Developing Teachers' Reasoning about Comparing Distributions: A Cross-Institutional Effort

Dung Tran
Victoria University Melbourne
<dung.tran1@vu.edu.au>

Hollylynne Lee
NC State University USA
<hstohl@ncsu.edu>

Helen Doerr
Syracuse University USA
<hmdoerr@syr.edu>

The research reported here uses a pre/post-test model and stimulated recall interviews to assess teachers' statistical reasoning about comparing distributions, when enrolled in a graduate-level statistics education course. We discuss key aspects of the course design aimed at improving teachers' learning and teaching of statistics, and the resulting different ways of reasoning about comparing distributions that teachers exhibited before and after the course.

Statistics continues to leverage its roles in the curriculum standards documents in Australia and in the United States (ACARA, 2013; Common Core State Standards Initiative, 2010). Any effort to improve the learning of school statistics depends on professional development of teachers. However, the research base on teachers' statistical reasoning remains sparse (Batanero, Burrill, & Reading, 2011), and thus limits the capacity for teacher educators to design effective professional development experiences. Our research is situated within the context of a graduate course whose design and implementation was influenced by Pfannkuch and Ben-Zvi's (2011) recommendations for designing experiences to develop teachers' statistical thinking, the GAISE report (Franklin et al., 2007), and the Mathematical Education of Teachers II report (CBMS, 2012). The course, developed and implemented at two US institutions, engaged teachers through experiences as learners with the key concepts of distribution, variation, sampling, and informal inference and through explicit discussions of such learning experiences from a teacher's perspective. In this paper, we examine how teachers' reasoning about comparing distributions changed through their participation in the graduate course. In particular, this study addresses the questions:

(1) What aspects do teachers attend to when comparing distributions?

(2) How does the teachers' reasoning about comparing distributions change as they participate in the course on teaching and learning statistics?

Related Literature

A teachers' statistical reasoning is foundational to their ability to teach statistics (e.g. Groth, 2007; Lee & Hollebrands, 2011) and researchers underscore the importance of comparing distributions tasks to engage students in rich reasoning about contexts. Comparison of distributions offers opportunities for students to reason with key ideas in statistics including distribution, variability, sampling distribution, and statistical inference (e.g. Konold & Higgins, 2003). Studies have examined a range of students' thinking when dealing with comparison tasks, from elementary and middle school (Gal, Rothschild, & Wagner, 1989; Watson & Moritz, 1999) to high school (Estepa, Batanero, & Sanchez, 1999) and university-level students (Ciancetta, 2007). In these studies, students were asked to decide which group did better as well as what evidence they used to support their claims. While some studies presented graphical representations and/or numerical statistics for students to reason with, in other cases students worked in a technological environment to compare the data sets. A consistent finding across these studies reveals that students

2016. In White, B., Chinnappan, M. & Trenholm, S. (Eds.). Opening up mathematics education research (Proceedings of the 39th annual conference of the Mathematics Education Research Group of Australasia), pp. 608–615. Adelaide: MERGA.

generally have difficulties with these tasks and often fail to incorporate measure of centre in their reasoning. Students' performance is worse when tasks include distributions with similar measure of centre, but different variation and/or shape, and unequal sizes of groups. While younger students have difficulties because they often rely on context, individual data points, or a specific measure (e.g. mode) in making the comparison. The abilities to reason about comparisons increase as students incorporate more strategies as they get older (Gal et al., 1989; Watson & Moritz, 1999).

Another research trend is in the use of comparison tasks to examine students' understanding of other statistical concepts, such as variation (e.g. Ben-Zvi, 2004; Lee, Zeleke, & Wachtel, 2002) and distribution (e.g. Ciancetta, 2007; Cobb, 1999). Most of these studies were conducted in the context of teaching experiments with extensive focus on exploratory data analysis in a technological environment. The studies underscore a wide range of understandings of variation and the spread of abilities that students have in estimating and using various measures of variation. There exists a wide range of misconceptions, such as equating the range or other measures of variation (variance and standard deviation) with variation itself, and associating more data points with greater variation. Instead, with a well-designed learning environment, research shows a positive improvement in students' performance in these tasks. Students attend to more sophisticated understanding of variation towards global understanding of variation and distributions that incorporate both measure of centre and variation in their reasoning (e.g. Cobb, 1999). These results motivated the inclusion of rich contexts for distribution comparisons in technological based environments in the design of the graduate course.

Additional research focuses on teachers' reasoning with comparison tasks for both preservice (Makar & Confrey, 2005) and inservice teachers (Madden, 2008; Makar & Confrey, 2002). These studies have different foci such as documenting teachers' language and their informal inferential reasoning. This study adds to this research about teachers' reasoning about comparison of distributions and how these understandings change as they engage in a course focusing on teaching and learning statistics.

Methodology

Context

A team of four instructors from two institutions met weekly via videoconference for an academic year to design and plan a 15-week course. The course consisted of opportunities for teachers to engage in statistical investigations with real data, while using dynamic statistical tools (e.g. *Fathom, Tinkerplots*). There were several opportunities for teachers to examine both small and large data sets about which they posed and investigated questions about comparing groups. The teachers used software tools that enabled them to represent data in both dot plots and box plots and were able to overlay statistical measures such as the mean or median in the graphs. Across institutions, the course served 27 participants, 8 in Course1 and 19 in Course 2. Twenty-one participants were female and six were male, with six participants for whom English was a second language. Most participants had completed an equivalent of an undergraduate major in mathematics, with all but two having had at least one course in statistics. We refer to course participants as teachers.

Instruments and Analysis

We report on the data from two sources: 3 multiple-choice items related to distribution comparison from a pre- and post-test statistics concept test, and stimulated recall interviews after the completion of the course. On the first day of class and during the final week of the course, all participants completed a 20 item multiple-choice test with items in five categories: distributions (5 items), comparing distributions (3 items), probability (2 items), sampling variability (7 items), and formal inference (3 items). Items were drawn from validated instruments and adapted from research (e.g. delMas et al., 2007; Garfield, 2003; Watson & Kelly, 2004; Zieffler et al., 2008). The 20-item test was agreed upon by instructors during the planning phase to ensure items had content validity to measure concepts to be addressed in the course. The post-test was followed by semi-structured interviews with selected teachers (n=13) to understand changes in their reasoning and perceptions of what might have influenced those changes. Both quantitative analyses such as summary statistics and qualitative analyses with open coding were completed by the research team.

Results and Discussion

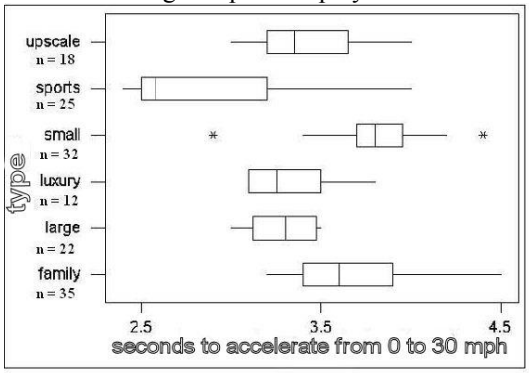
The category of comparing distributions included three items that assessed teachers' reasoning about differences between groups given graphical displays (boxplots or dotplots). Each item received a score of one for a correct answer; one item consisted of two parts, each of which was scored for a half point for a correct answer. The comparing distribution score mean improved by 0.26 points (S.D.=0.82) from the pre- to the post-test. Table 1 summarizes the performance of 27 teachers in the study.

Table 1
Summary Statistics for Pre- and Post-Tests

	Pre-Test Mean (S.D)	Post-test Mean (S.D)
Comparing Distribution Score (out of 3)	1.89 (0.73)	2.15 (0.82)

The teachers generally performed better on the boxplot items than on the dotplot item.

The 1999 Consumer Reports new Car Buying Guide reported on the number of seconds required for a variety of cars to accelerate from 0 to 30 mph. The cars were also classified in six categories according to type. The following box plots display these distributions:



Which statement provides the best reasoning for the type of car that tends to accelerate the slowest?

	PRE-TEST	POST-TEST
a. Family cars because the maximum value is 4.5.	3.7% (1)	0.0% (0)
b. Small cars because the median is greatest.	11.1% (3)	3.7% (1)
c. Small cars because it has both the greatest median and the smallest interquartile range.	**74.1% (20)**	**81.5% (22)**

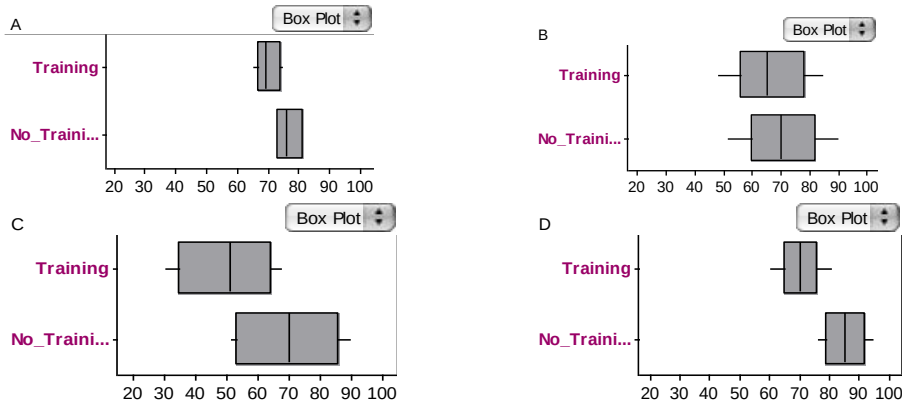
d. Sports cars because it has the smallest median.	0.0% (0)	3.7% (1)
e. Undetermined because one cannot compare different distributions with unequal numbers of observations.	11.1% (3)	11.1% (3)

Figure 1. Teacher Performance on Car Buying Guide Item on Pre- and Post-Test

In the Car Buying Guide item (Figure 1), stacked boxplots for the acceleration of six different types of cars were provided and teachers were asked for which car type has the slowest acceleration. The teachers performed quite well on this item on the pre-test, in which 74.1% of the teachers chose a correct option. This number increased by 7.4% on the post-test. Three teachers claimed the groups were not comparable on the post-test (two of whom also chose this option and one changed from a correct answer on the pre-test).

The Weight Training item has two parts with similar structures that provided four possible pairs of boxplots representing students' running times in two different training programs in a well-designed experiment, and asks students to decide which pair shows most convincing and which pair shows least convincing evidence that the training programs led to different results. Despite the similarity of the structures of two parts in the Weight Training item (see Figure 2), the teachers' performances on the two parts were dramatically different. For part I, 59.3% teachers selected a correct response on the pre-test. The performance was worse with less than half choosing the correct option on the post-test. Almost all teachers chose the correct option on the pre-test for part II (88.9%), and all but two teachers (92.6%) selected a correct option on the post-test. Further interviews explain the discrepancy in the performances in the two similar sub-items.

Suppose that there is a special summer camp for track athletes. There is one group of 100 athletes that run a particular race, and they are all pretty similar in their height, weight, and strength. They are randomly assigned to one of two groups. One group gets an additional weight-training program. The other group gets the regular training program without weights. All the students from both groups run the race and their times are recorded, so that the data could be used to compare the effectiveness of the two training programs.
Presented below are some possible graphs that show boxplots for different scenarios, where the running times are compared for the students in the two different training programs (one with weight training and one with no weight training). Examine each pair of graphs and think about whether or not the sample data would lead you to believe that the difference in running times is caused by these two different training programs. (Assume that everything else was the same for the students and this was a true, well-designed experiment.)



I. Which set of boxplots show the MOST convincing evidence that the weight-training program was more effective in DECREASING athletes' running times?

	PRE-TEST	POST-TEST
A	3.7% (1)	7.4% (2)
B	7.4% (2)	3.7% (1)
C	29.6% (8)	40.7% (11)

D	**59.3% (16)**	**48.1 (13)**
II. Which set of boxplots shows the LEAST convincing evidence that the weight-training program was more effective?		
A	3.7% (1)	3.7% (1)
B	**88.9% (24)**	**92.6% (25)**
C	3.7% (1)	0.0% (0)
D	3.7% (1)	3.7% (1)

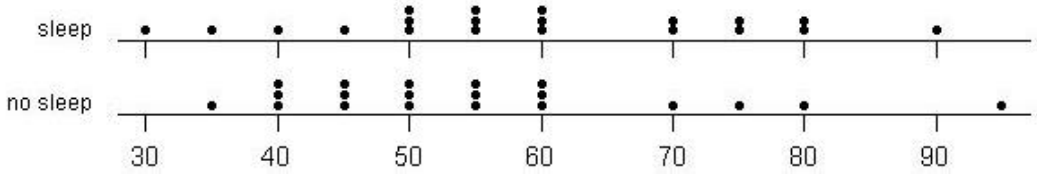
Figure 2. Teacher Performance of Weight Training Item on Pre- and Post-Test

The interviews show two lines of reasoning explaining for the teachers' selection. The teachers who attended to the overlap of the two boxplots to inform their decision in Part I identified the correct option (D). T11 said, "I am looking for boxplot of training to the left of the boxplot for no training, so I see 'A' has some improvement, 'B' I am skeptical, C looks possible, but 'D' jumps out of it because I see less overlap … here [pointing to 'D']." Others (nine of the) teachers in the interviews attended to medians in comparing these groups by estimating the difference between the two medians to decide which boxplots provided the most convincing evidence of the weight training. This type of reasoning helped most teachers choose the correct option in part II, in which the difference of the two medians is smallest in the most overlapped boxplots. This also explains why Part I appears to be much more difficult than Part II. Actually, with an incomplete reasoning (e.g. attending to merely the medians in comparison), the teachers could select a correct option in Part II. The interviews also help explain why more teachers chose an incorrect option on the post- compared with the pre-test: It was because the teachers could read the boxplots and knew how to estimate the medians by the end of the course.

When further prompted for possible challenges working on this item, some teachers revealed that during the exam they merely took a measure of central tendency (medians) into consideration. They simply compared the difference between two medians to get the smallest/largest value for evidence of least/most convincing of the weight training. Language was another challenge when teachers solved this item. Some teachers who selected the incorrect option misread the question by interpreting the word "plausible" as impossible. Still another challenge was the unfamiliarity with boxplots and six teachers acknowledged the course experience gave them many opportunities to engage with this type of display. "That is funny. Because during the pre-test I didn't have a slight idea what really a boxplot is." T1 said.

The dotplot item, Sleep Effect (see Figure 3), was challenging for the teachers even after extensive experience with comparing distributions using graphical displays in the course. Although the number of teachers who got the correct option improved by 22.3%, only 63% chose the correct option after the course. The teachers who identified incorrect options mostly attended to simply comparing means or how individual values of data might influence the mean of each distribution. No one selected the option that focused on an individual data point (a) and used the range alone to make a comparison (c).

Forty college students participated in a study of the effect of sleep on test scores. Twenty of the students volunteered to stay up all night studying the night before the test (no-sleep group). The other 20 students (the control group) went to bed by 11:00 pm on the evening before the test. The test scores for each group are shown on the graph below. Each dot on the graph represents a particular student's score. For example, the two dots above 80 in the bottom graph indicate that two students in the sleep group scored 80 on the test.



Examine the two graphs carefully. From the 6 possible conclusions listed below, **choose the one with which you most agree.**

	PRE-TEST	POST-TEST
a. The no-sleep group did better because none of these students scored below 35 and a student in this group achieved the highest score.	0.0% (0)	0.0% (0)
b. The no-sleep group did better because its mean appears to be a little higher than the mean of the sleep group.	3.7% (1)	3.7% (1)
c. There is no difference between the two groups because the range in both groups is the same.	0.0% (0)	0.0% (0)
d. There is little difference between the two groups because the difference between their means appears to be small compared to the amount of variation in the scores.	**40.7% (11)**	**63% (17)**
e. The sleep group did better because more students in this group scored 80 or above.	3.7% (1)	7.4% (2)
f. The sleep group did better because its mean appears to be a little higher than the mean of the no-sleep group.	51.9% (14)	25.9% (7)

*Figure 3.*Teacher Performance on Sleep Effect Item on Pre- and Post-Test

Seven of the interviewed teachers focused exclusively on the means when comparing the two distributions and ended up with an incorrect option. They estimated how the individual values influenced the mean of each distribution, and used them to compare the two groups. However, in the interviews teachers attended more to the variation of the data and used the spread of data for estimating and comparing means. Considering mean was a good indicator for comparisons, three teachers found it hard to estimate the difference of the means from the graphs. These teachers indicated a need for numerical computations to compare groups and they ended up with an incorrect option.

Six of the interviewed teachers, while attending to both means and variability, chose the correct option in the interview. These teachers considered whether the difference of the two means was significant and indicated a need for a simulation or randomization test to confirm the significant difference, which was similar to what they experienced in the class. T27 stated:

> I think the mean isn't enough to make the statement of the two groups, if it is significant enough… If we have enough information, you can do a randomization sample, and get the differences, and see if these differences are obviously likely, unlikely what happens over and over again.

The role of context in this problem seemed to both hinder and help the teachers in their decision-making. Two teachers referred to their own experience with testing when reasoning about this problem. "You want to mean something that I stay up all night [so my score is better] or you gave up and got some sleep… You don't want it to be D. It [sleep] matters", said T13. Personal experience was among the challenges the teachers identified as to why this item might be difficult. In addition, several teachers attended to only means or wanted the numerical values of the means to help make a decision. Additionally, the long wording of the item seemed to be tricky for some teachers, perhaps causing them to overthink the wording and intended meaning of the item as some teachers indicate.

Conclusion and Implications

We discuss three aspects: (a) students' reasoning in relation to the features of items, (b) the validity of multiple-choice items, and (c) how the learning experience changes their ways of reasoning regarding distribution comparison. First, the types of given graph seemed to influence the features that teachers attended to when reasoning about the comparisons. Most teachers attended to the measure of centre in the dotplot item without taking variation into account. By contrast, with the boxplot items, most teachers attended to the variation (the overlap of boxplots in this case), followed by the centre (the medians). Notwithstanding, for most teachers who made the wrong choice for the Weight Training item, they focused just on the magnitude of the difference of the medians to find the most convincing graph. This confirms previous findings (e.g. Watson & Moritz, 1999) that students did not consistently use the same type of strategy for comparing data sets across a series of tasks. Second, the exclusion of numerical statistics on the graphs might suggest another level of difficulty when reasoning about these problems. The teachers attended to the means for comparison, however they attempted to estimate means from the dotplots, which was not trivial. One teacher admitted that if the numerical statistics were provided, she might reason differently about the item. This finding corroborates the difference in students' performance with and without numerical statistics in Ciancetta's (2007) study on undergraduate and postgraduate students. Therefore, it is imperative to expose students to a variety of tasks features (graphical displays, numerical statistics) in such tasks.

Some teachers chose correct options without appropriate reasoning. For example, in the Weight Training item, teachers could get part II correctly by focusing merely on the difference between the medians. This finding leads us to question validity of the item, does it really measure what it aims to measure?

Understanding other key statistical ideas such as randomness, bias, sampling, and measures of centre is essential for making statistical comparisons (Franklin et al., 2007). Some teachers, who did not show evidence that they could coordinate both measures of centre and variability when comparing groups, failed to identify the correct option. However, the interviews demonstrated all teachers held more than a local view when comparing graphical displays of distributions (cf. Ciancetta, 2007). They were able to focus on measures of centre as representativeness to compare groups (Konold & Higgins, 2003). Some teachers moved beyond variation within distribution (variability of data in a single distribution) to variation between distributions (whether the difference between the numerical statistics is meaningful or significant) (cf. Makar & Confrey, 2004). Such advanced thinking could be associated with experiences they had in a semester-long course focusing exclusively on teaching and learning statistics and forms the basis for further professional development experiences. The question is how we, mathematics and statistics educators, can create such learning opportunities for preservice and inservice secondary teachers to develop understanding of teaching statistics.

References

Australian Curriculum, Assessment and Reporting Authority [ACARA]. (2013). *The Australian curriculum: Mathematics, version 5.0*. Sydney, NSW: Author.

Batanero, C., Burrill, G., & Reading, C. (Eds.) (2011). *Teaching statistics in school mathematics—Challenges for teaching and teacher education: A joint ICMI/IASE study*. New York: Springer.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, 3(2), 42-63.

Ciancetta, M. A. (2007). *Statistics students reasoning when comparing distributions of data.* (Doctoral of Philosophy), Portland State University. Retrieved from http://search.proquest.com/docview/304823182

ProQuest Dissertations & Theses database.
Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1*(1), 5-43.
Common Core State Standards Initiative (CCSSI). (2010). *Common core state standards for mathematics*. Retrieved from http://www.corestandards.org
Conference Board of the Mathematical Sciences [CBMS] (2012). *The mathematical education of teachers II*. Providence RI and Washington DC: AMS & MAA.
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.
Estepa, A., Batanero, C., & Sanchez, F. (1999). Students' intuitive strategies in judging association when comparing two samples. *Hiroshima Journal of Mathematics Education, 7*, 17-30.
Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., et al. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A Pre-K-12 curriculum framework*. Alexandria, VA: ASA.
Gal, I., Rothschild, K., & Wagner, D. A. (1989). *Which Group Is Better? The Development of Statistical Reasoning in Elementary School Children*. Paper presented at the Meeting of the Society for Research in Child Development, Kansas City, MO, USA.
Garfield, J. B. (2003). Assessing statistical reasoning. Statistics Education Research Journal, 2(1), 22-38.
Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, 38, 427-437.
Konold, C., & Higgins, T. L. (2003). Reasoning with data. In J. Kilpatrick, G. W. Martin & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193-215). Reston, VA: NCTM.
Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289. doi:10.2307/749741
Lee, H. S., & Hollebrands, K. F. (2011). Characterizing and developing teachers' knowledge for teaching statistics. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics—Challenges for teaching and teacher education*: *A joint ICMI/IASE study* (pp. 359-369). New York: Springer.
Lee, C., Zeleke, A., & Wachtel, H. (2002). *Where do students get lost? The concept of variation.* Paper presented at the B. Phillips (Chief Ed.), Developing a Statistically Literate Society: Proceedings of the Sixth International Conference on Teaching Statistics. Voorburg: The Netherlands (CD-ROM).
Madden, S. R. (2008). *High school mathematics teachers' evolving understanding of comparing distributions.* (Doctor of Philosophy), Westeran Michigan University. Retrieved from http://scholarworks.wmich.edu/dissertations/792/
Makar, K., & Confrey, J. (2002). *Comparing two distributions: Investigating secondary teachers' statistical thinking.* Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS-6), Cape Town, South Africa.
Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups *The challenge of developing statistical literacy, reasoning and thinking* (pp. 353-373): Dordrecht: Kluwer Academic Publishers.
Pfannkuch, M., Ben-Zvi, D. (2011). Developing teachers' statistical thinking. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics—Challenges for teaching and teacher education: A Joint ICMI/IASE Study* (pp. 323-333). New York: Springer.
Watson, J., & Moritz, J. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, 37(2), 145-168.
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.