CONTRASTING CASES: THE "B VERSUS C" ASSESSMENT TOOL FOR ACTIVATING TRANSFER

Rachelle Kisst Hackett
Benerd School of Education, University of the Pacific, United States of America
RHackett@Pacific.edu

This paper focuses on an assessment method that has been employed on exams given to education students in an applied graduate-level statistics course, but could be incorporated as a class activity or given as homework in undergraduate or graduate courses in other fields. Students review the work of two presumably competent statistical consultants labeled "B" and "C", who have each attempted to address the same research hypothesis using the same data. After contrasting the cases, the students write letters to either consultant (or to both) who they think is in error, explaining the nature of the mistakes. Sample "B vs. C" problems are presented including descriptions of the consultants' work and key features upon which the scoring of student answers focus. In addition to identifying theoretical underpinnings (especially Bransford & Schwartz, 1999), student reactions to this assessment method are shared.

BACKGROUND

During my graduate training, but particularly as a post-graduate research assistant, when working with prominent American cognitive psychologists (James Greeno and John Bransford, for example) the importance of considering prior knowledge and working towards conceptual change had been emphasized. Making students' thinking visible and building from it were valued activities in the learning process. Getting students to articulate their thinking, whether orally or in writing, was important for both formative and summative evaluation purposes. As part of the CTGV (Cognition and Technology Group at Vanderbilt), I was involved in the creation of "Blueprint for Success," one of the adventures within the Jasper Woodbury series (videodisc-based anchored instruction materials that focus on mathematical problem finding and problem solving; See http://peabody.vanderbilt.edu/projects/funded/jasper/). Associated with Jasper adventures were "SMART" tools which featured a reporter who would talk to students about the approaches they were taking to solve the problem (Zech et al., 1998). These students were actually paid actors/ actresses but the processes being modeled were based on what we observed students doing in the classroom. Thus, while some of the solutions being modeled were productive approaches, many highlighted student misconceptions (but without explicitly labeling them as such.) The idea was that students could look at others' work and decide whether they wanted to revise their own mathematical work. The intent was that, by contrasting cases, students' thinking could be refined, ultimately enabling them to engage in a productive solution to the problem.

When I began teaching applied statistics to graduate students over 10 years ago within a school of education, I wanted to emphasize conceptual understanding and rely on technology to perform mathematical calculations (particularly since many students in this field typically lack confidence in their mathematical ability). However, I feared that menu-driven statistical software interfaces might lead to inappropriate uses for those who overly relied upon such seeming ease without understanding various options and subsequent interpretation of the statistical output. I was also aware that many of the students in our program were not required to take another course thereby limiting the types of analyses they might competently perform by themselves. Thus, it seemed important to give them a solid foundation upon which to build their statistical thinking rather than to expose them to several techniques that they would only superficially understand. And, my hope was that an emphasis on statistical reasoning and communication would better position them to productively engage with professors and/or statistical consultants with whom they could then collaborate rather than upon whom they would become overly dependent. These goals complement those articulated for the statistics education reform movement which emphasized attention to statistical literacy, reasoning, and thinking (delMas, Garfield, Ooms & Chance, 2007).

Thus, shortly after becoming a professor, my post-doctoral training with the CTGV began to inform my own teaching practice. In particular, the use of contrasting cases for assessment purposes became a hallmark of the intermediate statistics course I taught to masters- and doctoral-level students, as most of my take-home exams have included one "B vs. C" problem that students were required to address.

In C. Reading (Ed.), *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, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

GENERAL DESCRIPTION OF THE "B VS. C" ASSESSMENT METHOD

In short, students are asked to review the work of two presumably competent statistical consultants (professors) who are simply referred to as "B" and "C", each of whom has attempted to address the same research hypothesis using the same data. The statistical software output and commentary generated by each consultant is shown. After contrasting the cases, the students write letters to either consultant (or to both) who they think is in error, explaining the nature of the mistakes. It is acknowledged that some of the errors may be exacerbated by the particular statistical software (Statistical Packages for the Social Sciences which is also called "SPSS") being used in the course but the errors, nonetheless, reflect more than procedural issues.

The types of errors that I showcase in the hypothetical work of individuals "B" and "C" are typically ones that I have noticed students making in my course when solving problems involving the use of inferential hypotheses testing and/or related estimates of confidence intervals. And, sometimes I model correct or incorrect approaches in an attempt to stimulate students' thinking, whether or not my students have actually committed the error themselves. The "B vs. C" assessment format was an adaptation of ideas upon which the CTGV based its "SMART" tools. In other words, my intent was that, by contrasting cases, students' thinking could be refined, enabling them not only to engage in a productive solution to the "B vs. C" problem at hand, but, ultimately leading them to rethink, and possibly revise their own mathematical work on other non "B vs. C" statistical inference problems. Thus, as an assessment of learning, the method is not strictly a paradigm "that characterizes transfer as the ability to directly apply one's previous learning to a new setting or problem," but one that "broadens the conception of transfer by including an emphasis on people's 'preparation for future learning' (PFL)" (Bransford & Schwartz, 1999).

Responses are graded holistically but focus on both the correct identification of any major errors and the quality of the written explanation. Exemplary responses demonstrate that the student can do more than point to procedural errors; students' explanations should include statistical reasoning (Garfield, 2002) about the connections between propositional knowledge. The latter results in conceptual understanding (Broers, 2006), and, coupled with procedural knowledge, both are needed by the student to evaluate the statistical thinking (Chance, 2002) of the consultant's work. Also, the "B vs. C" method can be seen as testing statistical literary because quality explanations require "the ability to interpret, critically evaluate, and communicate about statistical information and messages" (Gal, 2002).

IDENTIFYNG TYPES OF ERRORS TO MODEL

It seems that the "B vs. C" format may be incorporated into the assessment of many statistics course objectives because the errors being modeled can occur at either the generation of the statistical output or the interpretive level (as displayed by the hypothetical annotations that "B" or "C" add to their output). The course in which this method of assessment is used assumes the graduate student has had prior exposure to basic descriptive statistics; the course coverage focuses on inferential techniques commonly employed in the social sciences. In this paper, I will focus on "B vs. C" problems that have been used for Cluster A (one-sample, two-independent samples, and two-dependent samples t-tests involving means and/or their related confidence intervals) and Cluster C (Chi-Squared Goodness of Fit Tests, Chi-Squared Tests of Associations, and related tests of proportions: one-sample, two-independent samples, and two-dependent samples). In my own experience, I try to vary which, if either, consultant is correct, and to bundle errors that tend to be made simultaneously. The kinds of mistakes that can be modeled include:

- Drawing a conclusion based only on the descriptive statistics without considering sampling error and the associated p-value.
- Failing to alter the "test value" (from SPSS' default value of zero) in a one-sample t-test to match that implied by the null hypothesis.
- Using the two-tailed p-value when a directional alternative hypothesis is implied.
- Being inconsistent when setting up a directional alternative hypothesis, visualizing the rejection region in the corresponding tail of the sampling distribution, and calculating the test value using the corresponding order of subtraction.
- Failing to use the modified t-test when the assumption of homogeneity of variance is not met.

- Disregarding necessary modifications to the level of confidence when relying on 2-sided CI output for testing directional hypotheses (i.e., an alpha=0.05 one-tailed test is related a 90%, not a 95%, 2-sided CI).
- Analyzing nominal data as a chi-square goodness of fit test when a chi-square test of association (also known as a test of independence) is needed, or vice versa.
- Failing to correctly specify the model (i.e., expected values) when using the chi-square goodness of fit test (from SPSS default that all categories are equal).
- Mistaking the row versus column percentages when checking that the sample statistics are consistent with that predicted by a directional alternative hypothesis concerning differences between independent samples.

EXAMPLE ONE: A DIRECTIONAL ONE-SAMPLE T-TEST

Instructions (for Example One) as Presented to Students

On the exam, the instructions are as follows. “Research Question: Is there evidence to suggest third graders spend less than 60 minutes, on average, per week, doing science homework? Use alpha= 0.05. The research question has been analyzed by two consultants: Professor B and Professor C. They analyzed the data independently but agreed, from the start, to adopt an alpha level of .05 for this problem. The printouts of SPSS runs performed by Professor B and Professor C are attached. Each has annotated the printouts highlighting information s/he felt was critical for evaluating the research question being considered. Your task it to look over the printouts separately annotated by Professor B and Professor C and to determine who, if either, has correctly analyzed the data. Then you are to write a short letter (not to exceed one page) to the professor less competent in data analysis explaining exactly where s/he went wrong! If both either did the analysis improperly or drew the incorrect conclusion, write each professor a letter explaining the error(s). The grade you receive for this problem does depend on (1) the accuracy of your answer and (2) the quality of the explanation you provide in your letter(s). Be as specific as possible in pointing out any major error(s).”

Description of the SPSS Output and Annotations (for Example One) Produced by “B” and “C”

Professor “B” performs a one-sample t-test setting the test value to 60. The one-sample statistics show N=121, Mean= 64.96, Std. Deviation= 30.751, and Std. Error Mean= 2.796. The one-sample test output shows t= 1.774, df= 120, Sig. (2-tailed) = 0.079, Mean Difference = 4.959, and a 95% Confidence Interval whose lower and upper values are -.58 and 10.49, respectively. Professor “B” remarks: “The null hypothesis is mu=60, so the test value must be set to 60. The alternative hypothesis is directional, so the 2-tailed sig must be split in half. p= 0.079 / 2= 0.0395. p < 0.05 (alpha). Reject the null hypothesis. There is evidence to suggest third graders spend less than 60 minutes each week, on average, doing science homework, t(120)= 1.774, p= 0.0395.”

Professor “C” performs a one-sample t-test leaving the test value at 0 (zero). Just like those of Professor “B”, the one-sample statistics show N=121, Mean= 64.96, Std. Deviation= 30.751, and Std. Error Mean= 2.796. In contrast to those of Professor “B,” the one-sample test output of Professor “C” shows t= 23.237, df= 120, Sig. (2-tailed) = 0.000, Mean Difference = 64.959, and a 90% Confidence Interval whose lower and upper values are 60.32 and 69.59, respectively. Professor “C” remarks: “Since alpha=0.05 and our alternative hypothesis is directional, the related confidence interval is 90%. The null hypothesis states mu=60. I see that this value, 60, is not within the 90% CI. Thus, the null hypothesis should be rejected. There is evidence to suggest third graders spend less than 60 minutes each week, on average, doing science homework, t(120)= 23.237, p < 0.0005.”

Scoring Responses (to Example One)

The conclusion drawn by both Professors “B” and “C” is incorrect. Relatedly, some errors are made in the processes they used to arrive at their conclusion. Thus students would be expected to write separate letters to both professors. The letter to *both* professors should note that (1) the sample statistics were not even in the predicted direction; (2) the wrong conclusion had been reached because there is insufficient evidence to support the alternative hypothesis (i.e., there is insufficient evidence to suggest that 3rd graders spend less than 60 minutes each week, on average, doing science homework); and (3) the p-value is greater than .05, the alpha level. The letter to Professor "B" should point out that the process of splitting the two-tailed significance level in half is only appropriate when the results are in the predicted direction. The letter to Professor "C" should point out that (1) the numerical summary (calculated value of the test statistic and p-value) is incorrect because it is based on the incorrect test value, and (2) the process of checking to see that the test value is not contained within the related confidence interval as a basis for rejecting the null hypothesis does not directly apply for tests of directional alternative hypotheses. Furthermore, students' letters to Professor "B" and to Professor "C" should attempt to explain the conceptual reasons why the processes they used were misapplications. In other words, the response exhibited by Professor "B" suggests a failure to understand what the p-value actually represents whereas that of Professor "C" also suggests a failure to fully understand how hypothesis testing and confidence interval estimation relate, in the case of one-tailed tests. These conceptual points should be elaborated upon in the letters (i.e., students' responses to the exam problem).

EXAMPLE TWO: A DIRECTIONAL, INDEPENDENT SAMPLES PROPORTION TEST

Instructions (for Example Two) as Presented to Students

On the exam, the instructions are as follows. "Two professors (B and C) were approached by a student on whose dissertation committee they were serving. The student brought the crosstabulated output shown below and separately asked each professor what she or he would conclude in regard to the Research Hypothesis. Look over their comments and determine which professor, if any, has properly advised his/her doctoral student. Write a letter to any (maybe both) professor(s) who needs statistical instruction and explain the error(s) of his/her ways! Be as explicit as possible. Research Hypothesis: The proportion of students who believe that computer careers are more appropriate for men (than for women) is higher for those most interested in Computer Engineering than it is for those most interested in Webmastering. Use alpha=.05."

Description of the SPSS Output and Annotations (for Example Two) Produced by "B" and "C"

Both professors provide the same SPSS output: a crosstabulation of the careers in which the (hypothetical) survey respondents are most interested (Row 1= Webmaster versus Row 2= Computer Engineer) by belief regarding computer careers being more appropriate for men (Column 1= Disagree versus Column 2 = Agree) than for women. Among the 45 interested in Webmastering, 26 disagree while 19 agree with the belief. Among the 95 interested in Computer Engineering, 70 disagree while 25 agree with the belief. In addition to showing these four observed counts, the "% within career" (row percentages) and "% within belief" (column percentages) are provided. The chi-square test output includes a row labeled "Pearson Chi-Square" for which the Value, df, and Asymp. Sig. (2-sided) are given as 3.585, 1, and .058, respectively.

Professor "B" comments as follows: "Your research hypothesis is directional because one group's proportion is hypothesized to be higher than the other group's proportion. Notice that the asymptotic significance level for the Pearson Chi-Square test given by SPSS is noted to be 2-sided. Therefore, for a directional test, the .058 must be divided in half. Then the resulting p-value (.029) for the one-tailed test would be less than alpha (.05) and you can reject the null hypothesis. So your research hypothesis is supported. Your APA summary will read: χ^2 (1, N=140)= 3.585, p= .029 ."

While the SPSS output provided by Professor "C" is identical to that of Professor "B," his interpretation varies. Professor "C" comments as follows: "Your research hypothesis cannot be supported. It is true that the sample statistics are in the predicted direction. Notice that there are 56.8% of those interested in Computer Engineering who believe the gender stereotype. And there are just 43.2% of those interested in Webmastering who believe the gender stereotype. However, the p-value (.058) is larger than alpha (.05) so you must fail to reject the null hypothesis. Your APA summary will read: χ^2 (1, N=140)= 3.585, p= .058 ."

Scoring Responses to Example Two

The decision of Professor "B" to reject the null hypothesis is incorrect because the results were not in the predicted direction. The process of utilizing the output from a chi-square test to inform a conclusion regarding one population's proportion exceeding another population's proportion is partially understood. However, Professor "B" does not seem to recognize that, because the chi-square calculation effectively tests whether there is *any* difference between two proportions, the SPSS user must still check the sample proportions to determine whether they are consistent with the hypothesized direction. The z test statistic, can be easily obtained by taking the square root of the calculated chi-square value, but it is incumbent upon the user to apply the positive or negative sign to the (absolute) value of the square root before arriving at a decision and conclusion.

The errors made by Professor "C" are different. Unlike, Professor "B," Professor "C" does not show an awareness that the asymptotic significance corresponds to a non-directional (2-tailed) hypothesis test. As a result, Professor "C" correctly fails to reject the null hypothesis, but for the wrong reason. Coupled with this error, we see that Professor "C" does not recognize that it is the row percentages that should be compared (26.3% versus 42.2%, for those interested in computer engineering versus webmastering, respectively) rather than the column percentages. As a result, Professor "C" fails to realize that the sample statistics are not in the predicted direction. Ideally, the student should explain to Professor "C" why the row percentages, rather than the column percentages, address the research hypothesis, and, in particular, note that there are over twice as many respondents in the second group (n=95) compared to the first group (n=45).

STUDENT REACTIONS TO B VS. C ASSESSMENT METHOD

Students enrolled in my statistics course during summer 2009 (n=14) were asked to anonymously complete a short survey consisting of Likert-type items and an opportunity to provide comments. Despite the limitations of this small sample of convenience, some general impressions about student reactions to this method of assessment are suggested by the results which include:

- 93% agreed that answering the "B vs. C" type items was challenging for them.
- 86% agreed that having difficulty with the "B vs. C" type items made them question their abilities to competently analyze data.
- 77% agreed that responding to the "B vs. C" type items had taught them that over reliance on a statistical consultant may be unwise.
- 69% agreed that, by reviewing another person's work (such as that of "B" or "C") they were able to spot errors/ misconceptions that they had themselves made.
- 57% agreed that, in the process of answering the "B vs. C" type items, they learned more about statistical concepts.
- While just 46% agreed that they would recommend that statistics course instructors include "B vs. C" type items as part of their *assessment* practice, 69% would recommend such items as part of instructors' *instructional* practices.
- 62% agreed that having two cases to contrast is an effective way to help students learn.

When asked to comment on anything the student would like for me or other statistics course instructors to know about the practice of assessing students using "B vs. C" type items, three students remarked as follows:

- "While difficult, I can see the rationale. I would have liked more practice with this during the instructional portion."
- "I don't particularly look forward to these problems, but I think they are helpful. I would recommend completing some in class while learning the procedures / concepts."
- "Although the questions did require additional critical thinking, they did allow for deeper understanding."

DISCUSSION

While the "B vs. C" method can serve as an assessment for evaluating what students have learned, it was designed to function more as an assessment of students' capacity to learn, via reasoning. It bears a resemblance to dynamic assessment, if one considers the written work of the two consultants as an "intervention" that assists students in furthering their learning. Bransford and Schwartz (1999, p.94) "emphasized the importance of using dynamic assessments to measure the degree to which people's past experiences have prepared them for future learning".

It was expected that most students would be challenged by the "B vs. C" problems. These types of problems had purposefully *not* been modeled in class. At the time, I had been concerned that using them as part of instruction, would compromise the validity of interpreting performance on the "B vs. C" problems as an indicator of students' statistical reasoning and ability to transfer

(Bude, 2006). The majority of students agreed that, in the process of answering the “B vs. C” type items, they learned more about statistical concepts. This suggests, at least partially, the method may be a promising tool for activating transfer. However, one approach to the evaluation of transfer, it seems, would necessitate the explicit design of coupled “B vs. C” and non-“B vs. C” items placed within the same or a subsequent examination.

The validity of claims made as to what the “B vs. C” method measures rests upon further investigation. Videotapes of students thinking aloud while completing “B vs. C” problems have recently been collected and these response processes will be subjected to analysis. In addition, criterion-related evidence of differential performance on these problems corresponding to the examinee’s level of statistical training should be explored.

At first it may appear negative that the majority of the students agreed both that having difficulty with the “B vs. C” type items made them question their abilities to competently analyze data, and, responding to the “B vs. C” type items had taught them that over reliance on a statistical consultant may be unwise. However, consider that this is my students’ first graduate course in applied statistics. Those in our doctoral programs are either required to take an additional applied statistics course, or, are expected to utilize the services of a professional statistics consultant. In either case, a healthy skepticism or “critical stance”(Gal, 2002) as to the adequacy and accuracy of one’s analysis is a disposition that bodes well for the professional practice of statistics (Chance, 2002), particularly in an era where meaningless statistical output can be generated in a matter of seconds and even meaningful output is subject to misinterpretation.

NOTE: Copies of grading rubrics for select problems, along with the actual statistical output and annotations as presented on the exam, may be obtained from the author (rhackett@pacific.edu).

REFERENCES

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (Vol. 24, pp. 61-100). Washington, DC: American Educational Research Association.

Broers, N. J. (2006). Learning goals: The primacy of statistical knowledge. In A. Rossman & B. Chance (Eds.), *Proceedings of the 7th Annual Meeting of the International Conference on Teaching Statistics.* Auckland, New Zealand: International Association for Statistics Education. Online: www.stat.auckland.ac.nz/~iase/publications/17/6G2_BROE.pdf.

Bude, L. (2006). Assessing students’ understanding of statistics. In A. Rossman & B. Chance (Eds.), *Proceedings of the 7th Annual Meeting of the International Conference on Teaching Statistics.* Auckland, New Zealand: International Association for Statistics Education. Online: www.stat.auckland.ac.nz/~iase/publications/17/6G3_BUDE.pdf.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3). Online: www.amstat.org/publications/jse/v10n3/chance.html)

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. Online: www.stat.auckland.ac.nz/serj.

Gal, I. (2002). Adults’ statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*, 1-51. Online: www.stat.auckland.ac.nz/~iase/cblumberg/gal.pdf.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3). Online: www.amstat.org/publications/jse/v10n3/garfield.html.

Zech, L., Vye, N., Bransford, J., Goldman, S., Barron, B., Schwartz, D., Hackett, R., & Mayfield-Stewart, C. and the Cognition and Technology Group at Vanderbilt. (1998). An introduction to geometry through anchored instruction. In R. Lehrer & D. Chazen (Eds.), *Designing learning environments for developing understanding of geometry and space* (pp. 439-463). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.