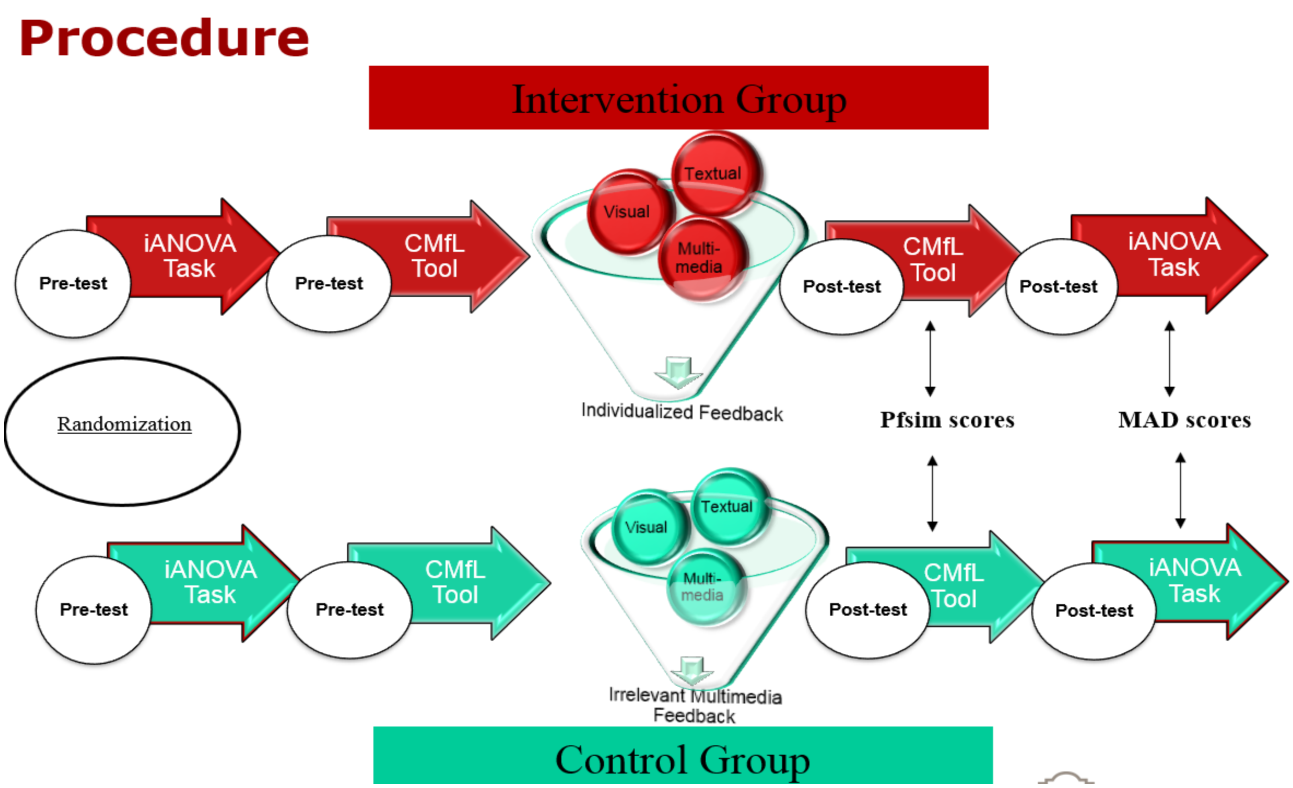


Figure 7: Flow chart for the experimental design

Procedure for intervention group participants. The procedure for the intervention group involved five phases. In the first phase, participants completed the iANOVA task available online as a pre-test (see Appendix A). Specifically, 12 hypothetical datasets were presented in the iANOVA tasks and participants were asked to rate their perceived strength of evidence on a 10-point scale for each dataset. Participants were also required to provide a justification for each rating. In the second phase of the process, participants rated the degree of relatedness of all concept pairs, as did the experts, using the CMfL tool. The participants had to complete twenty-two pairwise ratings in total. Each participant completed the relatedness rating task in random order. Figure 8 below demonstrates the pairwise rating task that each participant completed.

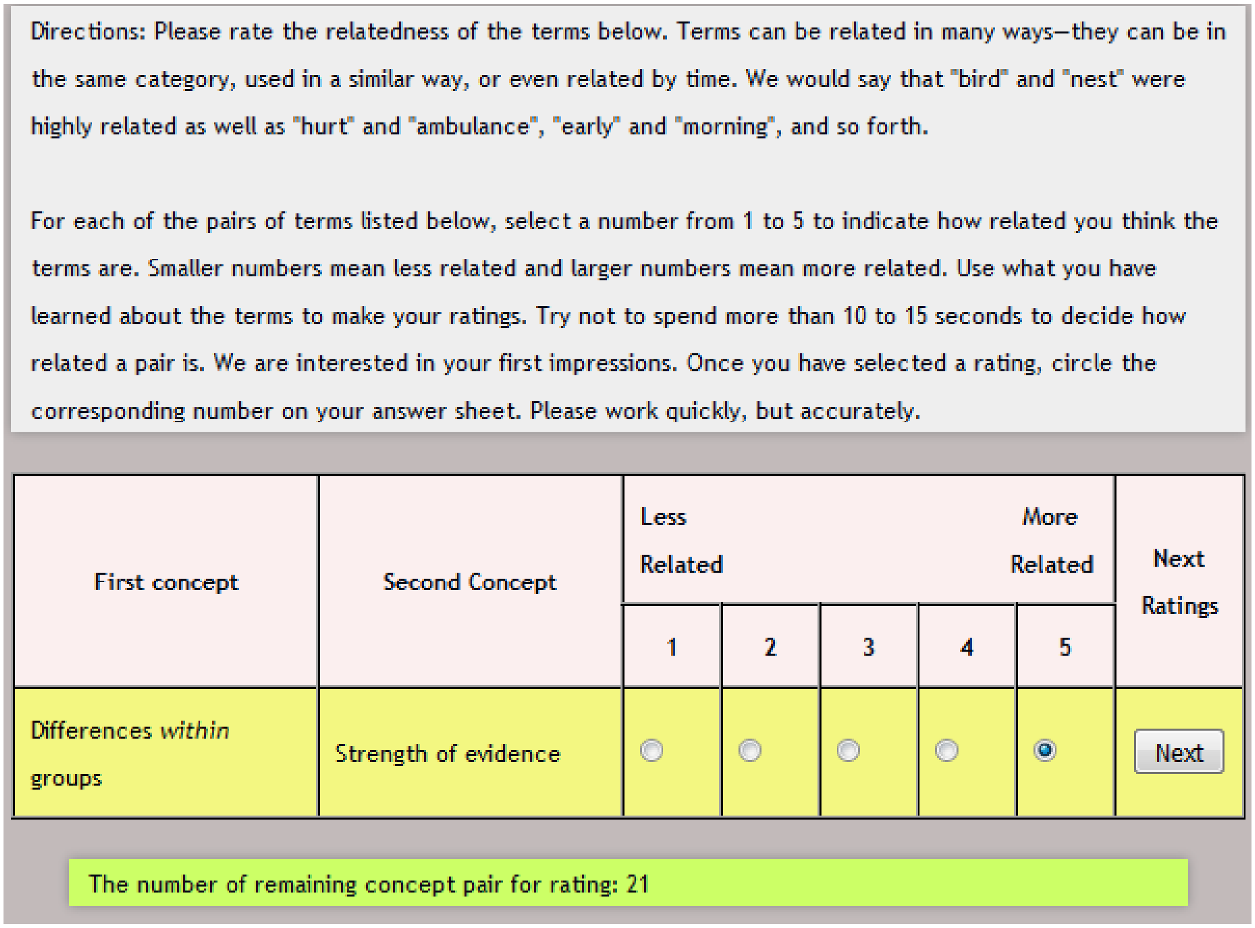


Figure 8. The pairwise rating task

As shown in Figure 8, the instructions on how to complete the rating task were provided on the top of the page followed by concept pairs shown on a 5-point scale. Also, the number of remaining concept pairs for rating was provided at the bottom. Again, the CMfL tool converted these ratings into a visual concept map (i.e., PFnet).

In the third phase, participants received immediate individual feedback (i.e., visual, textual, and linked feedback), based on their concept map through the website. As an initial stage in the feedback process, students received visual feedback on their concept map as a comparison between their own map and the referent map. As seen in Figure 9 below, visual feedback on the students' map appears in three forms: a black line, a grey dotted line, and a red dashed line.

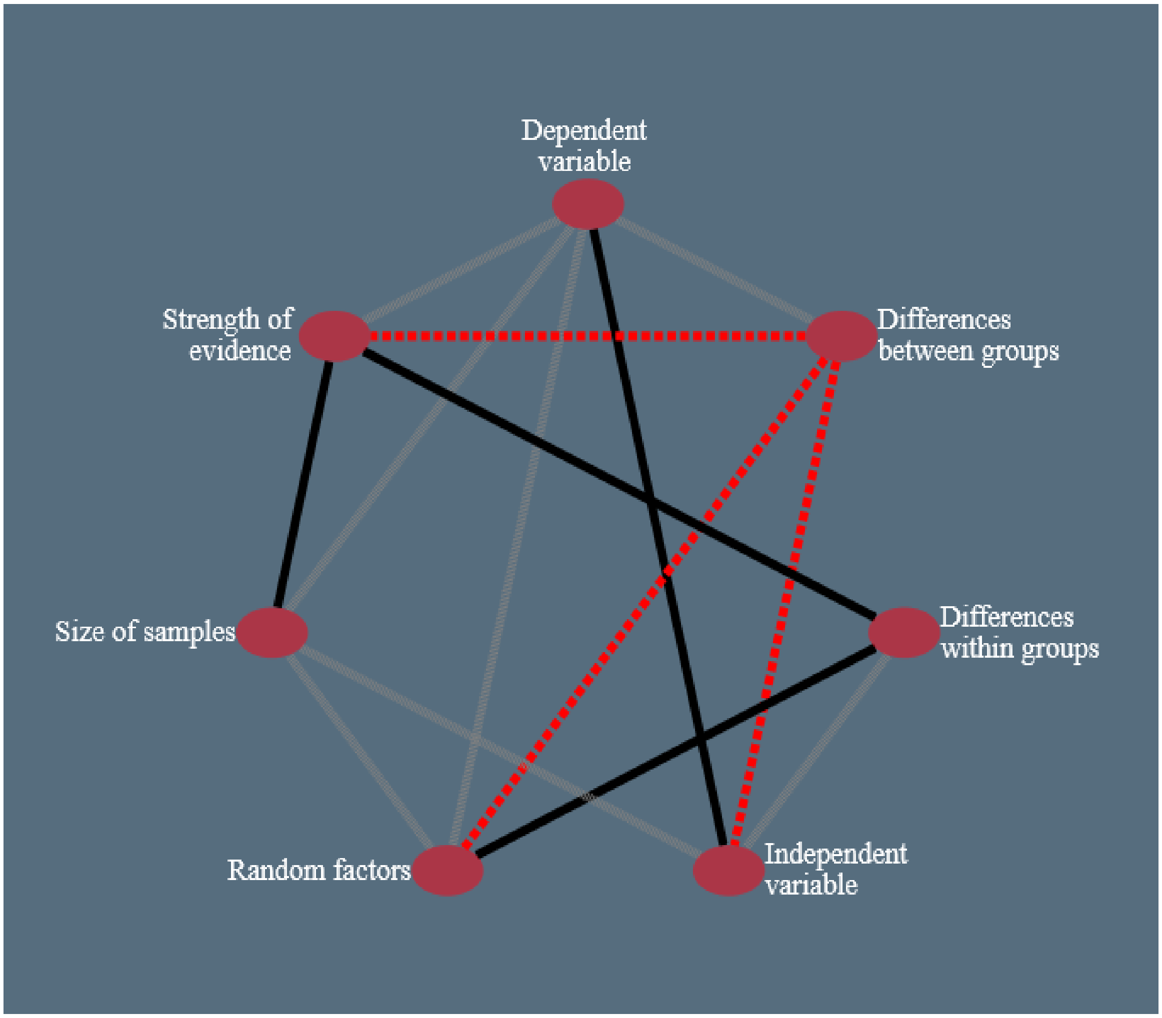


Figure 9. Visual feedback

A black line, called a "*relevant link*," refers to a link between two concepts in both the referent and the student maps (e.g., *random factors* and *differences within gr*oups). A grey line, called an "*extraneous link",* emerges if there is a link between two concepts in the student's map but not in the referent map (e.g., *independent variable* and *differences within groups*). Finally, a red dashed line, called a "*missing link",* appears if there is a link between two concepts in the referent knowledge map but there is no link between these concepts in the student map (e.g., *differences between groups* and *independent variable*). Participants of this study were shown a sample visual feedback before they start the task to highlight the significance of the black, red, and, grey lines on appears visual feedback.

Further to the visual feedback, students were also provided with additional instructional materials, which were linked to any missing links in their concept map. The linked feedback attempted to demonstrate the ways in which the associated concepts are related in more detail. When students clicked on a missing link on their map, they had access to two forms of linked instructional material: textual and multimedia. Both textual and multimedia feedback provided in the present study incorporated a soccer ball example used in the iANOVA task to highlight the conceptual understanding of ANOVA derived from an example that they were familiar with from the iANOVA task that they completed before the feedback phase. This type of feedback on students' missing links in their own concept maps was expected to lead to improvements in students' conceptual understanding of ANOVA, and subsequently to improvements in their Informal Inferential Reasoning performance. For example, if a student missed the relation between the concepts "random factors" and "differences within groups", the textual feedback would highlight the potential relationship between these concepts as illustrated below (see Appendix C for the complete set of textual feedback).

The Relation between *Random factors* and *Differences within groups*

To illustrate the relation between *random factors* and *differences within groups*, please recall the soccer ball experiment you have seen in the iANOVA task; now, suppose that the scores obtained by balls constructed from two different number of panels were as follows:

Scores of soccer balls by number of panels

Number of panels	
32-panel	**14-panel**
230	420
370	280
300	350

In this example, the balls made with 14 panels had 50-point higher average score than the balls made with 32 panels. At first glance, we could conclude that balls made with 14 panels are better balls with 32 panels. However, before concluding that 14-panel soccer balls are actually better than 32-panel ones, we need to recognize that not all of the balls kicked by the same force and made with the same number of panels obtained the same score. For instance, the three balls made with 32 panels differing by 70-point, on average, from one ball to another. This might have been due to factors such as gusts of wind and imperfections in the balls since all balls *within the same group* were made with the same number of panels. These factors are called "*random factors"* because they vary randomly across the experiment. Thus, it can be said that the three balls made with the 32 panels differed only due to *random factors* as did the three balls with the 14 panels. We could estimate the effect of *random factors* on scores obtained by balls by considering the average amount of deviation within the scores obtained by balls made with the same number of panel and kicked by the same amount of force. In this experiment, considering that 70-point average deviation from one ball to another, we could conclude that "*random factors*" are entirely responsible for the 70-point "*difference within groups*" because the difference cannot be due to different number of panels on a ball since all balls in a group were made with the *same* number of panels.

In addition to the written feedback, multimedia feedback explaining the nature of relation between linked concepts on the referent map was also prepared. Essentially, the multimedia feedback featured animated videos created by the researcher of this study using an online tool called "POWTOON". The length of video clips ranged between 1.30 minutes and 2.20 minutes. For instance, Figure 10 below demonstrates screenshots from an example of multimedia feedback prepared to emphasize the relation between "*independent variable"* and "*dependent variable"*.

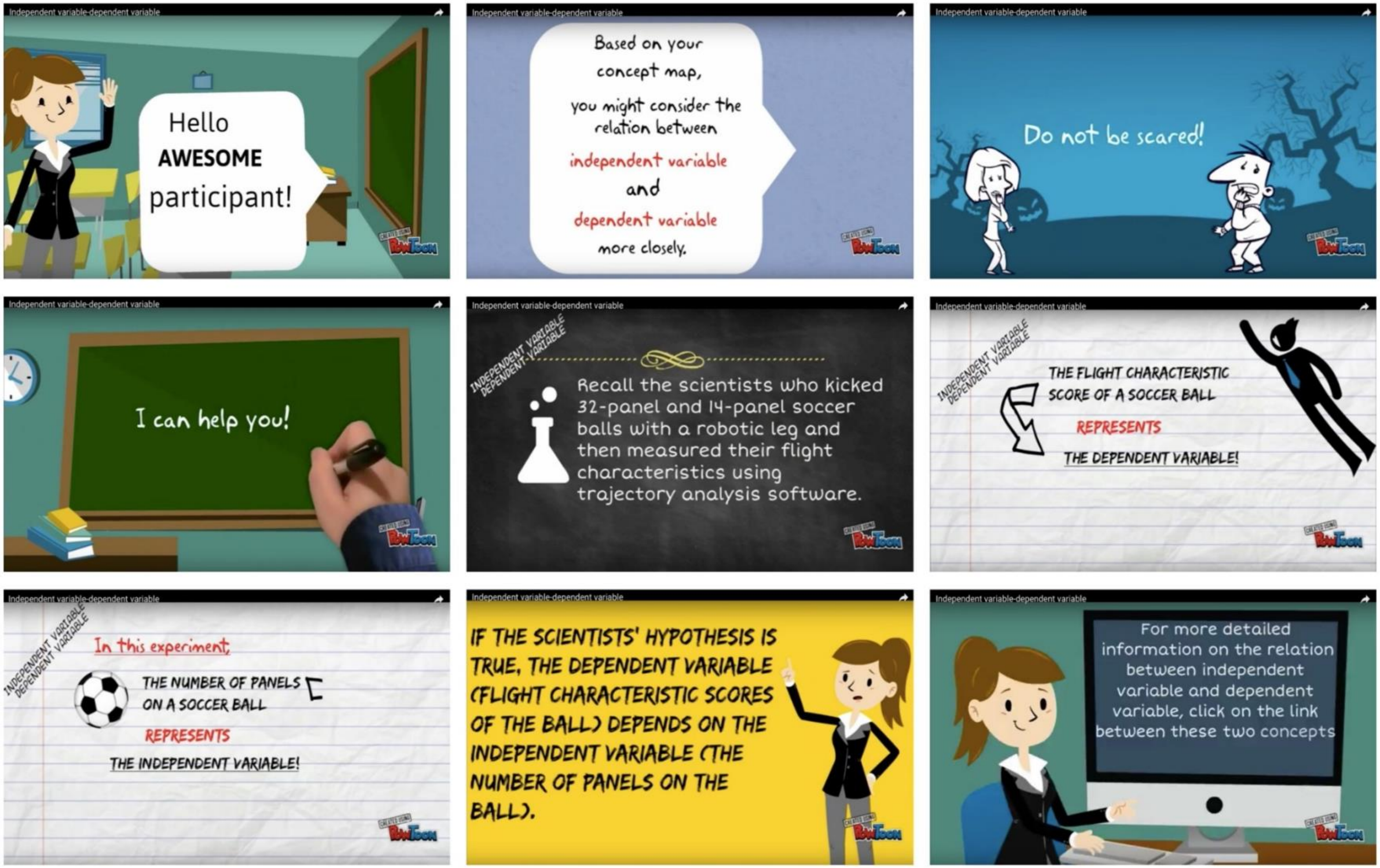


Figure 10. Screenshots from multimedia feedback

In the fourth phase of the experiment, participants completed the iANOVA task and concept relatedness rating task again right after the individualized feedback by following the identical procedures as in the pre-tests completed during the first two phases. One week after completing the fourth phase, participants were asked to complete the iANOVA and relatedness rating tasks once more in the fifth and final phase.

Procedure for control group participants. Participants in the control group completed the iANOVA and concept relatedness rating tasks as pre-, post-, and one-week follow-up tests, just as did the participants in the intervention group. However, participants in the control group only saw their own concept map and received irrelevant multimedia feedback (i.e., a short video clip about superfood) as a filler activity during the third phase of the study. However, participants in the control group had the opportunity to receive the same feedback as the intervention group after they completed the study.

As a result of the overall procedure for both intervention and control group participants, the researcher of the present study was able to collect both quantitative and qualitative data. A detailed description of how quantitative and qualitative data from the study were analyzed are presented in each associated part in the results section below

CHAPTER 3

RESULTS AND DISCUSSION

QUANTITATIVE DATA ANALYSIS

Measuring Performance on the iANOVA Task. As indicated, 12 hypothetical datasets were presented in each of the iANOVA tasks and participants were asked to rate their perceived strength of evidence on a 10-point scale for each dataset. Each of these datasets rated by participants was denoted by their size and between- and within-group differences. For example, one dataset in the iANOVA task involved small samples of soccer balls and had a mean of the scores obtained by soccer balls with 32 panels of 305 with a standard deviation of 2 and a mean of the scores obtained by soccer balls with 14 panels of 300 with a standard deviation of 2. Thus, the dataset had a between groups difference in means of 305–300 = 5 and an average within-group standard deviation of (2+2)/2 = 2. Because the sample sizes were relatively small (i.e., comprised of 3 rather than 10 data points per group), it was referred to as Sb5w2, where "S" refers to small size, "b" refers to between-group difference, and "w" refers to within-group differences. Likewise, a dataset with the same amount of between- and within-group differences, but comprised of 10 data points per group, would be referred to as Lb5w2.

Participants' performances on the smaller and larger datasets were considered separately. Considering first the six smaller datasets, for example, if we were to conduct an actual ANOVA on those six smaller datasets to be rated, the resultant *F*-ratios (i.e., formal calculation of the ratio of between-group variation to within-group variation) and associated *p*-values would give us a normative indication of the effect of the independent variable, (i.e., number of panels), relative to the effect of random factors. Bigger ratios would provide stronger evidence to support the conclusion that one particular number of panels is better than the other. Conducting the

ANOVAs, the *F*-ratios of the six smaller datasets are rank ordered from smallest to largest as in Table 1. Thus, while data set Sb5w20 would provide the weakest evidence to support the hypothesis, the data set Sb25w2 would provide the strongest evidence that one type of ball is better than the other one.

Table 1

F-ratios and p-values of Small Hypothetical Datasets used in the iANOVA tasks

Size	**Data sets**	***F*-ratio**	***p*-values**
S	B5W20	.094	.775
S	B15W20	.844	.410
S	B25W20	2.34	.201
S	B5W2	9.375	.038
S	B15W2	84.375	.001
S	B25W2	234.375	.000

Each participant's performance for smaller datasets was quantified as the mean absolute difference (MAD) between their rankings of the six smaller datasets and the normative ranks determined by actual ANOVA results shown in Table 1. The following example will help to understand how the MAD score for each participant's performance on the smaller datasets was calculated. Suppose that Table 2 below shows the ratings for the six datasets by one participant, along with the ranks of the datasets as ordered based on the participant's ratings, and the normative ranking of datasets as displayed in Table 1. For example, as seen below, both datasets Sb15w2 and Sb25w2 were rated highest by the participant and thus were tied for the 1st and 2nd ranks, so each got a rank of (1+2)/2= 1.5. Next, the dataset Sb5w20 was rated as the third and the dataset Sb25w20 was ranked as the fourth highest, so these data sets got a rank of 3 and 4. Then, the next two datasets receiving ratings of 6 were tied for the 5th and 6th ranks, so these data sets would receive a rank of (5+6)/2=5.5. Next, the differences between these ranks and the normative ranks of the datasets were computed and summed, resulting in a sum of absolute differences for this particular participant of 7. Dividing this sum by the total number of smaller datasets generates the MAD score of 7/6=1.17. Following this procedure, MAD scores for smaller and larger datasets were calculated for all participants both in the control and intervention group to determine their performance on the pre- and post- iANOVA tests.

Table 2

Example to Illustrate Calculation of the Mean Absolute Difference between Normative Rankings and a Participant's Rankings for the Six Small Datasets

	Normative rankings	Ratings by participant	Participant rankings	Absolute difference between normative and converted ranks
Sb5w20	6	6	3	3
Sb15w20	5	4	5.5	0.5
Sb25w20	4	5	4	0
Sb5w2	3	4	5.5	2.5
Sb15w2	2	7	1.5	0.5
Sb25w2	1	7	1.5	0.5
				Sum of absolute differences =7

Later, the MAD scores were examined to determine if students' reasoning about the strength of evidence provided by the hypothetical datasets was consistent with the normative logic of ANOVA, and if their reasoning became more consistent with normative logic after the intervention. In other words, MAD scores reflect the amount of deviation from the normative performance. Therefore, smaller differences (i.e., smaller MAD scores) between participants' rankings on the iANOVA task and normative rankings of datasets indicates better performance.

Measuring the Performance on the CMfL Task. Students' ratings of the degree of relatedness of critical ANOVA concepts were converted to a concept map by the CMfL tool described in the methodology section of this study. Then PFsim scores for each student's concept map were obtained by comparing the student's map with the referent map.

Although participants of this study completed both the iANOVA and CMfL tasks three times: before intervention (i.e., pre-test), right after intervention (i.e., post-test), and one week after intervention (i.e., follow-up test), the main hypotheses were tested only with pre/post-test data because of the small sample size at follow-up test since many participants failed to return for the follow-up. Later, however, results of the follow up test will also be reported.

QUANTITATIVE RESULTS

Initially, a comparison of perceived level of anxiety and statistics experience was conducted to see if there was any difference between participants in the control and intervention group regarding these two variables before they began the study. An independent-samples t-test revealed that there was not a significant difference in the number of statistics classes taken by participants in intervention group (M=1.29, SD=.66) and control group (M=1.32, SD=.69) conditions; t (51) =.012, p =.854, d = 0.04. Similarly, there was not a significant difference in the level of perceived anxiety in intervention group (M=2.76, SD=1.09) and control group (M=2.89, SD=1.06) conditions; t (51) =.448, p =.656, d = 0.12. These results suggest that participants both in control and intervention groups had the same level of statistics experience and anxiety before this study indicating that randomization of participants into two different groups was successful. Further, another independent samples t-test was conducted to explore if there is any statistically significant difference between men and women participants with respect to key measures such as perceived level of anxiety, statistics experience, and performance on the iANOVA and Concept Mapping tasks. The results revealed that there was a significant difference in the perceived level of anxiety reported by women (M=3.08, SD=1.03) and men (M=2.14, SD=.864); t (51) = 3.01, p =.004, d = 0.98. However, there was not a statistically significant difference in the number of statistics classes taken by women (M=1.26, SD=.59) and men (M=1.43, SD=.85); t (51) =.082, p

=.413, *d* = 0.23. Similarly, there was not a significant difference in the overall performance on the iANOVA task performed by women (*M*=2.75, *SD*=.88) and men (*M*=3.05, *SD*=.60); *t* (51) =1.15, *p* =.255, *d* = 0.39; nor were the performance on the Concept Mapping task performed by women (*M*=.31, *SD*=.12) and men (*M*=.29, *SD*=.12) before the intervention; *t* (51) =.519, *p* =.606, *d* = 0.16. These results suggest that although there was a significant difference in the perceived level of anxiety between women and men participants before the intervention, there was not a significant difference in their performances in terms of key measures such as iANOVA and Concept Mapping performance before the intervention.

In order to confirm the validity of any conclusions reported later in this section, normality and homogeneity of variance checks for ANOVA models were also inspected. As a visual method for checking the normality, Normal Q-Q plots were examined for all outcome variables including MAD scores for small and large datasets as well as PFsim scores within each group. Results revealed no obvious non-linear patterns which indicates the normality of outcome variables. In addition, the frequency distributions (i.e., histograms) for all observed values appear relatively normal. To test the homogeneity of variance, the rule of thumb suggested by Maxwell and Delaney (1990), $\frac{n_{maximum}}{n_{minimum}} * \frac{s^2_{maximum}}{s^2_{minimum}} < 4$ was used to see if the homogeneity of variance was met for both MAD analysis and PFSim analysis. The results indicated that the homogeneity of variance was not violated for the analyses conducted to test main hypothesizes. Further, residual versus predicted (i.e., group) plots revealed similar spread at each level of outcome variables. Now, let us have a closer look at the main quantitative results.

Comparison of iANOVA performance on pre- and post-tests. As indicated in the methodology section of this study, participants were provided with two different sizes of datasets (i.e., small versus large) in the iANOVA task to compare their intuitive reasoning ability. A 2 Size (small, large) X 2 Time (pre-test, post-test) X 2 Group (control, intervention) Split-plot Factorial ANOVA with repeated measures on Size and Time factors was conducted on participants' MAD scores in order to compare participants' performance on the iANOVA task for small and large datasets before and after the intervention. The interaction of Size X Time X Group was not significant, $F(1, 52) = 1.57$, $p > 0.0$, $\eta^2 = 0.029$, nor were the effect of time, $F(1, 52) = .45$, $p > 0.05$, $\eta^2 = 0.001$, the Time X Group interaction, $F(1, 52) = 1.10$, $p > 0.05$ $\eta^2 = 0.021$, and the main effect of Size, $F(1, 52) = 3.24$, $p > 0.05$, $\eta^2 = 0.058$. However, the Size X Group interaction was statistically significant, $F(1, 52) = 4.50$, $p < 0.05$, $\eta^2 = 0.079$.

Table 3

Mean MAD scores by Group and Data Size

Group	*Size*	*Mean*	*SD*
control	small	1.45	.093
	large	1.47	.121
intervention	small	1.46	.096
	large	1.20	.125

As seen in Table 3, MAD scores of participants in the intervention group were smaller for large size datasets (M=1.20) than small size datasets (M=1.46) while the participants' performance in the control group did not significantly differ for small (M=1.45) and large data sizes (M=1.47). Recall that smaller MAD scores indicate better performance on the iANOVA task as smaller differences represent a smaller gap between participants' performance and normative logic of ANOVA.

Comparison of Concept Mapping Performance before and after Intervention.

To compare participants PFsim scores in control and intervention groups before and after the intervention, A 2 Group (control, intervention) X 2 Time (pre-test, post-test) Split-plot Factorial ANOVA was performed with repeated measures on the second factor. The main effect of Group was not statistically significant, $F(1, 52) = .14$, $p > 0.05$, $\eta^2 = 0.003$. However, the main effect of Time was significant, $F(1, 52) = 3.25$, $p < 0.05$, $\eta^2 = 0.059$. Also, the Time X Group interaction was statistically significant, $F(1, 52) = 6.02$, $p < 0.05$, $\eta^2 = 0.104$.

Table 4

Mean PFsim scores by Group and Time

Group	*Time**	*Mean*	*SD*
control	1	.331	.023
	2	.318	.027
intervention	1	.269	.024
	2	.358	.028

* 1 refers to pre-test and 2 refers to post-test

As seen in Table 4, mean PFsim scores of the participants in the intervention group were higher after the intervention (*M*=0.358) than before (*M*=0.269) whereas the PFsim scores of participants in the control group did not increase from pre-test (*M*=0.331) to post-test (*M*=0.318). Although the control group's overall similarity scores were higher than the intervention group before on the pre-test, these results suggest that the Pathfinder networks of the participants in the intervention group were more like the referent network than those of the control group on the post-test. However, as will be seen below, these results did not hold up at the follow up test when participants came back a week later to complete the test a third time.

It is worth noting that up to this point, the main hypotheses of this study were tested only with data derived from the pre and post-tests (i.e., two time points) due to the small sample size resulting in lack of power at follow-up test. The following analyses are based on the data from that smaller subset of participants who returned for the follow-up.

Comparison of iANOVA Performance across Three Time Points. A 2 Size (small, large) x 2 Group (control, intervention) x 3 Time (pre, post, and follow up) Split-plot ANOVA was also performed on participants' MAD scores. The main effect of Time was significant, $F(2, 54) = 3.79$, $p < 0.05$, $\eta^2 = 0.122$. However, the main effect of Group $F(1, 27) = 0.510$, $p > 0.05$, $\eta^2 = 0.019$, the Size x Time x Group interaction, $F(2, 54) = 1.72$, $p > 0.05$, $\eta^2 = 0.061$, the Size × Group interaction, $F(1, 27) = 0.473$, $p > 0.05$, $\eta^2 = 0.017$, the interaction of Time × Group, $F(2, 54) = 1.03$, $p > 0.05$, $\eta^2 = 0.037$, and the Size × Time interaction, $F(2, 54) = 0.812$, $p > 0.05$, $\eta^2 = 0.029$, were not statistically significant.

Table 5

Mean MAD Scores by Time

Time*	*Mean*	*SD*
1	1.288	.093
2	1.295	.101
3	1.045	.107

* 1 refers to pre-test, 2 refers to post-test, and 3 refers to follow up test

Overall performance of participants both in the intervention and control conditions was better in the follow up test which took place a week after the intervention.

Comparison of Concept Mapping Performance across Three Time Points. In addition to the analysis on MAD scores, a 3 Time (pre, post, and follow up) X 2 Group (control, intervention) Split-plot ANOVA was performed on participants' PFsim scores. The main effect of Time, $F(2, 58) = .887$, $p > 0.05$, $\eta^2 = 0.030$, the main effect of Group, $F(2, 58) = .570$, $p > 0.05$, $\eta^2 = 0.019$, and the Time X Group interaction, $F(2, 58) = .681$, $p > 0.05$, $\eta^2 = 0.023$, were not statistically significant. Overall, it was observed that participants' PFsim scores in the control condition were staying flat across the three tests whereas PFsim scores of participants in the intervention group were increasing slightly during the study (see Figure 11 below). However, this increase in the intervention group was not large enough at the follow up test to be considered as statistically significant due to possibly lack of power as a result of small number of participants at the follow up test.

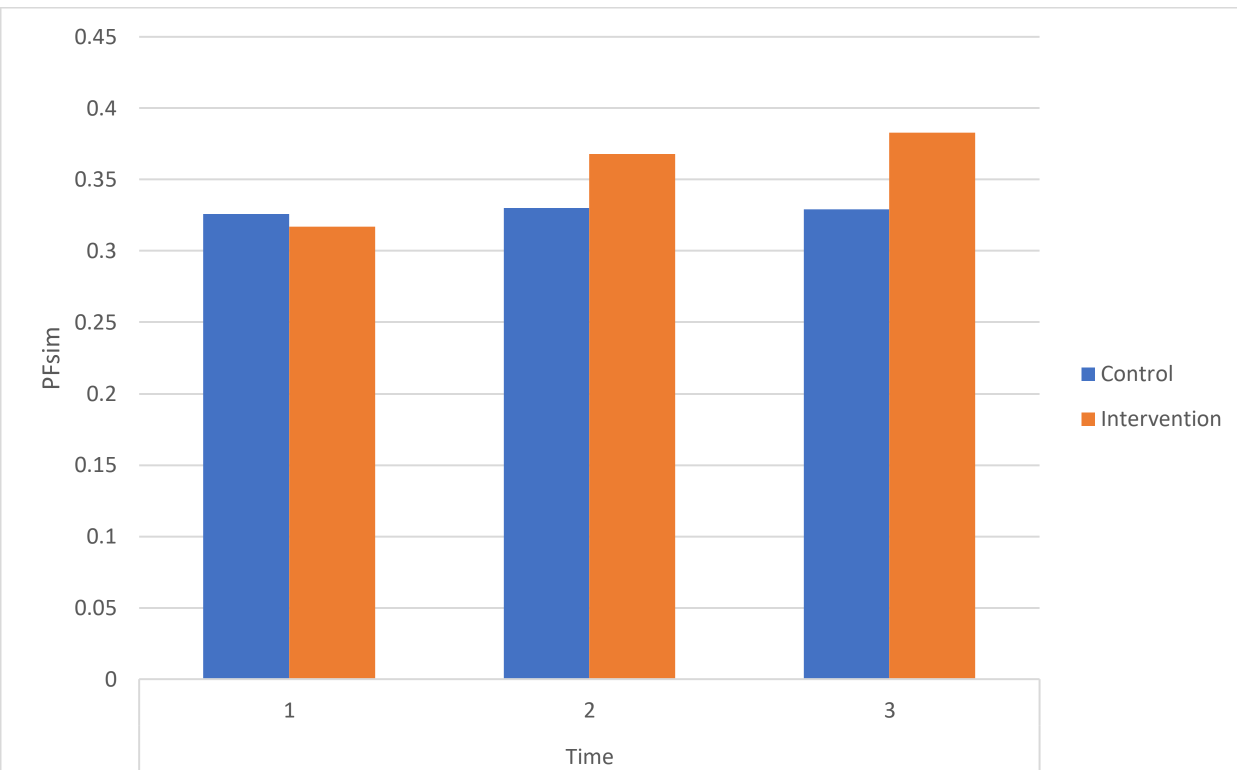


Figure 11. Mean PFsim scores for control and intervention group participants by three time points

Relation between MAD and PFsim scores. To answer the second research question, "Is there a relation between participants' performance on the Informal Inferential Reasoning task and conceptual understanding related to ANOVA measured through Structural Assessment of Knowledge?", Pearson correlations were computed between MAD scores for smaller and larger datasets and PFsim scores across three time points. As seen in Table 6 below, although pre- and post-test MAD scores were significantly, positively correlated with each other, as were pre- and post-test PFsim scores, PFsim scores were not significantly correlated with any of the MAD scores. Table 6 below displays the correlation coefficients between variables.

Table 6

Correlations between MAD and PFsim Scores

	1	2	3	4	5	6	7	8	9
1- SmMAD1	1	.685**	.168	.362**	.421*	.404*	-.011	-.148	-.021
2- LgMAD1	.685**	1	.256	.546**	.462*	.465*	.039	-.185	-.154
3- SmMAD2	.168	.256	1	.478**	.339	.272	.203	.008	-.010
4- LgMAD2	.362**	.546**	.478**	1	.390*	.287	.259	-.147	-.198
5- SmMAD3	.421*	.462*	.339	.390*	1	.444*	.136	-.096	.089
6- LgMAD3	.404*	.465*	.272	.287	.444*	1	-.072	.401*	.257
7- Pfsim1	-.011	.039	.203	.259	.136	-.072	1	.274*	.328
8- Pfsim2	-.148	-.185	.008	-.147	-.096	.401*	.274*	1	.408*
9- Pfsim3	-.021	-.154	-.010	-.198	.089	.257	.328	.408*	1

**Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).
1 refers to MAD scores for smaller datasets before intervention
2 refers to MAD scores for larger datasets before intervention
3 refers to MAD scores for smaller datasets after the intervention
4 refers to MAD scores for larger datasets after the intervention
5 refers to MAD scores for smaller datasets 1 week after intervention
6 refers to MAD scores for larger datasets 1 week after intervention
7 refers to PFsim scores before intervention
8- refers to PFsim scores after intervention
9- refers to PFsim scores 1 week after intervention.

QUANTITATIVE DISCUSSION

The purpose of the quantitative analysis was to determine if individualized formative feedback that focuses on conceptual understanding of ANOVA helps improve participants' performance on the Informal Inferential Reasoning task (i.e., intuitive ANOVA task). Before getting into the actual analysis, participants' level of perceived statistics anxiety on a scale of 1 to 5 and statistics experience (as recorded by the number of statistics classes taken) was measured. There was no statistically significant difference in the level of perceived statistics anxiety and number of statistics classes taken by participants in the control and intervention group. Overall, three main findings were obtained.

First, in an effort to improve conceptual understanding of ANOVA, participants in this study were provided with individualized feedback based on their pre-test knowledge structures of ANOVA concepts derived based on relatedness ratings of these concepts. Overall, it can be concluded that the conceptual networks of participants in the intervention group got closer to the referent network after the intervention. In other words, intervention group participants' knowledge related to critical concepts of ANOVA was increased significantly after the individualized feedback. Similarly, prior studies also found significant increase in students' knowledge structures after instruction (Goldsmith & Davenport, 1990; Jonassen, 1993; Kraiger & Cannon-Bowers, 1995) as well as after providing them with individualized feedback (Trumpower & Sarwar, 2010; Sarwar, 2011). However, this result did not retain at the follow up test one week after the intervention.

Second, participants' performance on the iANOVA task was measured to examine whether participants' Informal Inferential Reasoning abilities become more consistent with the normative logic of ANOVA after the intervention. This did not occur. That is said, consistent with the normative logic of ANOVA, participants in the intervention condition perceived larger datasets as providing stronger evidence than smaller datasets. However, this favorable result was not due to the effect of intervention since only the main effect of “Time” was statistically significant and the “Time by Group” interaction was not significant. Rather, the participants in the intervention group performed better on larger datasets both before and after the intervention. It was expected in this study that better conceptual understating related to ANOVA would result in enhanced Informal Inferential Reasoning which was measured by performance on the iANOVA task. But, even though participants’ conceptual understanding was somewhat improved through feedback on critical concepts, as indicated by Structural Assessment of Knowledge, this understanding did not transform into applied Informal Inferential Reasoning as indicated by iANOVA performance. This might be due to several reasons as follow:

(a) The relevant literature in statistics education often acknowledges the complexity of statistical terms such as variation and probability to students from non-statistics majors as well as the difficulty regarding overcoming misconceptions related to these key terms. For example, after conducting extensive research in statistics education, Konold (1995) marked three major findings: (1) students come into our courses with some strongly-held yet basically incorrect intuitions, (2) these intuitions prove extremely difficult to alter, and (3) altering them is complicated by the fact that a student can hold multiple and often contradictory beliefs about a situation" (Konold, pg. 2). Furthermore, he has concluded "we have a variety of data suggesting that these intuitions are persistent and, to this point, survive our best teaching efforts" (ibid, pg. 6). Thus, the feedback provided may not have been sufficient enough to alter these strongly held incorrect intuitions especially given that the feedback was not directly related to the Informal Inferential Reasoning required to complete the intuitive ANOVA task. Rather, the feedback was designed to highlight the logic behind procedures of ANOVA. It was expected that the increased conceptual understanding through feedback on the logic of ANOVA would translate into the better Informal Inferential Reasoning. It seems the conceptual understanding related to these terms did not transform into the Informal Inferential Reasoning as much as expected as this was evident in the results of the quantitative data analysis where there was no statistically significant correlation between students' MAD scores (i.e., designed to measure Informal Inferential Reasoning) and PFsim scores (i.e., designed to measure conceptual knowledge). Also, increasing conceptual understanding related to variation that transfers to Informal Inferential Reasoning may require more extensive training on the subject matter over prolonged time period through combination of various traditional and non-traditional teaching methods as Reid and Reading (2008) state that it is not realistic to expect students at the end of a one semester introductory course to show signs for strong consideration of variation. Overall, it can be concluded that altering misconceptions related to variation in statistics might be more difficult than one might think as related research highlighted that even secondary-school mathematics teachers demonstrated difficulties with the concept of variation while comparing two groups even after extensive training (Makar & Confrey, 2004).

(b) Related to the first reason, although some participants started to comprehend the critical concepts highlighted in the feedback as measured by the PFsim scores before and after the intervention, it is worthwhile to note that these participants did not have an opportunity to reflect on the feedback or practice the feedback with more real world problems due to the time constraints and online nature of the study. The total time required to complete the study was about two hours including completing the iANOVA and concept relatedness rating tasks twice and going through written and visual types of feedback. In order to make sure that participants paid attention to the individualized feedback throughout the study, the researcher designed the flow of the experiment in a way that each student had to spend at least five minutes with each type of feedback. The researcher did not want to extend the individual time spent with feedback considering the possible quantity of feedback each individual might receive was based on the common, missing, and extraneous links on their concept maps. Participants were required go through individualized written and visual feedback without having to explicitly demonstrate their conceptual understanding. As a consequence, the intervention in the present study may have missed the requisite depth of engagement/reflection as Trumpower (2013) highlights that effective formative feedback requires deeper and more engaging reflection on the conceptual understating of ANOVA. Consistently, Garfield and Ben-Zvi (2007) suggested that effective learning in statistics requires active involvement of students in constructing knowledge. In addition, Nicol and Macfarlane-Dick (2006) indicated that effective feedback presents opportunities for reflection during the learning process. Overall, participants did not have a chance to perform extensive reflection activities and, thus, lacked the essential deeper engagement with the feedback. Consequently, although intervention participants seemed to improve conceptual understanding somewhat, this was not good enough to translate into normative Informal Inferential Reasoning.

Third, participants were invited to complete both the Concept Mapping Task and iANOVA task a week after the intervention. Quantitative results indicated that both control and intervention group participants' performance on the iANOVA task at the follow-up was better compared to the previous two tests. One potential explanation for this result might relate to the small select number of participants from both groups in the follow up test. It is possible that only those participants who were really interested in the study joined the follow up test. Thus, participants from both groups had a better performance.

QUALITATIVE ANALYSIS

Variation is crucial for a deep and full understanding of ANOVA. Thus, identifying the degree to which students consider and reason about variation is important to determine about the cognitive development of this key term. As part of getting a better insight into participants' reasoning about variation in this study, participants' written explanations for their perceived strength of evidence for the hypothetical datasets presented in the iANOVA task were examined using a coding scheme, *Consideration of Variation Hierarchy*, adopted from Reid and Readings (2008), to look for evidence of any changes in level of consideration of variation indicative of deeper understanding. As defined in the methodology section of this study, Reid and Readings' hierarchy suggests analyzing students' responses related to variance in three categories: *weak, developing*, and *strong consideration of variation.* According to the hierarchy, *weak consideration of variation* includes consideration of either between-group or within-group differences, but not both. As a second level, *developing consideration of variation* includes explanations that consider within- and between-group differences but in a non-integrated manner. Finally, *strong consideration* of variation involves considering both within- and between-group differences relative to one another in a given comparison situation and is consistent with the normative logic underlying analysis of variance.

Although Reid and Reading's hierarchy is comprised of three levels and suggests classifying responses with respect to consideration of variation using these three levels, the present study expanded these levels as some participants' responses did not fit into these categories. For example, some of the participants of the present study did not consider variation at all. Specifically, they did not mention either between or within group variation in their responses. In addition, some participants indicated only some other relevant factors such as sample size and random factors that could affect the hypothetical results. Therefore, a new category titled "*unable to determine"* was created to classify such responses. As another example, according to Reid and Reading's hierarchy, weak consideration of variation involves recognizing either between-group or within-group variation. However, the present study expanded this category into two separate categories as *consideration of within group variation only* and *consideration of between group variation only* because responses which indicate only within group variation or only between group variation are quite distinct in term of a student's understanding related to ANOVA. That is said, while within group variation refers to the average standard deviation of scores within groups in a dataset (e.g., the average of the standard deviations of the 14-panel and 32-panel balls) and is a result of the effect of random factors only, between group variation refers to the difference between the group means in a dataset (e.g., the difference between the mean of the 14-panel balls and the mean of the 32-panel balls) which can be a result of both independent variable and random factors. Thus, students' responses which indicate only within group variation or only between group variation represent two different reasoning approaches while dealing with ANOVA cases. In addition, the category of consideration of between group variation was also divided into two categories as *consideration of only between group variation in a holistic manner* and *consideration of only between group variation individually* as these two approaches are also distinct and need a closer look. To further clarify the distinction between these two categories, the first one includes reasoning which focuses on the overall mean for each group and the difference between group means while the second one involves focusing on the magnitude of selected pairwise differences where the pairs are scores from each column (and not necessarily in the same row) are specifically selected to show large differences. Thus, the six-category hierarchy depicted in Table 7 below was used to examine students' responses for each dataset in the iANOVA tasks.

Table 7

Consideration of Variation Hierarchy adapted from Reid and Readings (2008)

1. Unable to determine
- Not enough information to make a holistic examination of a response (i.e., not a complete response).
- No mention of either between or within group variation
- Indication of small sample size only

2. Consideration of only between group variation in a holistic manner
- Identify only sources of between-groups variation with a holistic approach (i.e., focus on the overall mean for each group and the difference between group means).

3. Consideration of only between group variation based on individual data points
- Identify only sources of between-groups variation based on differences between distinct data points (i.e., focus on the magnitude of selected pairwise differences where the pairs are scores from each column [and not necessarily in the same row] specifically selected to show large differences between groups).

4. Consideration of within-group variation only
- Identify only sources of within group variation using related terms such as spread, variability within the dataset, standard deviation etc.

5. Developing consideration of variation
- Indicate both between-and within group variations in different ways (i.e., focus on between and within group variation as independent entities or focus on consistency of pairwise comparisons of individual data points)
- Does not use the within-group and between-group variation in a relative manner to support inferences (e.g., participants do not compare the magnitude of the within and between group variation. They use these two variation types as two independent entities).

6. Strong consideration of variation
- Link within-group and between-groups variation to support inference (e.g., participants consider the magnitude of the within group variation in relation to between group variation as the magnitude of big between group variation depends on the magnitude of big within group variation).

Given the expanded hierarchy above, each student's written explanation for their perceived strength of evidence ratings for the twelve hypothetical datasets on the iANOVA task was coded as indicating one of the following categories described above: *unable to determine*, *consideration of between group variation in a holistic manner*, *consideration of between group* *variation individually*, *consideration of within-group variation only*, *developing consideration of variation*, and *strong consideration of variation*. In general, an overview of responses in the control group for consideration of variation for small and larger datasets demonstrated similar patterns to those in the intervention group. Therefore, only the responses of participants from the intervention group were examined to observe the changes before and after intervention, if any, since the participants in the control group did not receive any feedback between pre- and post-tests. Initially, each participant's responses across datasets were explored to determine whether the participant employs a consistent strategy (e.g., focusing only on within group differences) across datasets regardless of data size and/or manipulated variation. However, preliminary results of such an analysis revealed that there is no visible pattern with respect to reasoning strategies used by participants across datasets. Most of the participants employed multiple strategies while dealing with different datasets. Therefore, the frequency of each strategy described above for a given dataset was examined to see if there is a consistency in terms of strategies used by participants for a given dataset. Although previous studies reported that participants were consistent with respect to reasoning strategies used across datasets (e.g., Read & Readings, 2008; Trumpower, 2015), the use of consistent reasoning strategies across datasets by participants of these studies might be due to the presentation of one type of dataset, namely only small or large datasets. Slightly different from previous studies, the present study asked participants to provide reasoning for both small and larger datasets with varying between/within group variation in a random manner within the same task.

The total number of participants included in the qualitative analysis was 27. Considering each participant provided justification for 12 datasets, once before and once after the intervention, the total number of responses examined were 648. Two independent coders classified each participant as using one of these six strategies based on the explanations that they provided for their ratings. An interrater analysis was conducted to determine the level of agreement between coders. The analysis of interrater reliability was Kappa=.61, p<.001. According to Cohen (1968), the Kappa result can be interpreted as follows: values ≤ 0 indicate no agreement, 0.01–0.20 none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. The reason why the agreement here falls onto the lower level of substantial agreement might be related to the use of a new framework to code the qualitative data. As reported before, the researcher had to use an extended version of the Reid and Reading's framework as most of participants' responses did not fit into the exiting framework. However, any discrepancies between the coders were discussed until consensus was reached (see Reid & Reading, 2008; Trumpower, 2015 for similar coding protocols).

Before delving into actual analysis of responses, it is worth noting the slight difference between iANOVA tasks used before and after the intervention in the present study. Although the size, the mean and the standard deviation of these datasets remained the same, the pre- and post-intervention iANOVA tasks used two different cover stories in which the datasets were presented. Participants' written responses incorporated the language from cover stories presented in these iANOVA tasks. Therefore, a reminder of the cover stories given in the iANOVA tasks may be useful with respect to understanding participants' responses better. The first iANOVA task presented a story where two scientists wanted to test if the number of panels on a soccer ball has an effect on the flight characteristics of balls, while the second cover story was about comparing the effectiveness of two different diets on lowering the triglyceride level.

QUALITATIVE RESULTS

Figures 12, 13, and 14 present an overall portrait regarding the distribution of reasoning strategies across different datasets, as well as the total number of responses under each strategy used before and after intervention across datasets. It should be noted that the aim of these three figures is to provide insights into the overall distribution of responses for each category determined above rather than exploring changes in participants' reasoning from before to after the intervention. Later, the detailed description of changes in reasoning will be provided in the related section for each category below. Whereas Figure 12 presents the frequency of reasoning strategies used by participants for each smaller dataset before and after the intervention, Figure 13 demonstrates the frequency of reasoning strategies used by participants for each larger dataset before and after the intervention. Additionally, Figure 14 shows the total number of responses for each reasoning strategy for smaller and larger datasets together before and after the intervention.

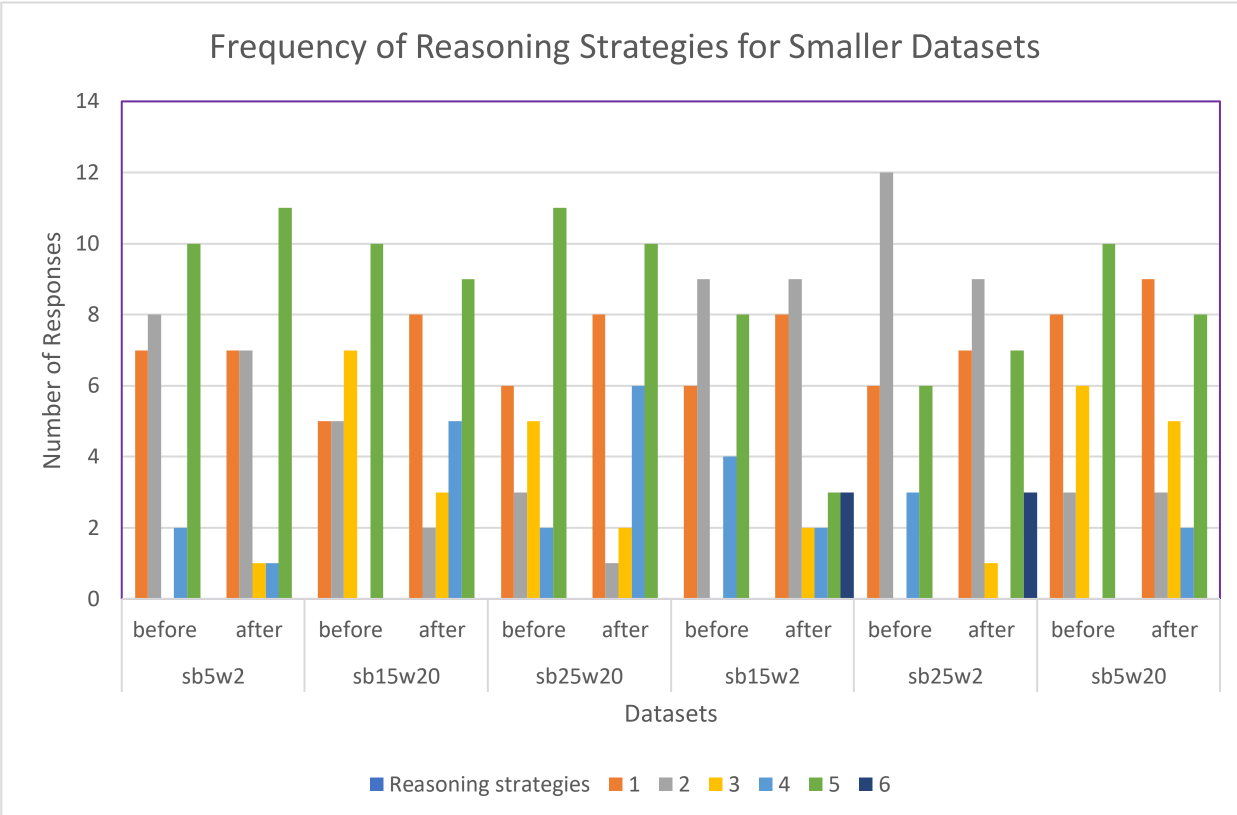


Figure 12. The frequency of reasoning strategies used by participants for six smaller datasets presented in the iANOVA tasks.

1.*Unable to determine*
2. *Consideration of between group variation in a holistic manner*
3. Consideration of between group variation individually
4. Consideration of within-group variation only
5. Developing consideration of variation
6. Strong consideration of variation

As a short summary highlighting the key results presented in Figure 12 above, for smaller datasets the number of responses belonging to category-1 where either the response was not complete or indicated small sample size as a main reason for their ratings were steady and visible across smaller datasets. Further, the number of responses classified as category-2 where participants' reasoning mainly focused on only the between group difference in a holistic manner were higher for datasets in which the within group difference was smaller regardless of the magnitude of between group difference as in datasets sb5w2, sb15w2, sb25w2. Additionally, the responses under category-3 where participants focused on selective pairwise differences were higher for datasets in which the within group variation is larger as in datasets sb15w20, sb25w20, and sb5w20. Moreover, the number of responses indicating a focus on only within group variation, category-4, as a reason varied across smaller datasets. As another distinctive finding, responses belonging to catgory-5, focusing on both within and between group differences in a non-relative manner, for small datasets were quite high compared to the other strategies across datasets before and after the intervention. A detailed discussion of possible reasons behind the high number of responses under the category-5 will be discussed in the associated section below. As category-6 represented strongly developed understanding of variation, there were very few responses under this category.

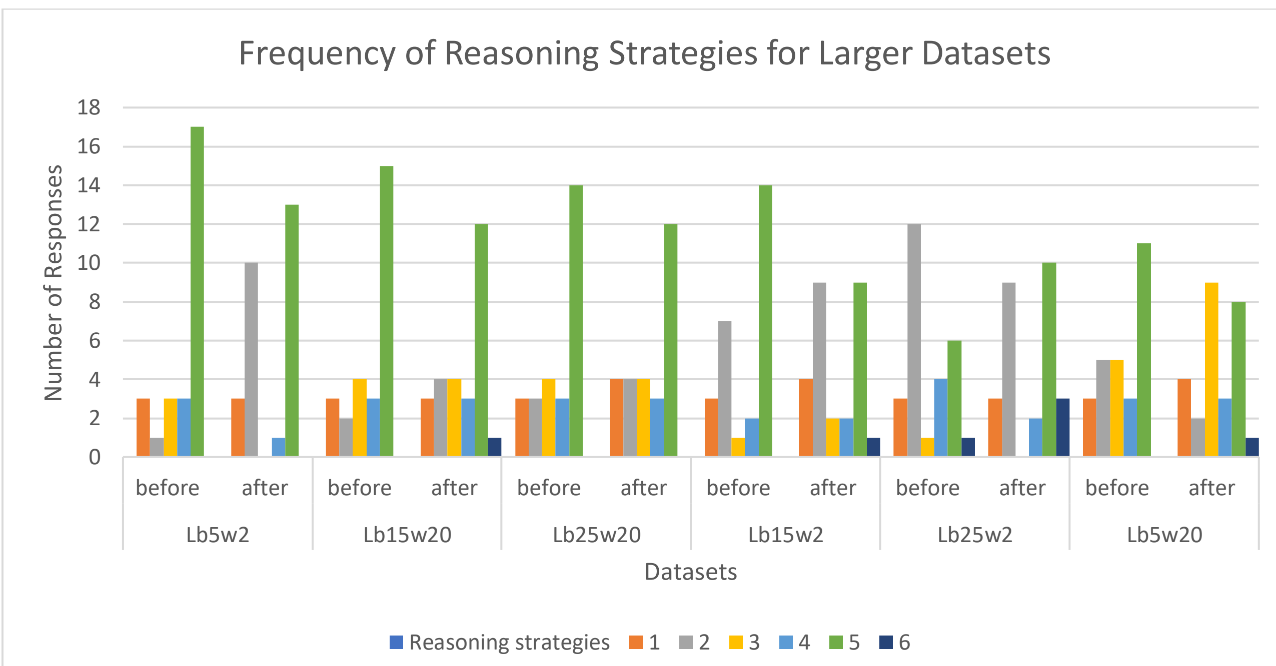


Figure 13. The frequency of reasoning strategies used by participants for six larger datasets presented in the iANOVA tasks.

1.*Unable to determine*
2. *Consideration of between group variation in a holistic manner*
3. *Consideration of between group variation individually*
4. *Consideration of within-group variation only*
5. *Developing consideration of variation*

6. Strong consideration of variation

Although the number of responses indicating small sample size for larger datasets, category-1, was noticeably lower than those for smaller datasets, results for the larger datasets also presented similar patterns of the reasoning strategies (see Figure 13). For example, the number of responses indicating a focus on only between group difference was higher for datasets in which the between group difference is larger than smaller.

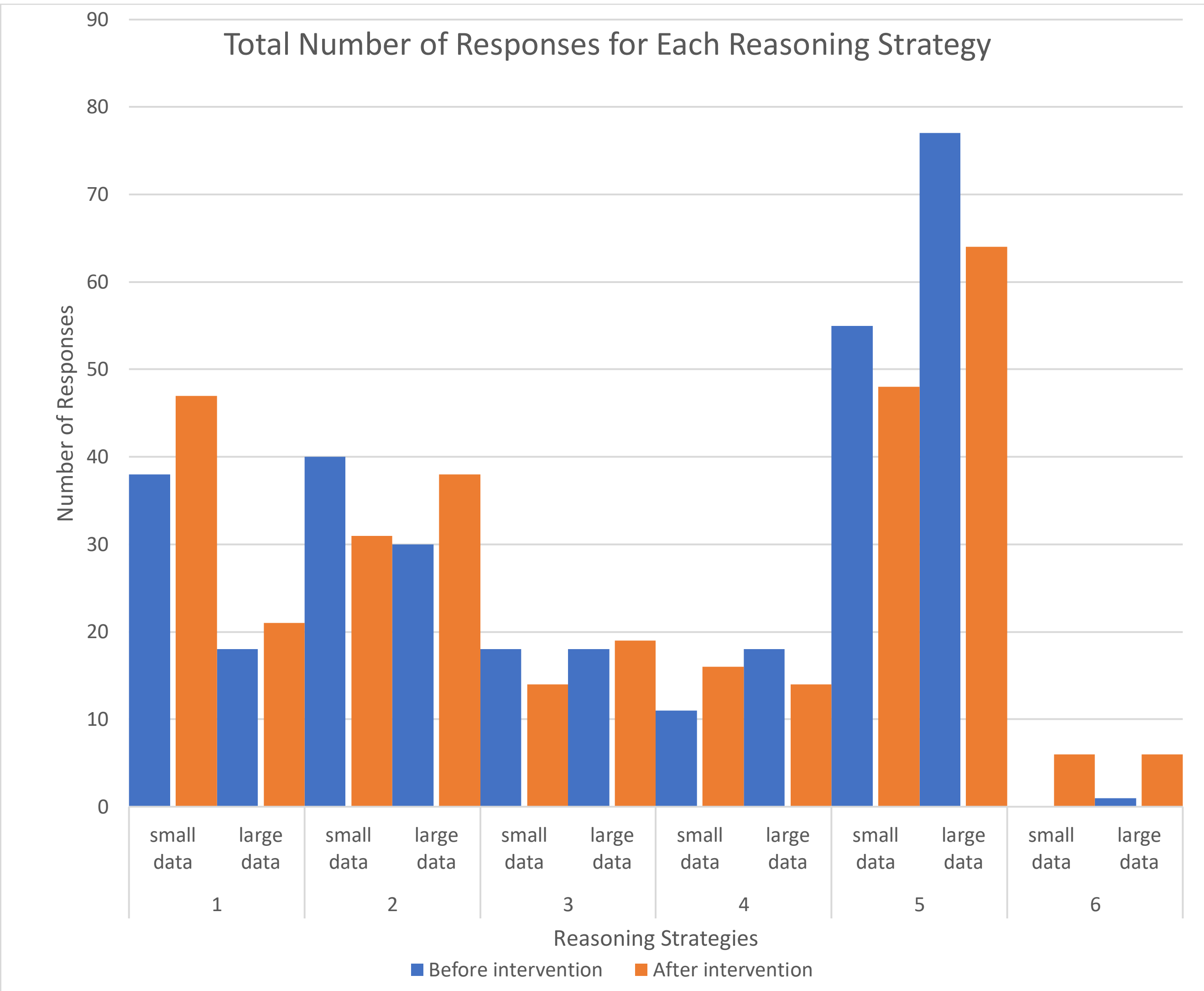


Figure 14. The total number of responses for each reasoning strategy used by participants before and after the intervention for datasets presented in the iANOVA tasks.

1.*Unable to determine*
2. *Consideration of between group variation in a holistic manner*
3. *Consideration of between group variation individually*

4. Consideration of within-group variation only
5. Developing consideration of variation
6. Strong consideration of variation

As seen in Figure 14, the total number of responses classified under category-5 was highest at 243 responses out of 648 in total followed by 191 responses at category-1, 139 responses at category-2, 69 responses at category-3, 59 responses at category-4, 14 responses at category 6. Overall, it is also worth to note that the language that participants in the intervention group used in their responses was also shifted following the intervention towards the terminology used in the feedback given through intervention. More detailed descriptions, examples, and discussion of the six strategies, as well as the number of responses classified as using each, follow.

Category-1: Unable to determine (n=124). As described before, participants' responses in this group either did not present a complete response for analysis or did not incorporate the variation as a main factor for their reasoning or indicated only "small sample size" as a justification for their low level of confidence in the evidence provided by the respected datasets in the iANOVA task. This was a unique finding from the qualitative data that some participants of the present study were acknowledging the effect of sample size on the strength of evidence provided by datasets in the iANOVA tasks.

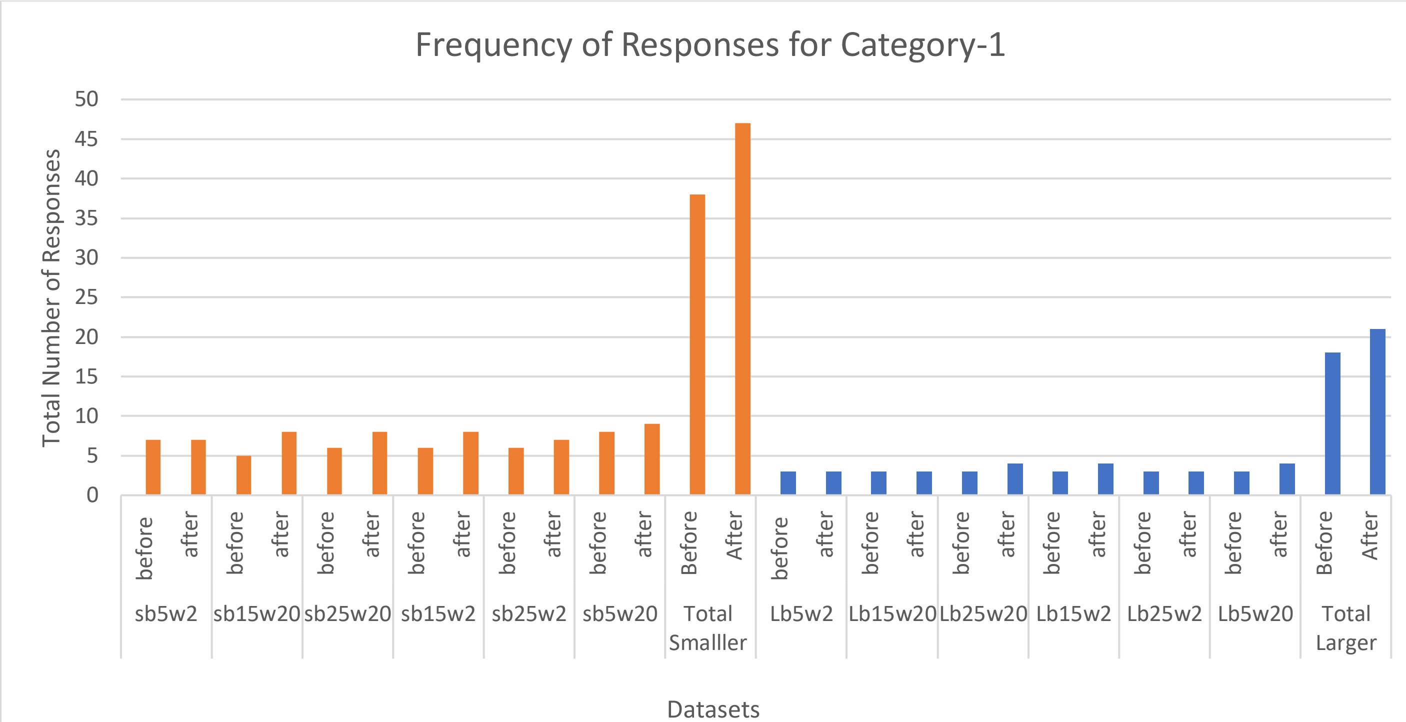


Figure 15. The frequency of responses for Category-1 across small and larger datasets.

As seen in Figure 15, for smaller datasets, the total number of responses classified as *unable to determine* was 38 before the intervention and 47 after the intervention while it was 18 before the intervention and 21 after the intervention for the larger datasets. It should be noted that almost 90 per cent of the responses in this category actually related to the indication of "*only small sample size*" while the rest was related to irrelevant or incomplete responses. While there was no significant change in the number of responses in this category over time, the total number of responses indicating lower level of confidence due to small sample size was more than twice as high for the smaller datasets (85) than the larger datasets (39). These results indicate that participants did not feel confident enough while judging the strength of evidence provided by smaller datasets even though some of the smaller datasets presented statistically significant evidence to support the hypothesis that one type of ball/diet is better than the other one by formal ANOVA procedures. Further, the lower level of perceived confidence for smaller datasets was also evident in participants' responses. For example, one of the participants stated very clearly that "Less numbers=less confidence" while dealing with a small size dataset. Similarly, participant #11 noted "there are less results and observations here, thus making me way less confident". Some of the other responses from participants were as follows:

"There are not enough results to make a proper conclusion", "More data is needed", "Sample size is too small", "Only three trials", and "I am more confident again because again there are more trials and results".

As highlighted before, this category also included a small number of participants whose responses were either irrelevant or incomplete. For instance, while judging the dataset sb25w20, participant #23 noted only, "It's not possible" or similarly another participant wrote, "Very strong evidence" for another dataset. These types of responses represented only a small portion of the responses in this category and were difficult to interpret.

Category 2: Consideration of only Between Group Variation in a Holistic Manner (n=140). The responses of the participants in this category focused on the magnitude of the overall mean for each group as well as the difference between the means of the two groups. In other words, these responses were mainly influenced by the between groups variation where participants used only between groups variation in a holistic manner for justifications of their ratings for hypothetical datasets. The holistic manner here refers to focusing on the magnitude of the *mean* difference between two groups regardless of the data size and within- and between groups variation as well as any individual data points within each group provided in the iANOVA task.

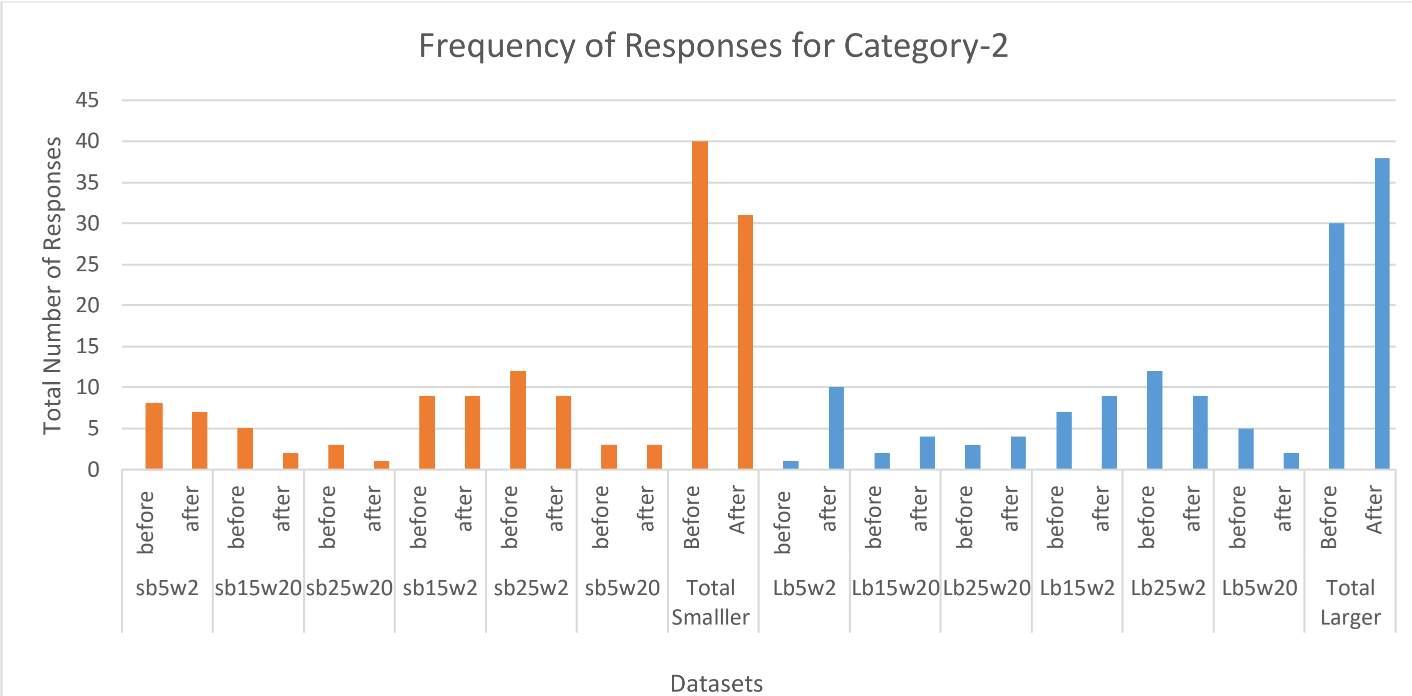


Figure 16. The frequency of responses for Category-2 across small and larger datasets.

As seen in Figure 16, for smaller datasets, the total number of responses classified as *consideration of only between group variation in a holistic manner* was 40 before the intervention while the total number of responses in the same category decreased to 31 after the intervention. The decrease in the number of responses indicating *only between group difference* in their responses after the intervention might be due to an increase in the number of participants who indicated "small sample size" as the main reason for their lower ratings for smaller datasets after the intervention. Recall that the frequency of responses for small size datasets which indicated *small sample size* as a justification for ratings increased from 32 responses to 46 after the intervention. Therefore, a further investigation of the qualitative data confirmed that those participants who considered a relatively small difference between group means in smaller datasets as a meaningful difference before the intervention shifted their mind to indicate that a smaller sample size actually does not provide strong evidence regardless of the magnitude of the between groups variation and indicated small sample size as a reason for the rating after the intervention. As for the larger datasets, the total number of responses classified as *consideration of only between group variation in a holistic manner* was 30 before the intervention while the total number of responses in the same category increased to 38 after the intervention.

Moreover, datasets Sb25w2 and L25w2 which depict larger between groups variation and smaller within group variation (i.e., provides stronger evidence) received higher responses than those datasets with higher between groups variation and higher within group variation (i.e., provides weaker evidence) in this category. In other words, responses focused merely on the between groups variation in cases where the between groups variation was visible regardless of the magnitude of within group variation. For example, participant #6's explanation for the small dataset sb25w2 was as follows: "The 32-panel held an average score of 300, and the 14-panel total average was 325. This was a difference of 25 in favor of the 14-panel". Similarly, while providing justifications for their ratings for the small dataset sb15w2, participant #9 noted that "Small sample size, 14 panel does have higher numbers and therefore higher average". Likewise, participant #17 explained his/her reasoning this way: "The two averages are 15 points apart, I would say this is close to supporting the hypothesis but it is only three scores". Although participants were discouraged from making any formal calculations since the iANOVA task was an informal inferential reasoning task, some participants calculated the mean for each dataset as well as the difference between means. While judging the strength of evidence for the small dataset sb5w2, participant #13 reported "the mean of the 14 panel vs the 32 panel was +5 therefore it is better". Along the same line, participant #13's reasoning for the large dataset Lb15w2 is as follows: "The mean of the 14 panel is even higher on this test so I'm even more confident".

Although a small number of responses in this category were made to datasets where the ratio of between groups variation to within groups variation was small, some responses still perceived larger difference between groups as providing stronger evidence regardless of the magnitude of the within group variation. For instance, to explain his/her ratings for the small dataset sb25w20, participant #20 wrote, “I would say that this data might be able to support the hypothesis because one group has a higher average than the other group”**.** Similarly, while explaining the reason for his/her rating for the larger dataset Lb15w20, participant #6 noted that, “An average of the two would show a higher value in regards towards the 14 panel”. Similarly, participant #20 explained his/her rating for the larger dataset sb25w20 as follows, “higher average than the other group”. It seems 25 units of difference between two samples were enough for this participant to be confident that the fixed factor really had an effect on the independent factor.

The majority of participants in this group continued to employ the same strategy after the intervention. For example, focusing on the average between datasets, participant #7 wrote for the dataset sb25w2, “Although smaller sample the averages are different”. As participant #23 also mentioned “The average of the Bird diet is 200 mg/dl. The average for the Tomkins is 205 mg/dl. This would give some support that the Bird diet reduces the triglyceride level more than the Tomkins diet”. Similarly, for the dataset Lb5w2, participant #7 indicated “Tomkins diet is barely higher than the Bird diet”.

To summarize, consideration of the between groups variation in a holistic manner remained as the main factor influencing most of the responses in this category on datasets with larger between groups variation and smaller within group variation (i.e., reliable differences between two groups). Furthermore, although a smaller number of responses, some responses considered datasets with larger between group variation and larger within group variation as providing relatively strong evidence while ignoring the magnitude of the within group variation. Earlier studies have also found that a majority of participants displayed this same misconception and perceived datasets with larger between groups variation and larger within group variation (i.e., provides weak evidence) as providing stronger evidence (see Masnick & Morris, 2008; Obrecht et al., 2007; Trumpower, 2013b; Trumpower, 2015Trumpower & Fellus, 2008).

Category 3: Consideration of only between groups variation individually (n=69)**.** Although responses in this category also focused on the between groups variation, they did not consider the between groups variation in a holistic manner as in responses presented in catergory-2. Responses in this category identified only sources of between-groups variation with an individual approach. The term "individual approach" here refers to individual, distinct data points (rather than the individuals themselves). All responses in this category focus on the magnitude of selected pairwise differences where the pairs are scores from each column (i.e., group) presumably selected to show large or small differences.

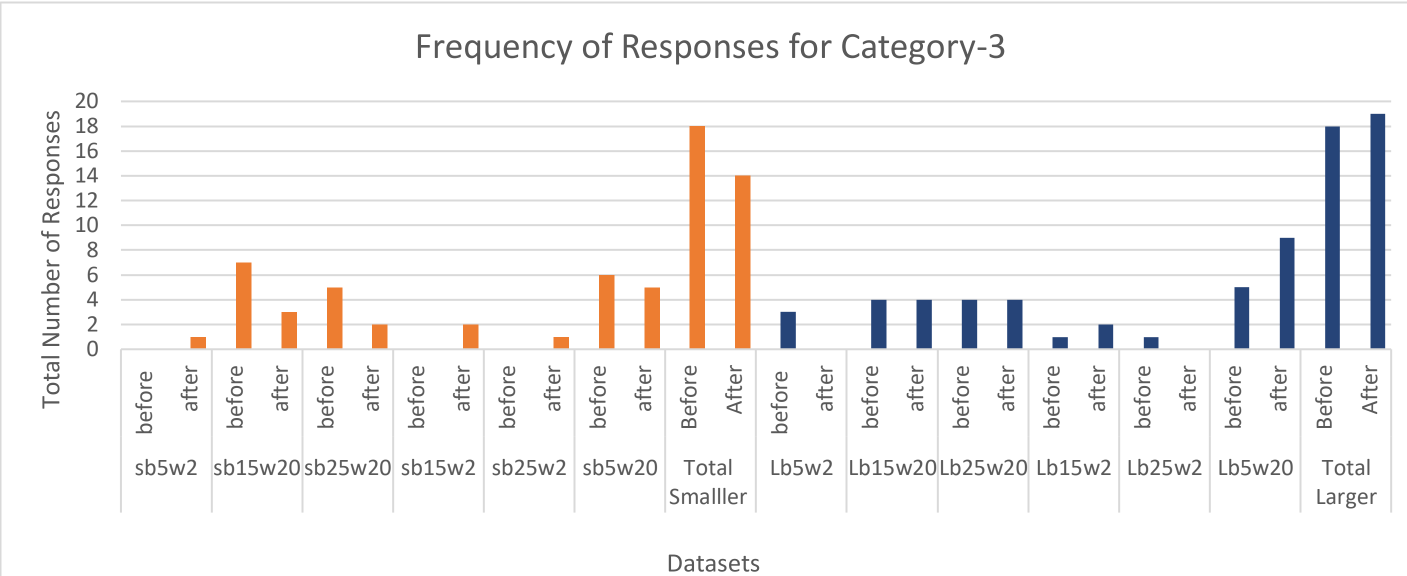


Figure 17. The frequency of responses for Category-3 across small and larger datasets.

As shown in Figure 17, the number of responses in this category did not change noticeably across time and datasets of different size (i.e., before vs. after the intervention, or smaller vs. larger datasets). While the total number of responses in this category was 32 (i.e., 18 before the intervention and 14 after the intervention) for smaller datasets, it was 37 (i.e., 18 before the intervention and 19 after the intervention) for the larger datasets. As seen in Figure 17, the total number of responses belong to this category in the smaller datasets were relatively higher for datasets with higher within group variation. For example, sb5w20 had the highest number of responses with 11 responses followed by sb15w20 with 10 responses, and sb25w20 with 7 responses. Similarly, one of the larger datasets, Lb5w20 with a relatively high within group variation had the highest number of responses in this category with 14 responses while other larger datasets with high within group variation had, on average, 4 responses. Typical responses in this category are as follow:

"It is slightly better evidence than none, however, because there is a higher number for the 14 panel ball even at the lowest (305) compared to the highest of the 32 panel ball (300). Though the difference isn't extreme, and over more trials there is a chance that those numbers will even out."

"The 14-panel has two really high scores and one low score that is the same as the 32-panel"

"Five of the sets of scores favor the 32 panel balls. The other five sets of scores favor the 14 panel balls. One of the sets of scores favor the 14 panel ball over the 32 panel ball by one point. I have no confidence with this dataset."

"Looking at the highest levels of component in Bird diet it was 220 while Tomkins diet was 285 this number highlights a major difference."

"Only one reading for the Bird diet is higher than the lowest reading for the Tomkins diet. I would like to know what the initial triglyceride level of participant of the Tomkins diet with 285 mg/dl. I would have some confidence that the Bird diet reduces the triglyceride level when compared to the Tomkins diet."

Overall, responses focusing on only between group differences from an individual data points perspective appeared to be selectively attaining a high score in one group of data and a low score in the other in a hypothesis confirming manner. That is, the high difference between select data points from two different groups led to seemingly larger between groups variation which in return confirmed the hypothesis that one type of ball or diet was actually better than the other one.

Category 4: Consideration of Within Group Variation Only (n=59). Participants who considered only within group variation in their responses mainly focused on the magnitude of differences that exist within each group but did not mention the between group variation at all while making inferences. These individuals might have considered the between groups variation implicitly. However, their written justifications only mentioned the within group variation. As with the previous category, the number of responses in this category did not change significantly across time and datasets of different size.

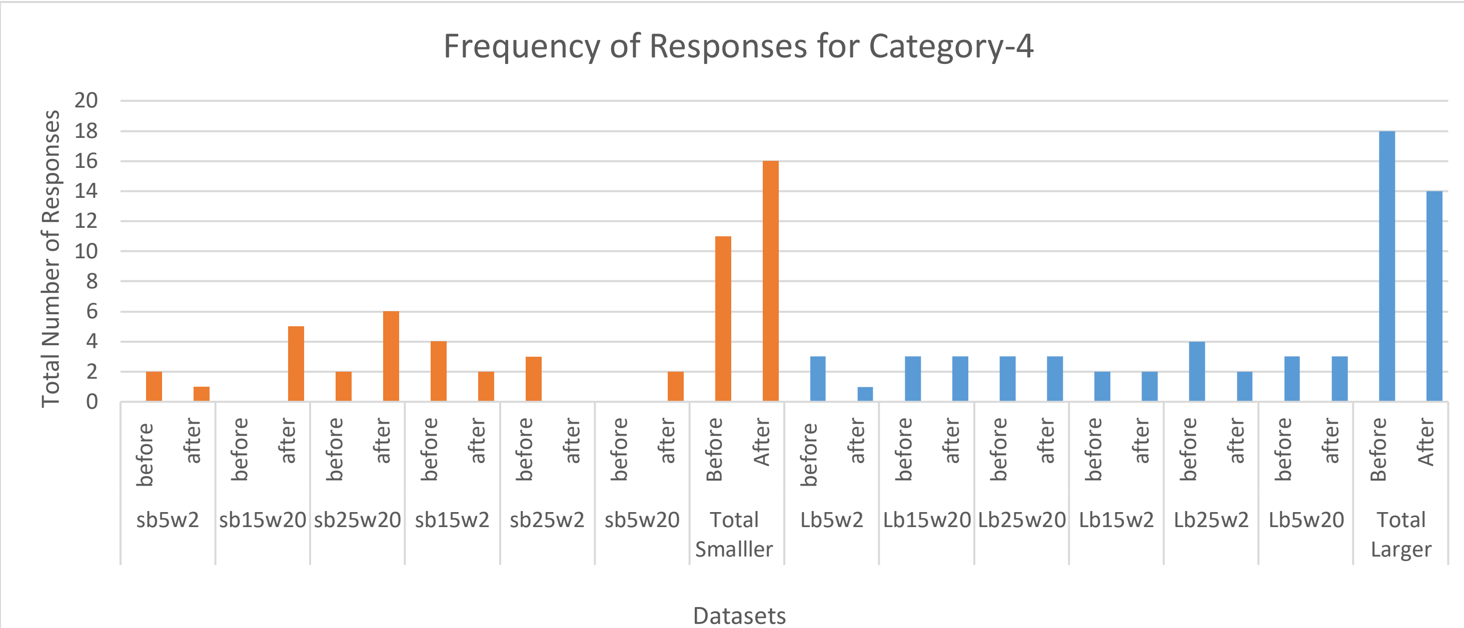


Figure 18. The frequency of responses for Category-4 across small and larger datasets.

As seen in Figure 18, the total number of responses in this category was 27 for smaller datasets (i.e., 11 before the intervention and 16 after the intervention) and 32 for the larger datasets (i.e., 18 before the intervention and 14 after the intervention). Further, the total number of responses belonging to this category in the smaller datasets were relatively higher for datasets sb15w20 and sb25w20. Even though it seems mostly datasets with high within group variation received larger number of responses in this category, especially smaller datasets such as Sb15w20 and Sb25w20, datasets with smaller within group variation also had responses in this category which indicated that participants identified sources of within group variation for both datasets with higher within group variation as well as smaller within group variation. Specifically, consistent with normative understanding of variation, participants noted higher within group variation as an indication of inconsistency when the within group variation is higher and noted smaller within group variation as an indication of consistency when the within group variation is smaller.

For example, while providing explanation for his/her rating for the dataset, sb15w20 which has a high within group variation, participant #2 explained “There is too much variation in the 32-panel to say the hypothesis was supported”. Similarly, participant #7 explained the reason for a low rating for the inconsistent results from this experiment as “Because the numbers vary a lot”. While explaining the reasoning for the same dataset, participant #4 indicated “The dataset is displaying *mixed results* and shows no significant difference nor relationship with number of goals and number of panels on the soccer ball” (emphasis added). It seems the high within variability in the dataset reduced the amount of confidence in supporting the hypothesis. In a similar vein, participant #10 also highlighted the high within group variability for his/her relatively low rating for the datasets sb5w20 as s/he noted “Score are inconsistent and very few trials”. Again, a small number of participants also referred to datasets with small within group variation such as sb15w2 in this category. For example, participant #17 noted “Consistent results, but not enough” for the sb15w2 dataset while participant #16 wrote “Somewhat consistent, still only three trials” for the same dataset.

As for the larger datasets, participants’ reasoning about the datasets with higher within group variation remained consistent with those used for smaller datasets. Individuals documented the within group variation in their responses when this variation is large enough in terms of visibility. For example, one participant stated for the same dataset that “There is too much variation in the 32-panel to say the hypothesis was supported”. As an example of how high variability within datasets affected participants’ confidence level with datasets, participant #15’s response for the dataset Lb15w20 is as follows: “The numbers are so varied it's hard to make a good conclusion from the information”. Some of the participants’ reasoning strategies for the datasets with high within group variability and low between groups variability remained consistent after the intervention as well. For instance, after the intervention, participant #4 wrote for the larger dataset Lb5w20 "This dataset shows mixed results demonstrating no significant difference nor relationship between the Bird diet and the Tomkins diet".

Overall, it could be concluded that participants recognized within group variation when the variation is higher than smaller regardless of the size of datasets and time. However, participants' understanding of within group variation is not complete even after the intervention. That is, these participants failed to consider the magnitude of the within group variation relative to the magnitude of the between group variation. Let's have a closer look at the participant #13's response quoted above where s/he explained the rating for the dataset Lb15w20 and wrote "This was clearly a windy day for testing but the superior ball still took home the gold. #14panelgreatness". While this particular participant noticed, some random factors causing the variation within the group, s/he still favored the 14-panel ball in the response even though this dataset provides very weak evidence compared to other datasets to support the idea that 14-panel balls are actually better than 32-panels.

Category 5: Developing consideration of variation (n=243)**.** Although responses in this category used both within-group and between-group variation, they did not use the within- and between-groups variation in a relative manner to support inferences (e.g., participants do not compare the magnitude of the within and between group variation. They use these two variations as two independent entities).

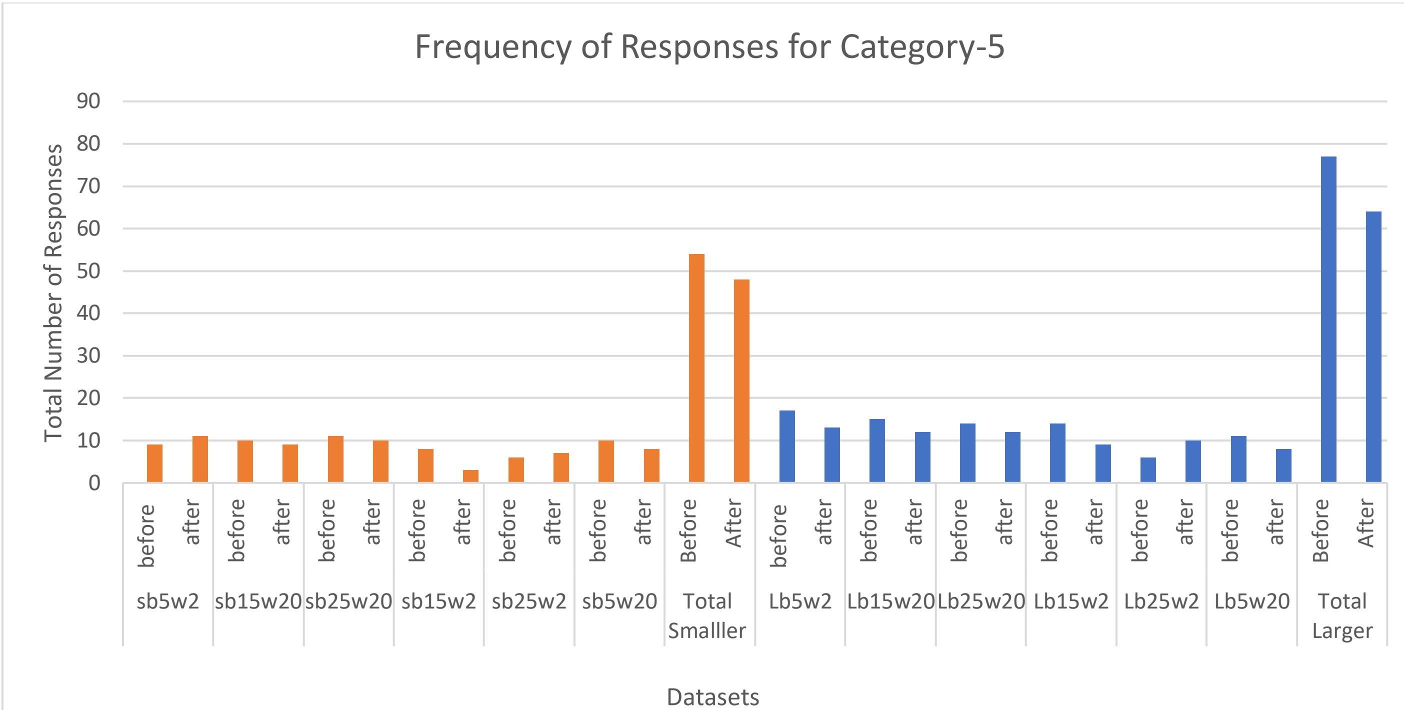


Figure 19. The frequency of responses for Category-5 across small and larger datasets.

As seen in Figure 19, the total number of responses in this category was highest compared to the total number of responses in the other five categories. While the total number of responses for small datasets was 54 before the intervention and 48 after the intervention, larger datasets received 77 responses before the intervention and 64 after the intervention. On average, smaller and larger datasets received similar amount of responses before and after the intervention in this category. The responses of participants in this category focused on between- and within group variation in a counter-normative fashion in their responses. To be specific, these participants' responses were mainly influenced by the consistency of pairwise differences between two data points, one from each group. It is worth to note that the consideration of pairwise differences is based on a misconception about data being meaningfully paired. Explicitly, these responses focused on the consistency of pairwise comparisons of individual data points where the pairs are scores aligned in rows. Indeed, there was no real basis for pairing data points in these datasets – the order of data points in independent groups designs, as were the hypothetical scenarios described in the iANOVA task used in this study, is completely random. A sample rating task along with dataset from the iANOVA task below will help to illustrate what it means to focus on the consistency of pairwise comparisons.

As seen in the sample rating task below, the iANOVA tasks highlighted the fact that balls used for each type of panel were independent and kicked in random order by stating "…*three 32-panel soccer balls*…*three 14-panel soccer balls kicked in random order*…". Also, the cover stories presented in the iANOVA tasks did not indicate that the groups were paired or dependent where balls used in the experiment were identical. The annotated notes on the example (indicated in blue and red) shows how participants reason and make pairwise calculations. Based on pairwise comparison, 14 panel balls were consistently higher than 32 panel balls.

Experiment #1 -											
Scores obtained by three 32-panel soccer balls and three 14-panel soccer balls kicked in random order by the kick-robot:											
32-panel		*14-panel*	**Pairwise Difference**								
298		305		**-7**							
302		303		**-1**							
300		307		**-7**							
Weak								Strong			
1	2	3	4	5	6	7	8	9	10		
Why?											

Thus, as in this example, the order of data points in independent groups designs is completely random. However, most participants in this group had a misconception related to the design of the study rather than being related to variation itself. Further, they calculated the difference between the two data points on each row and considered the consistency of these differences while judging the amount of evidence provided by these datasets. As a result, these participants used the between- and within group variation in a counter-normative fashion in their responses. That is, participants tended to focus on pairwise differences caused by within group variations in the datasets in a hypothesis confirming manner. For example, participant #23 explained "In four of the ten sets of scores, the 32-panel ball was equal to, or greater than the 14-panel ball. The differences are not consistent. I would have a low confidence with this data". This participant had an identical reasoning strategy for one of the larger datasets in the iANOVA task. The term 'consistency' that the participant refers to in his/her response is different than the consistency of scores within a group. The participant refers to the consistency of the pairwise differences. If interpreted correctly, one should have a low confidence level towards datasets with high within group variability while the participant had a low confidence due to inconsistency in the magnitude of pairwise differences. Similarly, the following responses before and after the intervention, to the different datasets clearly show that participants tend to make pairwise comparisons regardless of the size of datasets.

"There aren't very enough trials and there scores are varied, sometimes there is large difference and other times the scores a very close"

"Again, there are inconsistencies with this trial as sometimes the 32 panel ball does better than the 14 panel ball".

"This test has some inconsistencies as sometimes the 32 panel ball had better results than the 14 panel ball"

"The 32-panel trial one is smaller than the 14-panel. The second trial 32 panel is larger than the 14 panel, and the third trial is smaller than the 14 panel. The results are not all completely supportive"

"All but two Bird Diet triglyceride level were lower than the Tomkins diet counterpart"

Overall, it could be concluded that the responses presented in this category were considering both between- and within group variation yet the majority was looking for evidence by making pairwise comparisons of data points to verify the hypothesis that one type is better than the other one. A related discussion will follow in the discussion section.

Category 6: Strong consideration of variation (n=13). In this category, participants consider the magnitude of the within group variation in relation to between group variation. This type of understanding is deemed as strong consideration of variation.

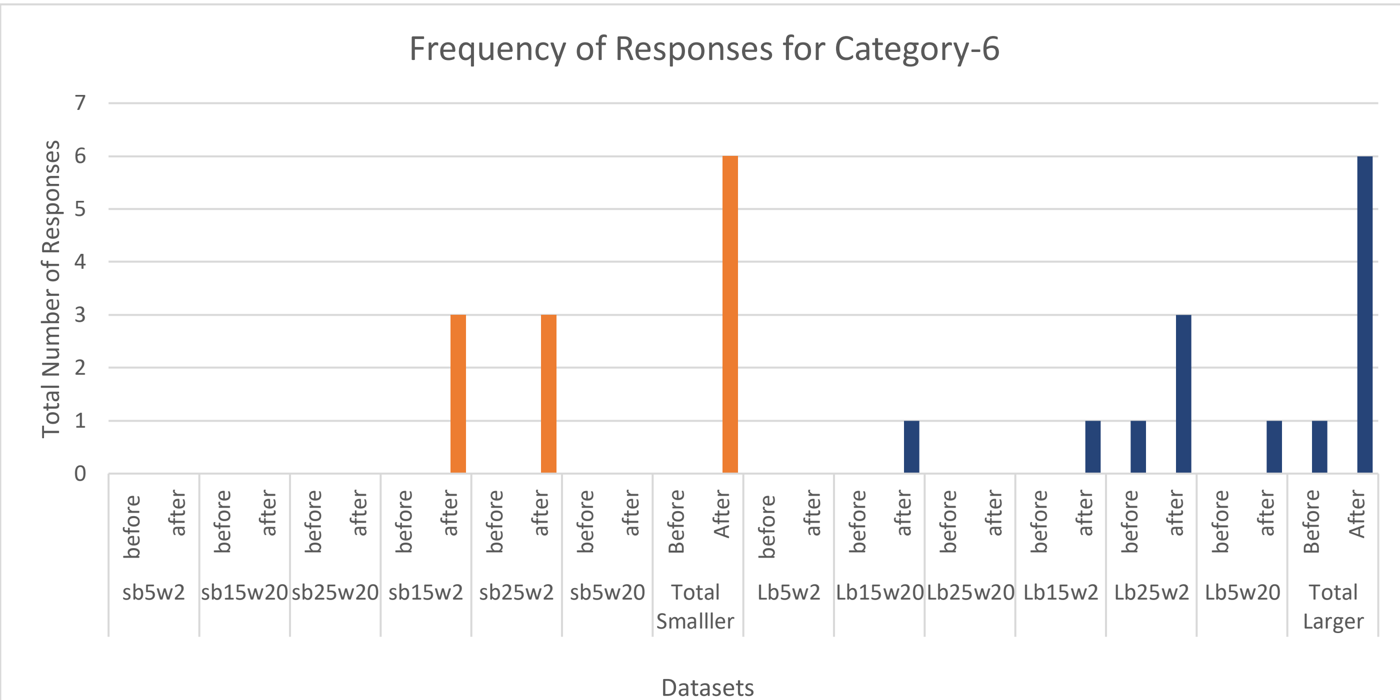


Figure 20. The frequency of responses for Category-6 across small and larger datasets.

While there was not any response considered as strong consideration of variation before the intervention, there were six responses recorded in this category for smaller datasets after the intervention. Further, the number of responses for larger datasets was one before the intervention and six after the intervention. These results are consistent with those obtained by Trumpower (2015) and Reid and Readings (2008) as these authors also indicated that the number of the participants registered as having strong consideration of variation are usually low or there is none compared to other categories.

Participants' responses revealing this strategy indicated that these participants hold a more complete understanding related to variation while judging the strength of evidence provided by two groups of data. For example, while examining the smaller dataset sb15w2, one participant noted "The variation between and within the groups is consistent and significant". This participant not only recognized the variation between and within the groups, s/he also related the small within group variation to the consistency of these groups. Similarly, while working with the larger dataset Lb15w20, another participant remarked "On the average, the Bird diet has an average of 200 mg/dl, and the Tomkins diet has an average of 215 mg/dl. The problem is that I am seeing the numbers vary large from the average of each group. I would not have any confidence that the Bird diet reduces the triglyceride more than the Tomkins diet". As consistent with a more complete understanding of variation, this participant highlighted the relatively higher within- than between-group variation as a reason for his/her lower rating for the dataset. In a similar manner, another participant noted for the dataset Lb5w20 "No consistent results on both samples, data collected fall within similar ranges". Recognizing the importance of the magnitude of the difference between group means as well as the consistency of the difference within each group, another participant noted for the larger dataset Lb25w2 that "Both diets yielded consistent results and with this big sample and observable difference there is more evidence to indicate effectivity of diets".

Overall, although only a small number of responses could be classified into this category, these results indicate that at least some participants were able to show an improved understanding of variation. Twelve responses out of thirteen in this category were recorded *after* the intervention. The follow up analysis in this category showed that these thirteen responses were provided by four individuals.

QUALITATIVE DISCUSSION

The aim of the qualitative analysis was to observe changes in responses before and after intervention as well as to get a deeper insight into some of the reasons behind students' difficulty in Informal Inferential Reasoning. Over the past 10 years, the Consideration of Variation Hierarchy by Reid and Readings (2008) has been useful in understanding the key features of learners' cognitive development related to variation in statistics education. It has also provided a robust framework for understanding the various levels of understating related to variation in statistics. However, responses of the participants in this study required a more nuanced classification scheme with respect to the levels of consideration of variation than those proposed in Reid and Readings' Hierarchy. The study by Reid and Readings and following studies using the same hierarchy typically used only one type of dataset (i.e., small or larger dataset) in their research studies. Further, the categories related to understanding of variation presented in these studies are very broad while the qualitative results of the present study provide more details about participants' reasoning with respect to variation not only because participants were provided with different sizes of datasets, but also with various scenarios regarding variation (i.e., different ratios of between groups- to within group-variation). Therefore, an in-depth examination of responses in the present study revealed that responses would be classified using an extended coding schema with six categories. Discussions related to each category will be presented in order below except for category-1 which will be revisited later in this chapter.

Category-2 represented responses which focused on only between groups variation from a holistic perspective. Responses in this category were most prevalent when this type of variation was salient through manipulations in datasets such as sb5w2, sb15w2, and sb25w2, Lb5w2, Lb15w2, and Lb25w2 (i.e., in datasets with smaller within-group variation). The mean difference between datasets was the only source of variation identified in the responses of this group; they did not mention the within group variation for each dataset (i.e., spread of dataset or standard deviation) even though it is as important as the difference between these groups from a normative perspective when comparing two groups of data.

Further, responses in category-3 focused on “Only between groups variation from an individual data point view”. In general, responses using this strategy for their reasoning were selectively picking up a higher score in one sample and a lower score in the other to confirm their biases that the independent factor (e.g., the number of panels on a ball) actually had an effect on the dependent factor (e.g., the flight characteristics of balls). This result was consistent with those obtained by Trumpower (2015). As part of the misconception related to evidence provided by within and between group differences, Trumpower (2015) also found that students using this strategy can be misled into thinking that datasets with larger within group variation provide stronger evidence against the null hypothesis by selectively attaining high scores in one sample and low scores in another one. He further noted that this may have been due to lack of an understanding related to the fact that within group variation (exceptionally high/low scores within a group) is the result of other random, uncontrolled factors rather than being a result of the independent factor. Although, a slightly different categorical system was used in the present study, responses in categories-2 and -3 would correspond to the weak consideration of variation in the Reid and Readings (2008) Consideration of Variation framework, as responses in both categories are based on only one type of variation, namely between groups variation.

As for the responses in category-4 where the emphasis was “Only on the within group variation”, the total number of responses was relatively small compared to the total number of responses in other categories. Although results obtained by Kramer, Caitlin, and Telferb (2017) found that students were able to identify within group difference correctly when the salience of the variability information is increased through manipulation, responses in this category were able to report the within group variation of a dataset where both within group variability was larger and smaller regardless of the magnitude of the between group variation. Specifically, the responses reported the decreased amount of confidence in supporting the hypothesis when the within group difference is large and marked increased amount of confidence when the within group difference is smaller. However, these responses failed to consider how the magnitude of differences between groups relative to this within group variability would affect their confidence. Thus, responses in this category would also link to the weak understanding of variation considering the Consideration of Variation Hierarchy by Reid and Readings (2008).

Considering the responses classified in categories-2, -3, and -4, the total number of responses influenced by only between group variation (category-2 plus category-3) were higher than those influenced by only within group variation in category-4. Overall, judgements of the strength of evidence provided by hypothetical datasets were more likely to be influenced by the between group variation than within group variation. This result was consistent with the findings reported by earlier studies. That is, these studies found that, overall, between groups variability had a larger effect on confidence judgements than did within-group variability (see Obrecht et al. 2007; 2010).

Across all six categories, responses most frequently fell into category-5, “Developing consideration of variation”. The responses in category-5 accounted for 37.5 percent of all responses. Although responses in this category were able to identify both between groups- and within group variation, they were not able to consider one type of the variation relative to the other type as required to demonstrate a more complete conceptual understanding of ANOVA. Further examination of responses in this category revealed two different forms: consideration of the difference between group means and within group variation as independent sources of evidence and consideration of the variability of differences between paired data points from each group. The first type of responses in this category was less frequent than the second type. The tendency to correlate the pairs of data from groups may have resulted from a misconception that the data were obtained in a within-subjects design, rather than an independent-groups design as was the case for the hypothetical datasets presented in the iANOVA task used in this study. Most responses were focusing on the consistency of differences between paired data points one from each row while comparing the two groups of data. However, this misconception was related to the design of the study rather than being related to the variation itself since participants using this strategy are actually considering both within and between group variability – they are looking at differences between groups in each pair, but they are also looking at the consistency of these differences which is at least partially due to random factors like within group variability. The tendency towards pairwise comparison is persistent across small and larger datasets. These results are consistent with the findings by Trumpower (2013a; 2013b). As Trumpower noticed, participants often performed pairwise comparisons of individual data points in the iANOVA task, which he believed tended to obscure within group variation. Thus, he devised a ‘stacked’ presentation format that would discourage such pairwise comparisons. The stacked presentation of datasets involves presentation of data columns on the top of each other rather than presenting beside each other as in un-stacked presentations. Participants, nonetheless, displayed the same non-normative performance in both versions of the iANOVA task. One might think that students had such a tendency towards pairwise comparison because they were only presented with raw data and did not have anything else to use in both iANOVA tasks. However, previous studies have shown that the types of data presentations such as only graphical presentations and presentation of summary statistics and/or raw data did not have a significant effect on students' reasoning in similar data comparison cases (Atas & Trumpower, 2015; Kramer, Caitlin, Telfer, 2017; Obrecht et al. 2007).

Further examination of responses in category-5 demonstrated that participants' understanding of consistency of results (i.e., standard deviation or spread of a dataset) contradicts the normative standards. While referring to consistency, the majority of students refer to consistency of the difference score between the individual data pairs from each sample rather than referring to the consistency of the data points within a group. If each data point within a group is larger than the one in the counterpart, the dataset is being considered as providing reliable and/or consistent results regardless of the spread within each group. However, in a formal sense, consistency of results depends on the average amount of deviation within a group rather than depending on the magnitude of the difference score between individual data pairs between two groups.

Responses under "Strong Consideration of Variation" fall into category-6. Overall, some of the students were able to integrate both between groups- and within group-variation successfully after the intervention as reported in category-6. In their written justifications for their judgements, thirteen responses displayed indications that they were able to consider the magnitude of between group variation in relation to the magnitude of within group variation after the intervention while it was only one response before the intervention. This result is consistent with the findings reported by Reid and Readings (2008). These authors also noted that usually only a small number of participants have been reported to show strong consideration of variation in a research study.

Although, Obrecht, Chapman, and Gelman (2007) found that participants gave little consideration to either the sample size or the standard deviation when presented with two lists of raw data, responses in Category 1 of "Unable to determine/Small Sample Size" was the second most common type of response in this study. As indicated in the results section, ninety percent of the responses in this category related to the indication of small sample size. Indeed, this category helped the researcher of the present study to understand the reasons behind misconceptions related to understanding of ANOVA. Specifically, the qualitative results of the presented study present a unique finding related to factors that might cause the overall misconception related to reasoning of smaller datasets. Overall, consistent with the normative logic of ANOVA, the participants of this study were aware of the power of sample size that could influence the strength of evidence provided by small size datasets. Further, it seems the effect of sample size was more visible for participants while reasoning about small sample size datasets than larger size datasets. Earlier studies (e.g., Trumpower, 2013; 2015) also reported lack of understanding related to students' reasoning abilities when it comes to providing justification for smaller datasets. Thus, faulty reasoning strategies reported in these earlier studies might be partially due to participants' lack of confidence in smaller datasets while judging the strength of evidence against the null hypothesis.

GENERAL DISCUSSION

This study not only presented research that explored the development of undergraduate students' consideration of variation while they engaged in intuitive ANOVA tasks but also investigated the effect of formative feedback that focuses on conceptual understanding on improving students' Informal Inferential Reasoning. An informal ANOVA task was used to encourage statistics learners to engage in inferential reasoning ideas without any need for formal calculations. That said, the Informal Inferential Reasoning task used in this study can be considered as a precursor to a more formal Analysis of Variance procedure.

Both quantitative and qualitative results from the present study indicated that statistics learners demonstrate diverse patterns of inferential reasoning when they are provided with different size of datasets with varying amount of variability. Even though previous studies noted that students were often consistent with their reasoning strategies regarding variation across datasets, this might be due to the fact that participants in these studies were each provided with only one size of dataset (see, Reid & Readings, 2008: Trumpower, 2013; 2015). However, both quantitative and qualitative results from the current study revealed that participants employed multiple strategies for different datasets depending on the manipulated variability within and between these datasets. The existing framework by Reid and Readings (2008) for classification of responses with respect to consideration of variation was found to have limitations. Their categories are broad and some of the participants' responses in the present study did not fit into their categories. This study has shown that at least six different types of consideration of variation exist to characterize cognitive development of the variation. The proposed new categories related to consideration of variation provide more detailed information about the cognitive structure of reasoning about variation. Participants' responses in five out of six categories reported in the present study showed non-normative patterns compared to the normative logic of ANOVA, failing to realize the importance of both between group variation and within group variation in an integrated manner.

In addition, details revealed in the consideration of variation categories shed some light on the persistent misconceptions that participants hold regarding the concept of variability. That is, while comparing two groups of data as part of inferential reasoning practices, leaners attempt to selectively attain pairwise differences between data points from two groups of data in a manner confirming the hypothesis that the independent variable has a real effect on the dependent variable. Specifically, participants selectively attend to pairwise differences when there is larger within-group variation which allows them to selectively choose pairs of data points that result in large differences to confirm their hypothesis about an effect of the independent variable. Those who look at pairwise differences often seem to be considering the consistency of the pairwise differences; thus, they are considering both between and within group variation to a certain extent. However, those who selectively choose a large number in one column and a small number in the other column are only considering between-group variation. Overall, it seems these misconceptions may be the result of strongly held misunderstandings related to the design of the study from where the data originates. Therefore, inclusion of a participant-designed data collection project in an introductory statistics course might help to overcome the strong misconception that participants have about data being from a within-subjects design which result in a tendency to examine pairwise differences despite no logical basis for pairing scores. Supporting this argument, Bradstreet (1996) proposes engaging students in data collection processes for better outcomes in statistics education. Similarly, Moore (1998), claims that students need to experience firsthand the process of data collection and data exploration in order to deepen their conceptual understanding, and Garfield and Ben-Zvi (2007) suggest that an effective learning process in statistics requires active involvement of students in constructing knowledge.

Findings from the present study explored whether formative feedback helped participants to improve their conceptual understanding related to ANOVA concepts such as *between group difference, within group difference*, *sample size*, *strength of evidence*, *random factors*, *independent* and *dependent variable*. This was evident in the quantitative data that participants' knowledge maps related to these concepts were closer to the referent map after the intervention. In addition, the improvement was also marked in the qualitative data that a few of the statistics learners demonstrated signs for strong consideration of variation which is considered as an indication of complete understanding of ANOVA. However, this improvement in conceptual knowledge was not translated into Informal Inferential Reasoning as there was no statistically significant improvement on participants' performance on the iANOVA tasks. A possible reason behind the difficulty might relate to the nature of feedback given to participants. Although the feedback in this study was designed using the principals of formative feedback, it involved a low level of student engagement as supposed to a high level of engagement such as reflection and problem solving activities consistent with the nature of formative feedback. As a result, participants also did not have an opportunity to reflect on the feedback or to explain the concepts in their own words as part of a real problem-solving practice due to the pure online nature of this study. Moreover, the feedback was not providing direct information to the students instead it was only indirectly addressing the Informal Inferential Reasoning task which was expected to lead to a more formal understanding of ANOVA procedures. From a teaching standpoint, having participants explaining the phenomenon rather than calculating things give them more insight into meaningful understanding, as Sarwar (2011) noted in his study of physics learners. As for future consideration in statistics education, any feedback that involves actual formal calculations related to statistical procedures as well as highlights the logic behind these procedures may help statistics learners better integrate conceptual and procedural knowledge which can result in a more complete understanding of the phenomenon being studied. Also, consistently with the nature of formative feedback principles, the feedback combining both conceptual and procedural knowledge should be persistent throughout the semester as one-time feedback may not be sufficient enough to get desired outcomes especially with the complex ideas such as variance in statistics education.

LIMITATIONS

As in any other research study, this study is also limited by the quality and the context of the research. Although robust actions were taken to strengthen the quality of data collection and analysis process as well as to minimize concerns about the validity, the study presents some limitations derived from its own contextual nature. One limitation of the study relates to the design of the experiment where there was lack of opportunity for participants to reflect on the feedback and to practice the different types of feedback with real world problems after the intervention. Even though participants began to comprehend the critical concepts highlighted in the feedback as measured by the difference in PFsim scores before and after the intervention, it should be marked that these participants did not have much time to reflect on the feedback and did not apply the feedback with more real-world problems due to time constraints of the study. The total time required to complete the study was, on average, 2.5 hours including completing the iANOVA and Concept Mapping tasks twice and going through written and visual types of feedback. As such, participants went through individualized written and visual feedback without having to explicitly demonstrate their conceptual understanding of it, nor having to apply it. Therefore, the intervention in the study may have missed the requisite depth of engagement/reflection. Further, considering ANOVA as one of the more complex topics in basic statistics education, the difficulty related to grasping ideas behind ANOVA and applying these ideas to practical problems might lie in the degree of abstractness that ANOVA concepts involves. Consequently, although participants seemed to improve conceptual understanding somewhat, this was not enough to translate into Informal Inferential Reasoning performance in the study.

Also, although the online nature of the study was more efficient than a paper based approach in assessing participants' structural knowledge and providing immediate feedback to a large number of students in a short time period, it may have imposed some limitations as well. For instance, the lack of familiarity with the use of online assessment and multi-media tools might have hindered the benefits of proposed instructional materials in the study. Further, the flow of the experimental study was designed in a way to make sure that participants spent a certain amount of time with each feedback before they proceeded with the next step. But, there was no way to ensure that participants were actually paying attention to the feedback being displayed on the computer screen. So, it is possible that these disadvantages reduced any positive effect of instructional materials within the study. Further, although the sample size in this study would impose a potential limitation, it should be noted that similar results were replicated several times in other empirical studies even with smaller sample sizes. Therefore, it can be concluded that results reported in this study are likely representative of a more general phenomenon. In addition, the relatively small sample size in this study allowed me to conduct an in-depth exploration of the qualitative data. Also, despite the fact that the participants of the study were mostly female there were no statistical differences on key measures with respect to self-identified gender except for the level of statistics anxiety reported by female participants. Overall, women reported a statistically significant higher level of statistics anxiety than men. However, the reported level of anxiety by women were at the mid-point on the scale. Therefore, considering the mid-point level of statistics anxiety, this result was not expected to limit participants' ability to engage in statistical activities presented in the study. Further, considering the profile of participants in the study, one might question the generalizability of these findings to university programs in which the majority of students are typically male – e.g., physics or mathematics and statistics. But, the implications of this study are primarily intended for helping develop statistical reasoning in students who do not already possess solid IIR, as might be expected of students in such programs. The results can likely be generalized, however, to students from introductory statistics classrooms taught in many other fields, such as education or the social sciences.

IMPLICATIONS FOR FUTURE RESEARCH

Attempts to increase students' conceptual understanding and Informal Inferential Reasoning on such a complicated statistics topic like ANOVA have several implications for future studies. The complex nature of the topic selected may require extensive practice and reflection activities to help students transform increased conceptual understanding into Informal Inferential Reasoning and perhaps subsequently transfer to procedural performance. Further, considering students' fear towards statistics due to its perceived difficulty, a combination of face-to-face and online instructional activities might help students better in acquiring a deeper understanding of critical concepts in statistics rather than relying on a mere online instructional design. For example, an initial use of an online tool such as the Concept Mapping tool used in the study might assist in measuring students' initial knowledge structures and providing immediate multi-media feedback in a timely manner, and thus helps saving time and efforts which can allow more time for reflection/discussion activities in a classroom environment. Later, as suggested in the GAISE (2016) report, ANOVA activities and interventions that encourage students to collect their own data and to use technology to explore concepts and analyze data with concrete examples may benefit a deeper understanding of the phenomenon. Overall, it may have been possible to get more positive results with a lengthier experimental design with a hybrid approach (i.e., combination of online and face-to-face activities).

As another implication of the study, the employment of Pathfinder Networks (PFnets) as a scaling technique embedded in the online Concept Mapping Tool benefited the following stages of the study in an effective and efficient manner:

1. Elicitation of a large number of students' and referent knowledge networks,

2. Assessment of students' knowledge structures by comparing with the referent network and identification of areas for improvement (i.e., determining common, missing, and extraneous links between student and referent network),
3. Provision of individualized multi-media feedback to a large number of students in a short time,
4. Tracking conceptual knowledge at various points during the study (i.e., as measured during pre-, post-, and follow-up tests).

Overall, completing all four main steps described above would have been very difficult without the online Concept Mapping Tool. That said, the present methodology aids in overcoming the challenges related to the required amount of time and work to evaluate students' knowledge structures and to provide feedback based on the evaluation reported in previous studies (e.g., Sarwar, 2011). Therefore, instructors of larger classrooms may consider this methodology as an alternative technique to assess students' knowledge structures and provide feedback.

The present study also extends the literature on the Consideration of Variation Hierarchy by Reid and Readings (2008). The hierarchy by these authors suggests classification of students' responses related to the degree of consideration of variation in statistical terms in three main levels: weak, developing and strong consideration of variation. However, this study revealed that categories suggested in the literature may be too general as some of the participants' responses did not fit into those categories. This might partially be because the present study, differently than previous similar studies, employed different sizes of datasets with varying within/between group variations to get a better understating of students reasoning process regarding variation. Thus, a deeper examination of students' responses has shown that six different types of consideration of variation exist to characterize cognitive development of understanding variation. As a result, the recommended new six categories related to consideration of variation provide more detailed information about the cognitive structure of reasoning about variation. So, any researcher who is interested in studying the cognitive development of students' reasoning regarding the consideration of variation may consider using the new six-category consideration of variation hierarchy to obtain a more detailed understanding of the developmental stages. Further related to this point, both quantitative and qualitative results from the present study indicated that statistics learners demonstrate diverse patterns of inferential reasoning when they are provided with different size of datasets with varying amount of variability. Even though previous studies noted that students were consistent with their reasoning strategies regarding variation across datasets (see, Reid & Readings, 2008: Trumpower, 2013; 2015), both quantitative and qualitative results from the current study revealed that participants employed multiple strategies for different datasets depending on the manipulated variability within and between these datasets.

REFERENCES

Acton, W. H., Johnson, P. J., & Goldsmith, T. E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology*, 86(2), 303.

Allen, R., Folklar, A., Lancaster, G., & Abram, B. (2010). Statistics for the biological and environmental sciences: Improving service teaching for postgraduates, In C. Reading (Eds.), *Data and context in statistics education: Towards an evidence-based society*. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia. Voorburg.

Atas, S., Trumpower, D., Filiz, M. (2013). How a computer-based concept mapping can be used in statistics education: A case of Analysis of Variance (ANOVA), *United States Conference on Teaching Statistics* (USCOTS) - North Carolina, USA.

Ausubel, D. P. (1968). *The Psychology of Meaningful Learning; an Introduction to School Learning*. Grune and Stratton.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1978). *Educational psychology: A cognitive view*.

Azzarello, R. N. (2007). Use of the pathfinder scaling algorithm to measure students' structural knowledge of community health nursing. *Journal of Nursing Education*, *46*(*7*), 313.

Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education*, *38*(2), 115-131.

Ben-Zvi, D., & Garfield, J. (2005). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (p. 3-16). Springer.

Boldt, M. (2001). Assessing Students' Accounting Knowledge: A Structural Approach. *Journal of Education for Business,* 76(5). doi:10.1080/08832320109599646

Bradstreet, T. E. (1996), "Teaching Introductory Statistics Courses So That Nonstatisticians Experience Statistical Reasoning", The American Statistician, 50, 69–78.

Brookhart, S. M. (2007). Feedback that fits. *Educational Leadership*, *65*(4), 54-59.

Carless, D. (2006). Differing perceptions in the feedback process. Studies in Higher Education, 31(2), 219-233. doi:10.1080/03075070600572132.

Casas-García, L. M., & Luengo-González, R. (2013). The study of the pupil's cognitive structure: the concept of angle. *European journal of psychology of education*, *28*(2), 373-398.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Cohen, V. B. (1985). A reexamination of feedback in computer-based instruction: Implications for instructional design. Educational Technology, 25(1), 33–37.

Coladarci, T., Cobb, D. C., Minium, E. V., Clarke, R. B. (2011). *Fundamentals of Statistical Reasoning in Education,* John Wiley & Sons.

Cooke, N. J. (1992). Predicting judgment time from measures of psychological proximity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 640.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage publications.

Curtis, M. B. and.M.A. Davis (2003). "Assessing knowledge structure in accounting education: an application of Pathfinder Associative Networks." *Journal of Accounting Education, 21*, 185-195.

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. London: Sage.

Davis, M. A., Curtis, M. B., & Tschetter, J. D. (2003). Evaluating cognitive training outcomes: Validity and utility of structural knowledge assessment. *Journal of Business and Psychology*, *18*(2), 191-206.

De Jong, T., & Ferguson-Hessler, M. G. (1996). Types and qualities of knowledge. *Educational psychologist*, *31*(2), 105-113.

DelMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. Statistics Education Research Journal, 6(2), 28-58. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf

Eggen, P., & Kauchak, D. (2007). *Group and individual differences. Educational Psychology: Windows on Classrooms*, 103-116.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Filiz, M., Trumpower, D. L., & Atas, S. (2012). Analysis of how well a concept mapping website conforms to principles of effective assessment for learning. *In Concept Maps: Theory, Methodology, Technology-Proceedings of the 5th International Conference on Concept Mapping*.

Filiz, M., Trumpower, D., & Atas, S. (2013). The Contributions of Digital Concept Maps to Assessment for Learning Practices. *International Association for Development of the Information Society*.

Filiz, M., Trumpower, D., & Vanapalli, A. (2014). Exploring the mobile structural assessment tool: Concept maps for learning website. *Revista Colombiana de Estadística*, 37(2), 297-317.

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. Journal of Statistics Education, 2(2).

Garfield, J. (1995). How students learn statistics. *International Statistical Review*, 63, 25-34.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19(1)*, 44-63.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75(3)*, 372-396.

Garfield, J. B. & Ben-Zvi, D. (2008). *Developing students statistical reasoning: Connecting research and teaching practice.* Springer Publisher.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning, *Learning and Teaching in Higher Education, (1),* 1-31.

Grant, K. A. (2007). Tacit knowledge revisited–we can still learn from Polanyi. *The Electronic Journal of Knowledge Management*, *5*(2), 173-180.

Goldsmith, T. E., & Davenport, D. M. (1990). Assessing structural similarity of graphs. In R. W. Schvaneveldt (Ed.), *Ablex series in computational sciences. Pathfinder associative networks: Studies in knowledge organization* (pp. 75-87). Westport, CT, US: Ablex Publishing.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of educational psychology*, *83*(1), 88.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. doi:10.3102/003465430298487.

Hay, D. B., Tan, P. L., & Whaites, E. (2010). Non-traditional learners in higher education: comparison of a traditional MCQ examination with concept mapping to assess learning in a dental radiological science course. *Assessment & Evaluation in Higher Education*, *35*(5), 577-595.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. Conceptual and procedural knowledge: The case of mathematics, 2, 1-27.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. *Second handbook of research on mathematics teaching and learning*, *1*, 371-404.

Hiebert, J. & Lefevre P. (2013). Conceptual and procedural knowledge in mathematics: An introductory analysis. In Hiebert, J. (Ed.). (2013). *Conceptual and procedural knowledge: The case of mathematics*. Routledge.

Jonassen, D. H., Beissner, K., & Yacci, M. (2013). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Routledge.

Kelly, A. E., Sloane, F., & Whittaker, A. (1997). Simple approaches to assessing underlying understanding of statistical concepts. *The assessment challenge in statistics education*, 85-90.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78(2), 311.

Kramer, R. S. S., Telfer, C. G. R., Towle, A. (2017). Visual Comparison of Two Data Sets: Do People Use the Means and the Variability, *Journal of Numerical Cognition, 3 (1).*

MacGillivray, H. (2004). *Coherent and purposeful development in statistics across the education spectrum*. In G. Burrill, & M. Camden (Eds.), Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable. Voorburg, The Netherlands: International Statistical Institute.

Makar, K. & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, 8(1), 82–105.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont: Wadsworth.

McClure, J., Sonak, B., & Suen, H. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching*, 36, 475-492.

McGaghie, W. C., McCrimmon, D. R., Mitchell, G., Thompson, J. A., & Ravitch, M. M. (2000). Quantitative concept mapping in pulmonary physiology: comparison of student and faculty knowledge structures. *Advances in Physiology Education, 23*(1), S72-81.

Merriam, S. (2009). *Qualitative research: A guide to design and implementation*. San Francisco, CA: Jossey-Bass.

Meyer, D. E., & Schvaneveldt, R. W. (1976). Meaning, memory structure, and mental processes. *Science, 192*(4234), 27-33.

Michael, P. (1958). *Personal knowledge. Towards a post-critical philosophy*. With a New Foreword by Mary Jo Nye.-2015.-284 p.

Michael, P. (1966). *The tacit dimension*. Peter Smith, Gloucester, Mass.

Moore, D., & Notz, W., I. (2014). *Statistics: Concepts and controversies*, W.H. Freeman and Company, New York.

National Council for Teachers of Mathematics. (2009). *Focus in high school mathematics: Reasoning and sense making in statistics and probability*. Reston: Author.

Nash, J. G., & Nash, J. M. (2003). A structural representation of migraine diagnostic criteria: The experts' view. *Headache: The Journal of Head and Face Pain*, *43*(4), 322-329.

Neuweg, G. H., & Fothe, S. (2011). In Search of the Golden Mean: The Ambivalence of Knowledge Explication. *Electronic Journal of Knowledge Management*, 9(4).

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218. doi:10.1080/03075070600572090.

Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of research in science teaching*, *27*(10), 937-949.

Novak, J. D., & Cañas, A. J. (2008). The theory underlying concept maps and how to construct and use them. *Reflecting Education*, *3*(1), 29-42.

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning*, 16(1), 26-44.

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.

PAWTOON (2017) https://www.powtoon.com.

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures.(Trans A. Rosin)*. Viking.

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), Working Cooperatively in Statistics Education. Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute. [Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning*, *13*(1-2), 27-46.

Quillian, M. R. (1969). The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM*, *12*(8), 459-476.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal*, 5(2), 46-68.

Reber, A. S. (1993). Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious (*Oxford Psychology Series*,*19*).

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of educational psychology, 93*(2), 346.

Rossman, A. J. (2008). Reasoning about Informal Statistical Inference : One Statistician's View *Statistics Education Research Journal*, 7(2), 5-19

Rumsey, D. J. (2009). Teaching bits: Random thoughts on teaching. Journal of Statistics Education, 17(1). [Online: http://www.amstat.org/publications/jse/v17n1/rumsey.html]

Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. Journal of Research in Science Teaching, 33, 569- 600.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. Assessment in Education: Principles, Policy and Practice, 5, 77-84.

Sarwar, G. S. (2011). *Structural assessment of knowledge for misconceptions: Effectiveness of structural feedback provided by pathfinder networks in the domain of physics*. Kolln, Germany: LAP Lambert Academic Publishing.

Sarwar, G. S., & Trumpower, D. L. (2015). Effects of conceptual, procedural, and declarative reflection on students' structural knowledge in physics. *Educational Technology Research and Development*, 63(2), 185–201. Tall, D.O. & Vinner, S.

Schneider, M., Rittle-Johnson, B., & Star, J. R. (2011). Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. *Developmental Psychology, 47*(6), 1525-1538. doi:10.1037/a0024997

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization*. Ablex Publishing.

Schvaneveldt, R. W., & Durso, F. T. (1981). Generalized semantic networks. In *Bulletin of the Psychonomic Society* (Vol. 18, No. 2, pp. 76-77). Austin, TX 78704: Psychonomic Soc Inc.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In *Psychology of learning and motivation* (Vol. 24, pp. 249-284). Academic Press.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *The Second Handbook of Research on Mathematics* (pp. 957–1010). Reston, VA: National Council of Teachers of Mathematics (NCTM).

Star, J. R. (2005). Re-conceptualizing procedural knowledge. *Journal for research in mathematics education*, 404-411.

Sternberg, R. J. (Ed.). (1999). *Handbook of creativity*. Cambridge University Press.

Taricani, E. M., & Clariana, R. B. (2006). A technique for automatically scoring open-ended concept maps. *Educational Technology Research and Development*, *54*(1), 65-82.

Trumpower, D.L. (2013). Formative Use of Intuitive Analysis of Variance. *Mathematical Thinking and Learning*, 15(4), 291-313, DOI: 10.1080/10986065.2013.830952

Trumpower, D. L. (2015). Aspects of first year statistics students' reasoning when performing intuitive analysis of variance: effects of within-and between-group variability. *Educational Studies in Mathematics*, 88(1), 115-136.

Trumpower, D. L. & Atas, S., (2014). "*Probability and Statistics", Analysis of Variance*, In Barragués J.I., Morais A. & Guisasola J. (Eds.), San Sebastian, Spain, Science Publishers, Taylor & Francis Group company.

Trumpower, D.L.& Fellus, O. (2008). Naïve statistics: Intuitive analysis of variance. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 499-503).

Trumpower, D. L., & Sarwar, G. S. (2010). Effectiveness of structural feedback provided by Pathfinder networks. *Journal of Educational Computing Research*, 43, 7-24.

Trumpower, D. L., Sharara, H., & Goldsmith, T. E. (2010). Specificity of structural assessment of knowledge. *The Journal of Technology, Learning and Assessment*, *8*(5).

Usher, R. (1996). A critique of the neglected epistemologies of educational research. In D. Scott, & R. Usher (Eds.), *Understanding Educational Research* (pp. 33-52). London and New York: Routledge.

Weinberg, A., Wiesner, E., & Pfaff, T. J. (2010). Using informal and formal inferential reasoning. *Journal of Statistics Education*, *18*(2), 1-24.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, 7(2), 40–58.

APPENDIX A

Pre-test iANOVA Task

Cover Story

A soccer ball is the only essential piece of equipment in the game of soccer; it is a necessity and is typically constructed from 32 panels of material that are stitched together. Suppose two scientists/entrepreneurs believe that the *independent variable* of number of panels from which a soccer ball is constructed has an effect on the *dependent variable* of the flight characteristics of the ball. In particular, the scientists/entrepreneurs hypothesize that soccer balls made of 14 panels will have better flight characteristics than the currently popular soccer balls made of 32 panels. Thus, they are thinking about developing a new soccer ball with 14 panels instead of 32. To test their hypothesis, the scientists/entrepreneurs devise an experiment in which a robotic leg will be used to kick 32-panel and 14-panel soccer balls with the exact same force in a wind tunnel. Trajectory analysis software will be used to track the flight of the balls and will determine a score for each ball. Higher scores indicate better flight characteristics. Recognizing the existence of *random factors* that can be present in any experiment, the scientists/entrepreneurs decide to test *groups* of 32-panel and 14-panel balls (rather than testing just a single ball of each type).

Listed on the following page are potential results from several such experiments with different *samples* of soccer balls kicked with the robotic leg. For each potential result, rate the *strength of evidence* (1=weak, 10=strong) that you think the results provide to support the hypothesis that soccer balls made with 14 panels have better flight characteristics than those made with 32 panels. You do not have to actually perform any calculations or tests for this task.

Next, you will be asked to rate how related each of the following concepts are on a 1-5 scale (1=least related; 5=most related). You may recognize the concepts from the previous task that you just completed.

Dependent variable

Differences *between* groups

Differences *within* groups

Independent variable

Random factors

Size of samples

Strength of evidence

Experiment #1 - *Scores obtained by three 32-panel soccer balls and three 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
298	305
302	303
300	307

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

Why?

Experiment #2 - *Scores obtained by three 32-panel soccer balls and three 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
280	315
320	335
300	295

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

Why?

Experiment #3 - *Scores obtained by three 32-panel soccer balls and three 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
280	345
320	305
300	325

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

Why?

Experiment #4 - *Scores obtained by three 32-panel soccer balls and three 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
298	317
302	315
300	313

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

Why?

Experiment #5 - *Scores obtained by three 32-panel soccer balls and three 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
298	325
302	323
300	327

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

Why?

Experiment #6 - *Scores obtained by three 32-panel soccer balls and three 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
280	305
320	325

300	285

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

Why?

Experiment #7 - *Scores obtained by ten 32-panel soccer balls and ten 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
303	303
299	308
301	305
302	306
296	302
300	307
300	303
301	304
298	305
300	307

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

Why?

Experiment #8 - *Scores obtained by ten 32-panel soccer balls and ten 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
298	318
332	318
275	331
289	289
276	301
327	289
308	297
286	342
294	325
315	340

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

Why?

Experiment #9 - *Scores obtained by ten 32-panel soccer balls and ten 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
298	303
332	330
275	353
289	332
276	313
327	322
308	347
286	333
294	286
315	331

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

Why?

Experiment #10 - *Scores obtained by ten 32-panel soccer balls and ten 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
303	313
299	314
301	317
302	313
296	316
300	319
300	316
301	314
298	315
300	313

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

Why?

Experiment #11 - *Scores obtained by ten 32-panel soccer balls and ten 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
303	326

299	323
301	324
302	325
296	324
300	323
300	326
301	323
298	327
300	329

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

Why?

Experiment #12 - *Scores obtained by ten 32-panel soccer balls and ten 14-panel soccer balls kicked in random order by the kick-robot:*

32-panel	*14-panel*
298	283
332	315
275	283
289	286
276	342
327	298
308	291
286	325
294	311
315	316

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

Why?

Post-Test iANOVA Task

Cover story

Triglycerides are a type of fat that occurs in the bloodstream. Higher levels of triglycerides are a known risk factor for heart disease. Suppose that a physician believes that the *independent variable* of type of diet has an effect on the *dependent variable* of triglyceride levels. In particular, the physician hypothesizes that her newly developed diet, referred to as the Bird Diet, leads to lower triglyceride levels than the currently popular Tompkins Diet. Thus, the physician is thinking about publishing a book to promote her new Bird Diet. To test her hypothesis, the physician devises an experiment in which participants will be randomly assigned to follow either the Bird Diet or the Tompkins Diet for two months. Participants will be closely monitored to make sure that they are strictly following the diets. At the end of the two months, each participant's blood will be tested for triglyceride levels, as measured in mg/dL. Recognizing the existence of *random factors* that can be present in any experiment, the physician decides to test *groups* of participants who follow the Bird Diet and the Tompkins Diet (rather than testing just a single person following each type of diet).

Listed on the following page are potential results from several such experiments with different samples of participants. For each potential result, rate the *strength of evidence* (1=weak, 10=strong) that you think the results provide to support the hypothesis that type of diet has an effect on triglyceride levels. You do not have to actually perform any calculations or tests for this task.

Experiment #1 - Triglyceride levels (mg/dL) of three participants who followed the Bird Diet and three participants who followed the Tompkins Diet:

Bird Diet	*Tompkins Diet*

198	205
202	203
200	207

For this experiment, rate how strong the evidence is (1=weak, 10=strong) to support the hypothesis that type of diet has an effect on triglyceride levels.*

Follow-up Test iANOVA Task

Cover story

Corn is a source of food for many people in the world. Therefore, the ability to produce high yields of corn is important. In general, corn with more kernels per ear generates higher yields than corn with fewer kernels per ear.

Suppose that an agricultural scientist believes that the *independent variable* of type of fertilizer has an effect on the *dependent variable* of kernels per ear of corn. In particular, the scientist hypothesizes that his newly developed fertilizer, referred to as GrowMore, allows corn to produce more kernels per ear than the most popular current brand of fertilizer, BrandX. Thus, the scientist is thinking about marketing his newly developed fertilizer on a large scale. To test his hypothesis, the scientist devises an experiment in which corn plants will be randomly assigned to be treated with either GrowMore or BrandX fertilizer. At the end of the growing season, the number of kernels/ear on each corn plant will be measured. Recognizing the existence of *random factors* that can be present in any experiment, the scientist decides to test *groups* of corn plants treated with GrowMore and BrandX fertilizer (rather than testing just a single corn plant treated with each type of fertilizer).

Listed on the following page are potential results from several such experiments with different samples of corn plants. For each potential result, rate the *strength of evidence* (1=weak, 10=strong) that you think the results provide to support the hypothesis that type of fertilizer has an effect on the number of kernels/ear. You do not have to actually perform any calculations or tests for this task.

Experiment #1 – Number of kernels/ear of three corn plants treated with BrandX fertilizer and three corn plants treated with GrowMore fertilizer:

BrandX	*GrowMore*
498	505
502	503
500	507

For this experiment, rate how strong the evidence is (1=weak, 10=strong) to support the hypothesis that type of fertilizer has an effect on the number of kernels/ear.*

APPENDIX B

The University of New Mexico
Consent to Participate in Research

You are being asked to participate in a research study that is being done by Tim E. Goldsmith who is a professor from the Department of Psychology, University of New Mexico and by Sait Atas, who is a Ph.D. candidate at the University of Ottawa, Canada.

The aim of is study is to examine the effectiveness of computer-based feedback that can improve learning for statistical reasoning.

You are being asked to participate in this study because all registered students from the University of New Mexico's Psychology Department will be welcomed to participate in this research as the research question pertains to most students in the department.

This form will explain the research study, and will also explain the possible risks as well as the possible benefits to you. We encourage you to talk with your family and friends before you decide to take part in this research study. If you have any questions, please ask one of the study investigators.

What will happen if I decide to participate?

If you agree to participate, the following things will happen:

We will ask you to complete two web-based data collection tools, which can be accessed through the concept maps for learning website (i.e. www.conceptmapsforlearning.com). The URL along with a user name and password will be sent to you by email if you wish to participate in the study. Once you login by using your user name and password, you will be able to complete the data collection instruments. Then you will be provided with feedback based on your performance. After receiving the feedback, you will be asked to complete the two data collection instruments one week later as you did at the beginning of the study to help us understand the effectiveness of the feedback provided to you.

How long will I be in this study?

Participation in this study will take, on average, a total of 2.5 hours over a period of one session.

What are the risks or side effects of being in this study?

Participation in the research poses no potential harm beyond everyday life. A separate sheet will provide contact information of the Counselling Centre of the University of New Mexico should any need arise. Your participation will be anonymous and your data will be stored with a special code. There is no way to determine your identity in the study as you will not be asked for any personal individual identifiers such as name, address, or prior schools of attendance.

What are the benefits to being in this study?

There will be no benefit to you from participating in this study. However, it is hoped that information gained from this study will help identifying students' naïve abilities and biases in analysis of variance we can improve instruction. Making students aware of their accurate naïve knowledge may help to alleviate statistics anxiety. And, making instructors aware of pre-existing biases may help target instruction toward these more difficult areas.

What other choices do I have if I do not want to be in this study?

You have the option not to take part in this study. There will be no penalties involved if you choose not to take part in this study.

How will my information be kept confidential?

Your participation will be anonymous and your data will be stored with a special code. There is no way to determine your identity in the study as you will not be asked for any personal individual identifiers such as name, address, or prior schools of attendance. The data gathered form participants will be stored in the student investigator's computer and only the student investigator and thesis supervisor will have access to data. In order to ensure confidentiality of the data, completed data will be printed. The electronic data will be removed from the Internet. Printed data will be used to analyze and evaluate data. Upon project completion, data will be stored in a locked file cabinet in principal investigator's office at the University.

What are the costs of taking part in this study?

There is no cost associated with the study.

How will I know if you learn something new that may change my mind about participating?

You will be informed of any significant new findings that become available during the course of the study, such as changes in the risks or benefits resulting from participating in the research or new alternatives to participation that might change your mind about participating.

Can I stop being in the study once I begin?

Your participation in this study is completely voluntary. You have the right to choose not to participate or to withdraw your participation at any point in this study without affecting your future health care or other services to which you are entitled.

Whom can I call with questions or complaints about this study?

If you have any questions, concerns or complaints at any time about the research study, contact the PI, Tim E. Goldsmith.

If you need to contact someone after business hours or on weekends, please contact the student investigator, Sait Atas.

If you would like to speak with someone other than the research team, you may call the UNM Office of the IRB at (505) 277-2644.

Whom can I call with questions about my rights as a research participant?

If you have questions regarding your rights as a research participant, you may call the UNM Office of the IRB (OIRB) at (505) 277-2644. The IRB is a group of people from UNM and the community who provide independent oversight of safety and ethical issues related to research involving human participants. For more information, you may also access the OIRB website at http://irb.unm.edu.

CONSENT

You are making a decision whether to participate (or to have your child participate) in this study. Your signature below indicates that you/your child read the information provided (or the information was read to you/your child). By signing this consent form, you are not waiving any of your (your child's) legal rights as a research participant.

I have had an opportunity to ask questions and all questions have been answered to my satisfaction. By signing this consent form, I agree to participate (or let my child participate) in this study. A copy of this consent form will be provided to you.

____

Name of Adult Subject (print)

____ ____________________

Signature of Adult Subject Date

INVESTIGATOR SIGNATURE

I have explained the research to the participant and answered all of his/her questions. I believe that he/she understands the information described in this consent form and freely consents to participate.

____

Name of Investigator/ Study Team Member (print)

____ ____________________

Signature of Investigator/ Study Team Member Date

Letter of Consent

INVITATION TO PARTICIPATE

Title of the study: Effect of formative feedback via interactive concept maps on informal inferential reasoning and conceptual understanding of ANOVA

Principal Investigator
Sait Atas, Ph.D. Candidate, Faculty of Education, University of Ottawa

Thesis Supervisor
David Trumpower, Professor, Faculty of Education, University of Ottawa

Dear participant,
You are invited to participate in the Ph.D. thesis research project conducted by Sait Atas, under the supervision of David Trumpower. Thank you for considering this request.

Purpose of the Study: The purpose of the study is to examine the effectiveness of computer-based feedback that can improve learning for statistical reasoning.

Participation: Your participation will consist of a two-hour session during which you will be asked to provide demographic information including declared major, statistics background, and age and to complete two on-line tasks. Even though these tasks are related to statistics, you do not have to perform any calculations. The study will take about two hours and the participation will be entirely online. You can complete the study anytime, anywhere. You will only need to have access to a computer with the internet connection.

Risks: There are no risks associated with your decision to participate in this research.

Benefits: The intended product of this work- Effect of formative feedback via interactive concept maps on informal inferential reasoning and conceptual understanding of ANOVA will assist the course instructors in designing and delivering better statistics courses.

Confidentiality and anonymity: Your participation will be anonymous and your data will be stored with a special code. There is no way to determine your identity in the study as you will not be asked for any personal individual identifiers such as name, address, or prior schools of attendance.

Conservation of data: The data gathered form participants will be stored in the student investigator's computer and only the student investigator and thesis supervisor will have access to data. In order to ensure confidentiality of the data, completed data will be printed. The electronic data will be removed from the Internet. Printed data will be used to analyze and evaluate data. Upon project completion, data will be stored in a locked file cabinet in principal investigator's and supervisor's office for five years after the complete data collection at the University.

Voluntary Participation: You are under no obligation to participate and if you choose to participate, you may withdraw from the study at any time and/or refuse to answer any given question. If you choose to withdraw, all information you provided will be destroyed and not used in the report or subsequent publications. You will also have an option to withdraw your data as well.

There are two copies of the consent form, one of which is yours to keep.

If you have any questions or require more information about the research project, you may contact the researchers at the email addresses appearing below, or contact Sait Atas or David Trumpower.

If I have any questions regarding the ethical conduct of this study, I may contact the Protocol Officer for Ethics in Research, University of Ottawa, Tabaret Hall, 550 Cumberland Street, Room 154, Ottawa, ON K1N 6N5

Tel.: (613) 562-5387

Email: ethics@uottawa.ca

By clicking "I consent", I agree to participate in the study.

Please print this letter of consent for your records. You will also have option to save the document on your computer.

APPENDIX C

Textual Feedback

The Relation between *Random factors* and *Differences within groups*

To illustrate the relation between *random factors* and *differences within groups*, please recall the soccer ball experiment you have seen in the iANOVA task;
Now, suppose that the scores obtained by balls constructed from two different number of panels were as follows:

Scores of soccer balls by number of panels

Number of panels	
32-panel	**14-panel**
230	420
370	280
300	350

In this example, the balls made with 14 panels had 50-point higher average score than the balls made with 32 panels. At first glance, we could conclude that balls made with 14 panels are better balls with 32 panels. However, before concluding that 14-panel soccer balls are actually better than 32-panel ones, we need to recognize that not all of the balls kicked by the same force and made with the same number of panels obtained the same score. For instance, the three balls made with 32 panels differing by 70-point, on average, from one ball to another. This might have been due to factors such as gusts of wind and imperfections in the balls since all balls *within the same group* were made with the same number of panels. These factors are called "*random factors"* because they vary randomly across the experiment. Thus, it can said that the three balls made with the 32 panels differed only due to *random factors* as did the three balls with the 14 panels. We could estimate the effect of *random factors* on scores obtained by balls by considering the average amount of deviation within the scores obtained by balls made with the same number of panel and kicked by the same amount of force. In this experiment, considering that 70-point average deviation from one ball to another, we could conclude that "*random factors*" are entirely responsible for the 70-point "*difference within groups*" because the difference cannot be due to different number of panels on a ball since all balls in a group were made with the *same* number of panels.

The Relation between *Independent Variable* and *Difference between Groups*

To illustrate the relation between *independent variable* and *difference between groups*, please recall the soccer ball experiment you have seen in the iANOVA task;
Now, suppose that the scores obtained by balls constructed from two different number of panels were as follows:

Scores of soccer balls by number of panels

Number of panels	
32-panel	**14-panel**
298	325
302	323
300	327

Remember that the scientists were interested in measuring the effect of the different number of panels on a soccer ball's flight characteristics. In this experiment, you should have noticed that, the balls made out of 32 panels obtained 300-point average, whereas those made out of 14 panels obtained 325-point average. This 25 point average *difference between groups* might have been due to variable such as the number of panels with which soccer balls were constructed. This type of variable is called "*independent variable*" and was manipulated (32 versus 14) by the researchers to see the effect of the "*independent variable*" on balls' flight characteristics. The scientists could look at the average *difference between groups* in order to find out how large of an effect the "*independent variable*", number of panels, has on balls' flight characteristics. In other words, if the *independent variable*, number of panels, does indeed have an effect on a ball's flight characteristics, then there should be an observable difference between groups. For this experiment, considering 25-point average observed *difference between groups* of balls, we can conclude that the *difference between groups* is due to the *independent variable* of number of panels.

The Relation between *Random factors* and *Dependent Variable*

To illustrate the relation between *random factors* and *dependent variable*, please recall the soccer ball experiment you have seen in the iANOVA task;

In the soccer ball experiment, as you might remember, the researchers were interested in measuring the effect of different number of panels on soccer balls' flight characteristics. The flight characteristics of the soccer balls is a "*dependent variable*" in this experiment because we assume that balls' flight characteristics are dependent on many factors. For example, gusts of wind and imperfections in the balls can affect the flight of the balls. Factors such as these that have not been controlled by the researchers are called "*uncontrolled factors*" or "*random* *factors*" because they vary randomly across the experiment. For example, a strong crosswind altering the flight of the balls could occur at any time during the experiment. It is as likely to occur when a 32-panel soccer ball is being kicked with the robotic leg as when a 14-panel soccer ball is being kicked with the robotic leg. With this in mind, we could conclude that "*random factors*" such as gusts of wind and imperfections in the balls have an effect on the "*dependent variable*" that is soccer balls' flight characteristics in the soccer ball experiment. In any experiment, it is important to consider how large of an effect *random factors* have on the *dependent variable*.

The Relation between *Random factors* and *Differences between Groups*

To illustrate the relation between *random factors* and *difference between groups*, please recall the soccer ball experiment you have seen in the iANOVA task;
Now, suppose that the scores obtained by balls constructed from two different number of panels were as follows:

Scores of soccer balls by number of panels

Number of panels	
32-panel	**14-panel**
230	420
370	280
300	350

Could *difference between groups* be due to *random factors*? In answering this question, suppose that strong crosswinds were blowing when some of the 32-panel balls were being kicked, but that no such winds were blowing when the 14-panel balls were kicked. In such a scenario, the relatively lower scores of the 32-panel balls could have been due to the gusting winds. That is, the difference between the average scores of the groups of 32-panel and 14-panel balls could have been due to random factors alone, rather than being due to the number of panels. The answer is yes, it is possible that the 50-point average difference between groups is due to *random factors*.

The Relation between *Independent* and *Dependent Variable*

To illustrate the relation between *independent variable* and *dependent variable*, please recall the soccer ball experiment you have seen in the iANOVA task;

The scientists were attempting to measure the effect of the different number of panels on the flight characteristics of soccer balls. In that particular experiment, the number of panels represent the "independent variable" and a soccer ball's flight characteristics represents the "dependent variable". Remember that researchers believe that a soccer balls' flight characteristics (i.e., dependent variable) might have been affected by the number of panels (i.e., independent variable). In other words, the "d*ependent variable*" (i.e., soccer balls' flight characteristics), as the name suggests, depends on the change in the "*independent variable*" (i.e., number of panels on a soccer ball). For example, the scientists in the soccer ball experiment were trying to describe how much of the observed variation in the soccer balls' flight characteristics (i.e., dependent variable) can be explained by, or is dependent upon, the number of panels on a soccer ball (i.e., independent variable).

The Relation between *Strength of Evidence* and *Differences between Groups*

To illustrate the relation between *strength of evidence* and *difference between groups*, please recall the soccer ball experiment you have seen in the iANOVA task;
In determining the better type of soccer balls, the scientists look for evidence that makes the researchers confident that random factors such as such as gusts of wind and imperfections in the balls alone cannot produce the difference observed between the scores obtained by samples of 32-panel and 14-panel soccer balls. Now, suppose that the scores obtained by balls constructed from two different number of panels were as follows:

Scores of soccer balls by number of panels

Number of panels	
32-panel	**14-panel**
230	420
370	280
300	350

The *strength of evidence* that the number of panels had a real effect on the flight characteristics of balls depends on the size of the difference between the average scores of the two groups of balls (in this example 50 points). The *strength of evidence* that the 50-point difference between the average scores of the two groups dependents on the deviation within groups. For example, suppose that a strong crosswind happened to gust every time each 32-panel soccer ball was being kicked. In this case, each soccer ball within the same group of 32-panel balls would get different scores even though they were kicked with the same amount of force and made with same number of panels. In this example, there is a 70-point difference, on average, from one ball to another in within the two groups of balls. Now, think about another hypothetical result from soccer ball experiment where there is 50-point average difference between groups and 2-point average difference within the groups. Which one of the results would provide stronger evidence that random factors such as crosswind cannot produce the difference observed between the scores obtained by samples of 32-panel and 14-panel soccer balls?

If the differences within groups are just as large as the difference between groups, then it would seem plausible that the difference between the 32 and 14 panel balls was simply due to random factors; however, if the *difference between groups* is larger than the differences within groups, then it would seem that more than random factors are needed to explain the difference between the 32 and 14 panel balls. That is, the larger the *difference between groups* is, the *stronger is the evidence* that the difference between the 32 and 14 panel balls was not only due to random factors, but could have also been influenced by the number of panels.