INTERACTIVE ENGAGEMENT IN REASONING ABOUT VARIANCE

David L. Trumpower
University of Ottawa, Canada
david.trumpower@uottawa.ca

Students' reasoning about variation is often inconsistent with the normative logic underlying statistical procedures such as ANOVA, even after formal training. This study introduced an interactive classroom activity intended to support normative reasoning about variance. In small groups, students were shown a small dataset from a hypothetical, between-groups, experiment and asked to collaboratively consider how the within- and between-group variance could provide evidence about an effect of an independent variable. Before engaging in the interactive activity, most students considered between- and within-group variability in a non-normative manner. Following the activity, nearly all students' reasoning shifted to a normative pattern, recognizing the relevance of the ratio of between- to within-group variation for making judgements about group differences. The findings support Chi's ICAP theory of cognitive engagement and presents a novel interactive activity to support students' developing understanding of ANOVA, which can be very difficult to achieve in introductory courses.

INTRODUCTION

Students' reasoning about variation is often inconsistent with the normative logic underlying statistical procedures such as ANOVA, even after formal training (Reid & Reading, 2008; Trumpower & Fellus, 2008). For example, when weighing evidence about a possible difference between two independent groups based on sample data, students are prone to overemphasize the importance of the absolute magnitude of the between-group difference in sample means with less consideration for within-group variation (Obrecht, Chapman & Gelman, 2007; Masnick & Morris, 2008; Trumpower & Fellus, 2008; Trumpower, 2013a, Experiment 1; Trumpower, 2015). Even after instruction on ANOVA, students often have difficulty recognizing that it is the ratio of the difference in sample means *relative to* the within-group variation that provides evidence against a null hypothesis of no real effect of the independent variable. Whereas within-group variability provides an indication of the influence of random, uncontrolled factors, between-group variability can arise from any effect of the independent variable as well as effects of random, uncontrolled factors. Thus, the magnitude of a between-group difference in sample means alone does not provide useful information about the possibility of an effect of the independent variable, because the difference could be entirely due to random factors. However, by considering the magnitude of between-group variability relative to the variability attributable to random factors (as indicated by within-group variance), we can estimate the potential influence of the independent variable. This study introduced an interactive classroom activity, based on the Interactive Constructive Active Passive (ICAP) theory of cognitive engagement (Chi, 2009; Chi & Wylie, 2014), that was intended to support students' normative reasoning about variance.

Informal Analysis of Variance

In a series of studies using a task that has been deemed informal Analysis of Variance (iANOVA), Trumpower and colleagues have explored how individuals reason about variation when making informal inferences from samples of data. In the typical iANOVA task, participants are provided a cover story about an experiment in which small samples of data are collected in order to test for a potential difference between two distinct conditions (e.g., Frankenwood vs. Real wood baseball bats) and then shown several hypothetical sets of results from such an experiment and are asked to rate the strength of evidence provided by each. If using normative reasoning, ratings of the hypothetical results should correspond with the ratio of between to within-group variance, as would be given by an F-statistic (or Bayesian analysis). However, participants' pattern of ratings and associated verbal explanations consistently deviate from such normative reasoning. In particular, participants tend to rate datasets with a smaller absolute difference in sample means as weaker evidence than datasets with a larger absolute difference in sample means, even if the datasets with a smaller mean difference have a larger ratio of between to within-group variance (Obrecht, Chapman & Gelman,

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada. http://iase-web.org

2007; Masnick & Morris, 2008; Trumpower & Fellus, 2008; Trumpower, 2013a, Experiment 1; Trumpower, 2015). For example, a dataset in which the difference in means is equal to 5 and the average within-group standard deviation is 2 (resulting in a ratio of 5:2 = 2.5) is often viewed as weaker evidence than a dataset with a difference in means equal to 20 and average standard deviation of 20 (resulting in a ratio of 20:20 = 1.0). Qualitative analysis of participants' reasoning reveals that they often have difficulty integrating the between and within-group variation in a relative fashion (Trumpower & Fellus, 2008; Trumpower, 2015, Experiment 2). Instead, participants typically consider the two sources of variation independently with more weight given to between-group difference, sometimes with no weight at all given to within-group variation. It has been conjectured that this difficulty may stem from a lack of understanding that between-group differences can be attributed to both fixed factors *and* random factors, whereas within-group variation is due to random factors only (thus, providing a reference with which to compare the magnitude of between-group differences).

Although the difficulty of integrating between and within group variation is persistent, certain interventions have been proposed to support the development of normative reasoning about ANOVA (Trumpower, 2013b; Trumpower, 2013b), with some success. For instance, Trumpower (2013b) used the iANOVA task as a classroom tool to provide students with formative feedback regarding the normative logic underlying ANOVA and how their reasoning compared with it. Additionally, students were given a culminating assignment in which they were required to analyze and interpret other students' performance on the iANOVA task. It was found that the feedback combined with the assignment, which required active explanation of both normative and non-normative reasoning applied to the task, led to more normative inferential reasoning at the end of the course, whereas providing feedback alone did not. However, anecdotally, the culminating assignment requires much scaffolding on the part of the instructor and has been found to be quite confusing to many students. Chi's ICAP framework may provide a more robust and efficient strategy for supporting the development of students' reasoning about variance.

ICAP

The ICAP framework of active learning (Chi, 2009; Chi & Wylie, 2014) posits that *Interactive* engagement (as defined by co-generative collaborative behaviors) leads to deeper learning than *Constructive* engagement (defined by individual generative behaviors), which in turn leads to deeper learning than *Active* engagement (defined by manipulative behaviors), or *Passive* engagement (defined by attentive behaviors only) which results in the shallowest learning. Within the context of learning about variation, passive engagement could be exemplified by quietly listening to a lecture on variation or silently reading a text on the topic (and, involves only *storing* the information), whereas active engagement could be exemplified by taking verbatim notes during the lecture or highlighting passages in the text (and, involves *storing*, *activating* and *linking*). Both of these modes of engagement have been associated with rather shallow levels of learning, although passive modes are inferior to active modes in this respect. An example of constructive engagement, which is associated with a deeper level of learning, would be creating a summary of the lecture or text in one's own words and/or posing one's own examples of variability (involves *storing*, *activating*, *linking* and *inferring*). Associated with an even deeper level of learning, however, is interactive engagement which could be exemplified by collaboratively creating a summary with a partner, where both partners provide examples that they agree upon (also involving *storing*, *activating*, *linking* and *inferring*, but also includes inferring from both one's own knowledge and inferring from others' knowledge). Unfortunately, devising interactive and constructive activities is not always an easy task for instructors. Thus, this paper introduces and tests an interactive activity intended to support individuals' iANOVA reasoning.

The interactive learning activity developed for this study begins with presenting students with a version of the iANOVA task, comprised of a cover story and associated sample data from two comparison groups, and asking them to describe how the within- and between-group variation observed in the data can be used as evidence of the effect of an independent variable. This part of the iANOVA activity corresponds with *constructive* engagement, as individuals must activate their knowledge of variation and link it with the observed variation in the dataset to make inferences and generate answers to the question. Next, individuals are to be provided with several possible answers

to the question about how within- and between-group variability can be used to make inferences about the potential effect of an independent variable and, in small groups, asked to discuss and rank the possible answers from best to worst, and to justify their rankings. This part of the iANOVA activity corresponds with *interactive* engagement, as individuals must again activate their knowledge of variation, link it to the provided dataset and make inferences, but now must co-generate a justification with other group members.

METHOD

Ten students in an introductory-level university statistics course served as participants. As part of the course requirements, students were asked to complete a brief pretest, engage in small group discussions, and complete a homework assignment on analysis of variance, as described below.

On Halloween, kids typically choose to wear either scary or cute costumes. Suppose a researcher is interested in the potential effect that type of costume (scary or cute) has on how much candy one receives when trick-or-treating. They hypothesize that scary costumes will scare people into giving less candy than cute costumes. In order to test their hypothesis, they find 6 kids whose parents allow them to participate in an experiment. Next, they randomly assign 3 of the kids to wear a scary ghost costume and the other 3 kids to wear a cute teddy bear costume. They then allow the kids to trick-or-treat for two hours, after which they measure the amount of candy received by each kid.

Listed below are the results of the experiment. The numbers indicate the amount of candy (in ounces) received by the kids wearing the scary and cute costumes.

Scary	*Cute*
350	325
250	375
300	275

Figure 1. Cover story and dataset presented at pretest.

Pretest

On the first day of the course, students were shown a small dataset from a hypothetical experiment designed to test the effect of scary versus cute Halloween costumes on the amount of candy received by trick-or-treaters. Data were presented in two columns, along with the cover story, as shown in Figure 1. Individually, students were asked to, "Explain how the variability in amount of candy *within* either column along with the difference in amount of candy *between* the two columns, together, provide evidence about the researcher's hypothesis. Be sure to refer to BOTH the variability within AND between columns." Students submitted their responses electronically as an initial requirement for the course.

Intervention

Later in the semester, students were again shown the dataset from the pretest. As well, they were provided with the following possible responses to the question, "To what extent do these results support the researcher's hypothesis? Explain":

1. The average amount of candy received by kids wearing the cute costume (mean=325 ounces) was 25 ounces more than the average amount received by kids wearing the scary costume (mean = 300 ounces). But, there was also variation in how much candy was received by kids wearing the same costume - the kids wearing the scary costume differed by about 50 ounces from one another, and the kids wearing the cute costume differed from one another by 50 ounces, too. This shows that, regardless of the type of costume, there were random factors (e.g., individual differences) that affected how much candy each kid received. Because the difference between the means of the groups is fairly large (325 – 300 = 25 ounces), it looks like type of costume did have an effect, even though the differences within each group show that random factors had an effect, too.
2. The average amount of candy received by kids wearing the cute costume (mean=325 ounces) was 25 ounces more than the average amount received by kids wearing the scary costume (mean = 300

ounces). Because the difference between the means of the groups is fairly large (325 – 300 = 25 ounces), it looks like type of costume did have an effect.

3. The average amount of candy received by kids wearing the cute costume (mean=325 ounces) was 25 ounces more than the average amount received by kids wearing the scary costume (mean = 300 ounces). But, there was variation in how much candy was received by kids wearing the same costume - the kids wearing the scary costume differed by about 50 ounces from one another, and the kids wearing the cute costume differed from one another by 50 ounces, too. This shows that, regardless of type of costume, there were random factors (e.g., individual differences) that affected how much candy each kid received. In fact, the differences within each group (about 50 ounces) were twice as big as the difference between the means of the groups (325 – 300 = 25 ounces). So, it looks like the difference between the groups could have been due to random factors rather than type of costume.
4. There was variation in how much candy was received by kids wearing the same costume - the kids wearing the scary costume differed by about 50 ounces from one another, and the kids wearing the cute costume differed from one another by 50 ounces, too. This shows that, regardless of type of costume, there were random factors (e.g., individual differences) that affected how much candy each kid received. So, it looks like the difference between the groups could have been due to random factors rather than type of costume.
5. Every kid wearing the cute costume received 25 ounces more than a kid wearing the scary costume. This difference was consistent across all 3 pairs of kids. So, it looks like type of costume has an effect.
6. There was variation in how much candy was received by kids wearing the same costume - the kids wearing the scary costume differed by about 50 ounces from one another, and the kids wearing the cute costume differed from one another by 50 ounces, too. This shows that, regardless of type of costume, there were random factors (e.g., individual differences) that affected how much candy each kid received. Because the variation within each group was the same, it looks like type of costume did not have an effect.
7. The kids wearing the cute costume received up to 125 ounces more than kids wearing the scary costume. So, it looks like type of costume has an effect.

Each of the seven responses had been constructed to reflect a different type of reasoning about variation based on actual participant responses observed in previous studies using the iANOVA task. For example, some individuals with a weak understanding of variation focus solely on the magnitude of between-group differences while ignoring within-group variation altogether, whereas others with a developing understanding of variation consider both the between- and within-group variation independently. Those with a more complete understanding of variation also consider between- and within-group variation, but relative to one another rather than independently. That is, they consider the ratio of between- to within-group variation, consistent with the normative logic of ANOVA (see Trumpower, 2013a; 2015).

In small groups of 2-3 students, participants were asked to discuss and rank these responses from best to worst. Then, in a large group discussion, each group was asked to present and explain their top and bottom ranked responses in order to invoke Chi's interactive mode of cognitive engagement.

Posttest

Finally, for homework, students were individually asked to rate the strength of evidence provided by six additional hypothetical datasets and explain their responses in writing. The datasets presented for this homework assignment were analogous to, but slightly modified from, those used in prior versions of the iANOVA task (see, e.g., Trumpower, 2015). The normative strength of evidence, as indicated by the ratio of between- to within-group variance present in each, varied across the six datasets (as can be seen in the first three columns of Table 1 below).

RESULTS

Students' qualitative responses to the pretest question were analyzed to determine how they reason about variation before any formal instruction on the topic of inferential statistics. Responses

were coded as being either normative or non-normative reasoning. To be considered normative, a response was required to mention both between- and within-group variation and the magnitude of the between-group variability had to be considered *relative to* the magnitude of within-group variability (consistent with *strong* consideration of variation on Reid & Reading's hierarchy); all other responses were considered non-normative (consistent with *developing* or *weak* levels on Reid & Reading's hierarchy). Initially, most (9/10) students considered between- and within-group variability in a non-normative manner as distinct, independent sources of information.

During the interactive intervention, it was noted which of the six responses each group ranked as providing the best explanation. Initially, four of the five groups decided that the normative explanation, in which it is the ratio of between-group variation relative to within-group variation that provides evidence about the effect of an independent variable, was best. The only group that did not rank the normative explanation as best had ranked it second best. However, after the large group discussion, this group came to agree with the other groups that the normative explanation was indeed the best.

On the subsequent posttest, students' mean ratings of the six datasets were determined. As can be seen in Table 1, students' mean ratings followed a normative pattern. That is, the rank order of students' mean ratings for each of the six datasets corresponded perfectly with the rank order of the normative strength of evidence provided by each dataset as indicated by an F-ratio.

Table 1. Mean ratings of six datasets on posttest

Dataset[1]			
Difference in means	Standard deviation	F-ratio	Mean rating[2]
25	2	234.37	10.00 (0)
15	2	84.37	8.30 (1.27)
5	2	9.37	5.85 (1.10)
25	20	2.34	4.35 (2.17)
15	20	.84	2.20 (.75)
5	20	.09	1.00 (0)

[1]The first three columns describe different characteristics of the datasets that participants were asked to rate. As can be seen, the datasets are ordered by the F-ratio that would result from conducting a one-way, between groups ANOVA on that dataset (i.e., the normative strength of evidence provided by the dataset).

[2]Mean of students' ratings of the strength of evidence provided by each dataset on a scale of 1-10, where 1 indicates the weakest and 10 the strongest evidence in support of true effect of the independent factor described in the hypothetical experiment from which the datasets were derived.

Students' qualitative justifications of their ratings of the six datasets were also analyzed and categorized as either displaying normative or non-normative reasoning. It was revealed that a majority (8/10) of students' written explanations indicated that their reasoning was now normative, recognizing the relevance of the ratio of between- to within-group variation.

DISCUSSION

These findings support the ICAP theory prediction that activities which promote collaborative cognitive engagement can be effective in supporting students' developing understanding of ANOVA, which can be difficult to achieve in introductory courses. Whereas at pretest, only one student displayed normative reasoning about variation on an ANOVA task, all of the groups came to recognize normative reasoning during the collaborative activity which required interactive cognitive engagement. More importantly, nearly all of the class displayed normative reasoning on a subsequent ANOVA homework assignment. Thus, it appears that the interactive classroom activity was successful at supporting students' development of normative reasoning about variation.

When individuals perform informal ANOVA-type tasks without any formal training, they often focus solely on between-group variability, corresponding with Reid and Reading's weak level of consideration of variability. In the present study, students were prompted to consider both within- and between-group variability on the pretest. Therefore, the finding that almost all of them still failed to consider between- and within-group variability relationally at pretest attests to the strong predilection to reason non-normatively in group comparison problems such as presented in the iANOVA task. As

well, the interactive task utilized here only took approximately 30 minutes of class time. This makes the shift from non-normative to normative reasoning that was observed following the interactive task even more impressive. The findings presented here suggest that Chi's ICAP framework may serve as the basis for developing effective learning activities in other notoriously difficult subjects.

REFERENCES

Chi, M. T. H. (2009). Active-Constructive-Interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*, 73–105.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243.

Reid, J., & Reading, C. (2008). Measuring the development of students' consideration of variation. *Statistics Education Research Journal, 7*(1), 40–59.

Trumpower, D.L. (2013a). Formative use of intuitive analysis of variance. *Mathematical Thinking and Learning, 15*(4), 291-313.

Trumpower, D.L. (2013b). Intuitive analysis of variance - A formative assessment approach. *Teaching Statistics, 35*(1), 57-60.

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