The Challenges of Measuring Misconceptions in Middle Grades Statistics

Jessica Masters and Lisa Famularo

Measured Progress Innovation Lab

ABSTRACT

The Diagnostic Statistics Assessment (DSA) project is developing a formative, cognitively diagnostic assessment system designed to measure three misconceptions in middle grades statistics. The DSA consists of distracter-driven items, diagnostic score reports, and instructional activities designed to help students reorganize their understandings. The DSA researchers have conducted item development and item revision based on expert review and cognitive labs. Based on these activities, the authors describe the challenges and lessons learned in developing items to target statistical misconceptions at this grade level.

PROJECT OVERVIEW

Purpose: Develop a formative, cognitively diagnostic assessment system for middle grades teachers to identify and address three misconceptions their students hold in statistics.

Misconceptions: The DSA targets three misconceptions grounded in cognitive science research. The authors use the term misconception broadly to reflect the idea that misconceptions often represent knowledge that is productive in some contexts but is underdeveloped or overgeneralized.

- Datasets as Entities: Students do not perceive a dataset as a single, unified entity with its own characteristics, unique from those of its individual data points.
- Comparing Datasets: Students do not compare datasets based on representative or summary measures, but rather based on single data points, cutoff points, or slices.
- Overreliance on the Mean Procedure: Students rotely apply a procedure to calculate a measure without the corresponding conceptual understanding of that measure.

Components of the DSA:

- Parallel pre- and post-tests consisting of small sets of distracter-driven, closed-response items targeting a single misconception
- Instant diagnostic feedback reports for teachers
- Instructional resources that contain tutorials for teachers and classroom activities

Research and Development: The DSA items will undergo multiple iterative rounds of research and revision.

- Expert review to collect evidence of validity based on test content
- Cognitive labs using a "think aloud" protocol to collect evidence of validity based on response processes.
- Field test to collect evidence of validity based on internal structure and the relationship to other variables

SESSION PURPOSE

Throughout the first two phases of item development and revision (expert review and cognitive labs), the researchers encountered challenges unique to the measurement of statistical knowledge, particularly in the middle grades. Four of these challenges are presented, along with recommendations for action. The goal is to inform the development of future assessments and instructional materials related to statistics in the middle grades.

CHALLENGE: AMOUNT OF CONTEXT

How much information should be provided in order to create a real-world context without obscuring the statistical construct being measured?

The phrase "solve real-world and mathematical problems" is used repeatedly throughout the Common Core. The use of context in mathematics assessment is intended to both improve students' ability to demonstrate their understanding of mathematics and to measure a deeper understanding of mathematical principles. Context is particularly important in statistics. Decisions about data collection, analysis of data, and the interpretation of results are heavily dependent on the research questions which are grounded in a particular context or situation.

Choosing an appropriate level of context for an item can be particularly challenging when the goal is to measure the more mathematical aspects of statistical knowledge. Asking the student to find the mean of a list of numbers encourages students to apply the mean procedure in a vacuum. Students should be encouraged to consider the context of the situation and the data to evaluate whether the mean is an appropriate measure. But the inclusion of additional contextual information might place an unnecessary burden on students that could be distracting and/or lead to an inaccurate assessment of their abilities.

How much data and information to provide is also important and needs to be made with careful consideration of the grade level of the students. It is important to find a balance in providing enough data so that students can find the answers they need, but not too much data to overwhelm or confuse them. All possibilities and "what if" questions that might be asked by a statistician might not be considered by an average middle school student. They might, however, be considered by an above average student!

Item Version 1: This table shows the numbers of diners at a restaurant each day for one week.

Which of these best represents a typical number of diners per day?

Feedback from Experts: The item does not provide a real world context. Why would one want to know the typical number of diners per day?

Item Version 2: This table shows the numbers of diners at a restaurant each Friday for the last seven Fridays.

The manager is thinking about whether to hire a new waiter for Fridays. To help her decide, she wants to summarize the data with a single value that best represents how many diners come in on Fridays. Which number should the manager use?

Feedback from Cognitive Labs: The statement "the manager is thinking about whether to hire a new waiter for Fridays" seems irrelevant or could be misinterpreted. For example, , one student considered how the number of waiters depends on whether the people who come to the restaurant come in as individuals, couples, or families (because, for example, there is a different efficiency in serving 5 people across 5 tables versus 5 people at 1 table).

Item Version 3: This table shows the number of customers at a restaurant for seven Fridays in the summer.

The manager wants to know how many customers there are on a typical summer Friday. She will use this information to decide how much food to have in the kitchen.

What number best represents how many customers come in on Fridays?

Item Version 1: The tallest 100 skyscrapers in the United States range from 733 feet to 1776 feet tall. The median height is 1,020 feet.

A journalist is writing an article about a new skyscraper being built that will be 1000 feet tall. He wants to compare the height of the new building to the height of the existing skyscrapers.

Which statement should he use to describe the height of the existing skyscrapers?

Feedback from Experts: It is not clear why any of these comparisons will serve a purpose. It will depend on the intention of the journalist. Why does the height matter?

Item Version 1: At an aquarium a veterinarian took the temperature of a sea lion every night at 8pm for a month. This graph shows his results.

What is most likely the sea lion's **actual** body temperature?

Feedback from Experts: In an effort to make the item accessible to middle-school students, the item writer used the term *actual* body temperature as opposed to *estimated* body temperature or *true* body temperature. An expert reviewer noted that students who excel in science understand that there is a range in most mammals' temperatures and, thus, will be drawn to distracters that include a range. These students will be confused by the idea of an "actual" temperature.

Item Version 2: At an aquarium, a veterinarian weighs a sea lion at 3pm every day for 4 weeks. This graph shows her results.

The veterinarian wants to give her best estimate of the sea lion's weight. Based on this data, which estimate should she give?

Feedback from Experts: Specify that the sea lion is an adult and add text to indicate that the veterinarian knows that the sea lion is not gaining or losing weight and as such has a true weight hidden by minor day to day fluctuations.

Item Version 3: A veterinarian at an aquarium has to write a report each month that includes the weight of a sea lion named Chuck and how much food Chuck ate each day. During the month of February, the veterinarian weighed Chuck at the same time every day.

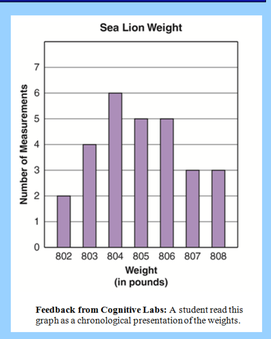
The veterinarian records that Chuck ate 20 pounds of food each day. Next, she has to determine the best estimate of Chuck's weight. Based on this data, which estimate should the veterinarian write in her report?

Feedback from Cognitive Labs: Students were confused about the eating of 20 pounds of food. This information threw students off. Why include how much Chuck ate? Seems like unnecessary information.

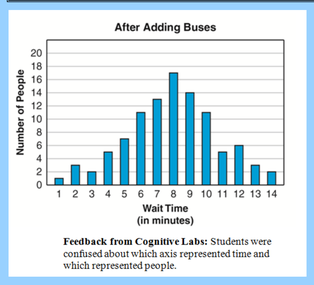
CHALLENGE: USE OF GRAPHS

How can statistical knowledge be measured without being confounded by the use of graphs and histograms?

The assessment of statistical knowledge often requires students to analyze data. Most meaningful datasets will be large enough to require a graphical presentation, as opposed to a list of raw values. Thus, students must be able to interpret the graph in order to access the data required to respond to the statistical inquiry. If a student is unable to interpret the graph (e.g., does not understand what the axes represent, incorrectly reads the scale, etc.), the student will not be able to successfully address the statistical question. Thus the ability to interpret a graph or histogram can serve as a barrier to the demonstration of statistical knowledge. Histograms are historically the most challenging type of graph to interpret but avoiding them places further restrictions on selecting meaningful contexts.



Feedback from Cognitive Labs: A student read this graph as a chronological presentation of the weights.

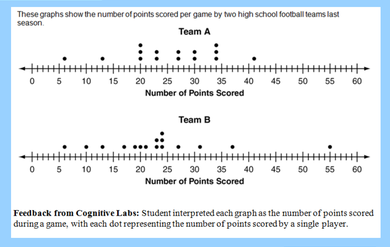


Feedback from Cognitive Labs: Students were confused about which axis represented time and which represented people.

Spinner Results

Spin Number	Color
1	Blue
2	Red
3	Blue
4	Green
5	Yellow
6	Red
7	Red
8	Red
9	Blue

Feedback from Cognitive Labs: Students thought the spin number was the number of *times* that color was spun.



Feedback from Cognitive Labs: Student interpreted each graph as the number of points scored during a game, with each dot representing the number of points scored by a single player.

CHALLENGE: MEANINGFUL, AGE-APPROPRIATE, AND REALISTIC CONTEXTS

The context must be meaningful for all students who will complete the assessment. Certain topics might be familiar to most students, but if they cause issues for one group, they cannot be used. For example, the context of an allowance is fairly common for many students, but is potentially unfamiliar to students not born in the U.S. Certain contexts might introduce bias or sensitivity issues or may make students particularly vulnerable to distraction because of personal circumstances, which can cause stress and affect students' ability to respond to the item (e.g., frequency of medical conditions, car accident data, weight loss results, etc.).

Contexts need to be grade-level and age-appropriate. Many situations that provide rich datasets are unfamiliar to middle-grade students. Even when a meaningful and age-appropriate context is found, it can be difficult to fully describe the context or data in language that is age-appropriate.

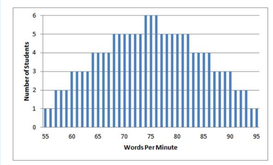
Contexts and datasets need to be realistic. One particular challenge in using real datasets or subsets of data that mimic real datasets is the *messiness* of the data. In an assessment, it is often desirable to have *clean* data so that students can apply statistical procedures with a clear result. But data in the real world is always noisy! Introducing noise will often result in data that is skewed, bimodal, or has outliers, which can change which analysis is appropriate. It is also a challenge to find or create a dataset with specific properties. For example, it might be desirable to have a dataset that is somewhat normally distributed but with a different mean and mode.

Items that use real world contexts often ask the student to put themselves in a position of an adult who would realistically be asked to use data to make a decision. Cognitive labs revealed that putting the student in the role of an adult decision maker increased the cognitive load thus possibly preventing the student from demonstrating his/her knowledge. However, there are not many meaningful contexts for which middle school students *truly* get to act as a decision maker!

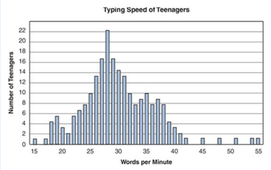
A final challenge in finding meaningful and realistic contexts is avoiding the use of decimals in datasets. Cognitive labs revealed that students' difficulties with decimals can be a confounding factor when they are making statistical calculations. But finding meaningful datasets that use only whole numbers is a challenge.

Terms that introduced difficulty in otherwise accessible contexts:
cell phone case
commute
conducted
course grade
diner (confused with dinner)
GPA
household income
nutritionist
policy
quality control
yearly income
yearly salary

Data Version 1:



Data Version 2:



Item: The principal asked 32 students to rate a new school cell phone policy on a scale of 1–10. This graph shows the results of the survey.

The principal wants to summarize the results to understand how this group of students feels about the new policy. Which statement provides the best summary of the data?

Item: A city recently added 5 new buses on one route. A group of 100 people reported how long they waited for a bus before and after the city added the new buses. The graphs below show the results.

The mayor wants to know whether buying the new buses was a good use of money. Which statement do you think best helps her decide?

CHALLENGE: STATISTICAL LANGUAGE

How can statistical knowledge be measured without being confounded by statistical language?

Because the Common Core has placed a greater emphasis on statistics and probability than previous standards and most of the statistical content is introduced in the Grades 6-8 standards, many middle grades students will be introduced to statistics for the first time. And many of their teachers will be teaching it for the first time. DSA cognitive labs revealed that there is a misunderstanding of even basic statistical language, including words like "data" and "value." More complex words such as "represents" and "summarize" were even more problematic. This challenge was more pronounced for ELL students, but was evident in most students. This kind of language is critical to conveying statistical ideas and questions.

Another challenge stems from use of the term *average*. Is *average* synonymous with *mean*? Or could it also mean *median* or *mode*? This issue is further complicated by the fact that *average* is both a mathematical and everyday term. Many experts recommend not using the term average and rather using the statistical term for what measure is required. However, what if the assessment seeks to determine whether students can choose which measure should be used?

Cognitive labs revealed difficulties with the following terms:
center
data value
dataset
distributed
independently
mean/median as an adjective
on average
represent
summarize
value

RECOMMENDATIONS

- When assessing statistical knowledge, provide students with context, even when targeting the more mathematical aspects of statistical knowledge. Context should include the statistical question being asked, not just a description of what the numbers represent. This is a better reflection of how students are encouraged to apply statistical procedures.
- In the classroom, instruction should focus on meaningful applications of mathematical processes within statistics. Assessments should mirror that instructional approach.
- Item writers must strike a balance between providing enough contextual information to create a realistic but not overly complex problem situation. Scenarios should support, not distract from, the statistical concepts.
- Items writers must understand more than the knowledge, skills, and abilities being assessed. They must have a strong understanding of what scenarios and language are developmentally appropriate for students at the targeted grade level.
- Item writers need feedback from various sources to find the appropriate balance of item context. Sources should include both statistical experts and students at varying levels of proficiency.
- Teachers should begin to use appropriate statistical terms in early grades and use these terms consistently throughout late elementary and middle school. Assessments can then reflect the language used in the classroom. In the interim, caution should be employed on the use of seemingly obvious statistical language in assessments.
- Careful attention should be paid to the amount of information, layout of information, and language used in labeling graphs. Keys and scales should be included to help students understand and interpret graphs. When possible, similar layout and formatting should be used for all items within a single assessment.

This material is based upon work supported by the National Science Foundation under Grant No. 1312133.