Lessons from the LOCUS Assessments (Part 1): Comparing Groups

Douglas Whitaker and Tim Jacobbe

Since its publication, the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K–12 Curriculum Framework* has been influential in the field of statistics education. The developmental levels—A, B, and C—through which students are hypothesized to progress provide convenient touchstones for curriculum and lesson design.

Despite the impressive contributions to statistics education in terms of instructional recommendations and the aforementioned developmental progression, the GAISE framework says less about what types of assessments are recommended or should be considered as a model. The NSF-funded Levels of Conceptual Understanding in Statistics (LOCUS) project focused on developing statistical assessments in the spirit of the GAISE framework. These assessments emphasize conceptual (rather than procedural) understanding and can be used to classify students as having understanding at level A, B, or C.

The assessments—which will be available in August 2014—consist of four forms: a pre- and post-test targeting the A and B levels and a pre- and post-test targeting the B and C levels. The A/B assessment was designed for students in grades 6–9, and the B/C assessment was designed for students in grades 10–12. Two versions of these are available—one with 23 multiple choice items and 5 free response items and another with 30 multiple choice items only. The items from which these assessments were constructed were piloted in spring 2013 with a total of 2,075 students for the A/B assessment and 1,249 students for the B/C assessment. (Although every item was not piloted with every student, each item was piloted with several hundred students.) While the pilot administration sample was large and included students of many backgrounds and ability levels, it was not selected to be a representative sample of students in the United States. We do report some overall performance indicators, but these are included to paint a more complete picture of the students and item and should not be over-interpreted.

Student work can be a valuable resource for teachers. The size and scope of the LOCUS pilot assessments yielded considerable variation in student responses. While there were some 'textbook' correct answers, students also were able to demonstrate correct statistical reasoning in imaginative ways. Incorrect answers often illustrated specific misunderstandings and, if identified as such, can suggest areas for more attention.

The LOCUS free response item being examined here is shown in Figure 1. This item addresses the following Common Core State Standards (CCSS):

- 6.SP.2 "Develop understanding of statistical variability."
- 6.SP.5 "Summarize and describe distributions."
- 7.SP.3 "Draw informal comparative inferences about two populations."
- S-ID.1-3 "Summarize, represent, and interpret data on single count or measurement variable."
- S-IC.3 "Making inferences and justifying conclusions."

The "Analyze" and "Interpret" components of the GAISE framework at Level B also are addressed by this item.

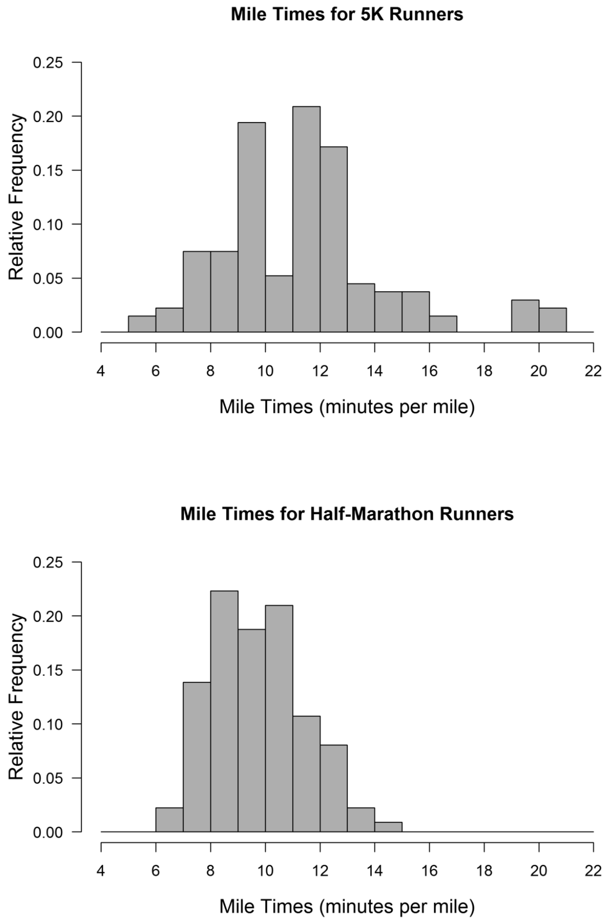

Figure 1: The city of Gainesville hosted two races last year on New Year's Day. Individual runners chose to run either a 5K (3.1 miles) or a half-marathon (13.1 miles). One hundred thirty four people ran in the 5K, and 224 people ran the half-marathon. The mile time, which is the average amount of time it takes a runner to run a mile, was calculated for each runner by dividing the time it took the runner to finish the race by the length of the race. The histograms show the distributions of mile times (in minutes per mile) for the runners in the two races.

Student Responses

This free response item was piloted with a total of 618 students in grades 9–12. Free response items were scored out of 4 points, with a 4 indicating a "complete" response (allowing for mistakes such as minor computational errors not indicative of a misunderstanding). For each item, a small team of graders established a rubric and conducted initial item scoring aloud as a group. Once every grader felt comfortable with applying the rubric, scoring continued individually. Any discrepancies or questions were brought to the group's attention. The distribution of students' scores for the item is given in Table 1: 4.8% earned a 4, 10.8% earned a 3, 33.8% earned a 2, 23.1% earned a 1, and the remaining 19.6% earned either a 0 or did not provide a response (these do not sum to 100% due to rounding).

Score	Percent of Students
4	4.8%
3	10.8%
2	33.8%
1	23.1%
0 or no response	19.6%

Table 1. The distribution of student scores for the item. (These do not sum to 100% due to rounding.)

Question related to free response item in Figure 1:

> *Jaron predicted that the mile times of runners in the 5K race would be more consistent than the mile times of runners in the half-marathon. Do these data support Jaron's statement? Explain why or why not.*

The first piece of this question related to Figure 1 is asking about how the mile times of 5K runners compare to the mile times of half-marathon runners—as a group. To answer the question correctly, students need to (1) recognize this is a question comparing groups and not individuals and (2) be able to correctly interpret histograms. Many students incorrectly focused on the heights of the bars representing relative frequency and concluded that variability in bar heights implies inconsistent mile times, as in this student response: "No, because there are more spikes in the graph for the 5K, and less in the graph for the half marathon."

Other researchers have discussed this misconception. Linda Cooper and Felice Shore, in a *Journal of Statistics Education* article, found that nearly 50% of students in their sample of 186 undergraduates judged variability in histograms by focusing on the heights of bars and attribute this to the visual similarity of histograms to bar charts and time-plots that use bars. The term "spread" is used often as a synonym for variability; while this word may seem accessible, students may be inclined to focus on the evenness of bars.

To score full marks on this question, a student must attend to some measure of the horizontal (rather than vertical) variability in the data. The measure of variability used need not be sophisticated. One student attempted to use the range with reasonable results: "No, the data in fact supports the opposite: that the half-marathon times are more consistent than the 5K times. The half-marathon times as you can see range from 6–15 minutes, whereas the range of 5K times is 5–21 minutes and 30 seconds, a much larger range (9 to 16.5)." This student also used the range, but made an implicit comparison: "No, the data doesn't support Jaron. The mile times for the 5K runners have a larger range and a higher standard deviation." Even though the student did not explicitly mention the half-marathon runners, language such as "larger" and "higher" indicate a comparison is being made.

Question related to free response item in Figure 1:

> *Sierra predicted that, on average, the mile time for runners of the half-marathon would be greater than the mile time for runners of the 5K race. Do these data support Sierra's statement? Explain why or why not.*

This question asks the students to directly compare the mile times for the two groups. Many students made appropriate arguments based on the mean or median: "No, the mean time for the half marathon is approximately about 9–10 minutes, where the 5K mean time is approximately about 11–12 minutes. And the 5K has some much higher times, which will increase the mean."

Of the students who did not compare the centers of the two groups, this response exemplifies a common misconception: "Yes, because the times stayed consistent during the 8–11 mile times." This student is focusing on the proportion of runners in the half-marathon group whose times were between 8 and 11 minutes. This approach attends to only part of the data in one group, does not make an appropriate comparison with the 5K runners group, and seems to be based on reasoning using the mode rather than a measure of center for quantitative data such as the mean or median.

Question related to free response item in Figure 1:

> *Recall that individual runners chose to run only one of the two races. Based on these data, is it reasonable to conclude that the mile time of a person would be less when that person runs a half-marathon than when he or she runs a 5K? Explain why or why not.*

The core component of this question is that the way runners were assigned to run either the 5K or half-marathon matters and has real implications for the conclusions that can be drawn from the data. As this student responded, "No, different races attract people of different abilities. Since these people ran their races voluntarily (they weren't randomly selected), nothing can be concluded." Although the student confused the terminology *random selection* with *random assignment*, they demonstrated a clear understanding that the runners chose either the 5K or half-marathon and that there could be a valid reason why a runner's mile time would not be less in a half-marathon than a 5K. Students were imaginative and many potentially valid reasons were given, all of which were scored as correct.

Some students correctly indicated concluding that mile times would be less when running a half-marathon than a 5K is inappropriate based on the given data, but their reasoning was incorrect: "No, because to do that you need to have an individual run both races to compare differences in time." While a matched-pairs design would work to answer the question at hand, it is not strictly necessary. This student's response ignores the lack of random assignment that could allow such a conclusion to be reached in a properly designed study.

In many cases, students can benefit from drawing on their experience and prior knowledge for answering questions. Other students, however, can be led astray by this if they rely on it too heavily: "Yes, if a person chose the half marathon over the 5K you can assume they are in better shape and are better runners than someone who would have chose to run less in the 5K." While such an assumption about the runners' abilities influencing their choice may turn out to be true in some cases, it is not appropriate to reach a conclusion on the basis of an untested assumption.

Discussion

This item broadly targets the "Analyze" and "Interpret" components of the GAISE framework and specifically targets several CCSS standards (6.SP.2, 6.SP.5, 7.SP.3, S-ID.1-3, S-IC.3). For students needing help with the content covered in this item, there are many resources available, including lesson plans on STEW (*www.amstat.org/education/stew*). Several lesson plans on STEW address the key aspect of comparing two groups using graphical displays (e.g., "How Long Is 30 Seconds?" and "Colors Challenge!"). Box plots are the graphical display used in these lesson plans, but histograms could be included and address the same CCSS standards (6.SP.4 and S-ID.1).

It is worth noting that, while this item was viewed favorably by the LOCUS development team and provided many interesting student responses, there were two complications that led to it not being included on the final version of the assessment. First, the histogram for the 5K runners has both a larger range (and other traditional measures of variability) and is 'bumpier' than the histogram for the half-marathon runners. Thus, when a student response included the half-marathon runners are more consistent with weak or confused justification, it was difficult to determine if the student was demonstrating a misconception about interpreting a histogram or simply not providing adequate justification.

Second, some students seemed to have difficulty with the concept of *mile time*—the measure used to compare the runners of races with different lengths in this problem. This presented confusion as to what was meant by a "greater" mile time: Does this mean larger in magnitude (slower) or better (faster)? The purpose of this question was not to test understanding of the mile time concept, but to test conceptual understanding of statistics. As such, it was replaced on the final forms of the assessment by questions that did not have these complications. The hundreds of student responses as written, though, still proved valuable.

Further Reading

Cooper, L. L., and F. S. Shore. 2008. Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education* 16(2). Retrieved from *www.amstat.org/publications/jse/v16n2/cooper.html*

Franklin, C., G. Kader, D. Mewborn, J. Moreno, R. Peck, M. Perry, and R. Scheaffer. 2007. *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-k–12 curriculum framework*. American Statistical Association: Alexandria, VA. Retrieved from *www.amstat.org/education/gaise*

Kaplan, J. J., D. G. Fisher, and N. T. Rogness. 2009. Lexical ambiguity in statistics: What do students know about the words association, average, confidence, random, and spread? *Journal of Statistics Education* 17(3):n3.

National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. *Common core state standards for mathematics*. National Governors Association Center for Best Practices and Council of Chief State School Officers: Washington, DC.

Shaughnessy, J. M. 2006. Student work and student thinking: An invaluable source for teaching and research. In *Proceedings of the Seventh International Conference on Teaching Statistics*. Retrieved from *http://iase-web.org/documents/papers/icots7/PL6_SHAU.pdf*