



Assessments and Rubrics

Dr. Laura Ziegler

IOWA STATE UNIVERSITY
Department of Statistics

Outline

- > Assessments
- > Rubrics
- > Lessons learned and advice

Assessment Experience

- > Basic Literacy In Statistics (BLIS) Assessment
 - Dissertation work
 - Assessment with 37 items
 - 3 expert reviewers
- > Statistics Learning Map Project (STAT-LM)
 - Large assessment with 100 items
 - 4 item writers, 3 content specialists, 3 bias/sensitivity reviewers, 1 editor
- > AP[®] Statistics Question Leader
 - Assist with Rubric Development
 - *AP[®] is a trademark registered by College Board, which is not affiliated with and does not endorse this presentation or Iowa State University's events, programs, etc.*

Assessment Types

- > One GAISE recommendation is to “Use assessments to improve and evaluate student learning” (2016)
- > Teaching
 - Improve: Formative assessment
 - Quizzes, homework assignments
 - Evaluate: Summative assessment
 - Exams, final course projects
- > Research
 - Focus on evaluation

Test Blueprint

- > Guides the assessment development
 - > “Responsibility for the process of learning cannot be shared unless the way to assess its’ quality is authentic, public, clearly understood, and accepted by the students.” (Cobb, 1993)
- > Includes topics and learning outcomes
- > Example:

Topic	Learning Outcome
Data	Understanding of the difference between a sample and population
Production	Understanding that statistics vary from sample to sample
Confidence	Understanding that a confidence interval provides plausible values of the population parameter
Intervals	Understanding of how the confidence level affects the width of a confidence interval

Test Blueprint

- How do you choose the learning outcomes?
 - “Students will value what you assess; therefore, assessments need to be aligned with learning goals.” (GAISE, 2016)
 - “Assessment should measure what is worth learning, not just what is easy to measure.” (MAA Instructional Practices Guide, 2018)

Test Blueprint Development Process

> Test Blueprint

- Create draft
- Expert review
- Edit
- Repeat as needed



Assessment Development Process

> Assessment

- Determine the type of items (forced-choice, free-response)
- Draft items or pull from existing banks/assessments
- Expert review
- Pilot items making edits after each pilot
 - Think-aloud Interviews with students
 - Small-scale pilot
 - Large-scale pilot



Example Item from BLIS

- > Topic = Data Production, Learning Outcome = Understanding of the difference between a sample and population
- > First Draft:
 - The Pew Research Center surveyed a nationally representative sample of 1,002 adults in 2013. The sample percent of internet users that have had an email or social networking account compromised was 21%. Identify the sample and population you would like to make inferences about.



Example Item

> Second Draft:

- The Pew Research Center surveyed a nationally representative sample of 1,002 U.S. adults in 2013. Of these adults, 21% have had an email or social networking account compromised. ~~The sample percent of internet users that have had an email or social networking account compromised was 21%.~~ Identify the ~~sample and population about which the Pew Research Center can you would like to~~ make inferences from the survey results and the sample from that population ~~about.~~



Grading Free-Response Items

- > How will responses be scored?
 - Correct/incorrect
 - Holistic score (e.g., essentially correct, partially correct, incorrect)
 - Numerical score
- > Scale
 - Too large of a scale = more chance for error in grading, more student complaints
 - Too small of a scale = not representative of student knowledge, misleading to students on their understanding

Rubrics

- > Guidelines for scoring a student response
- > Rubrics include:
 - Model solution
 - Rules for scoring
 - Additional notes (e.g., clarifications to rules)
- > Uses:
 - Provides consistency in scoring
 - Makes scoring transparent to students

Example Rubric

> This rubric is modeled after the AP[®] Statistics Exam Rubrics.

Question: The Pew Research Center surveyed a nationally representative sample of 1,002 U.S. adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population.

Model Solution: The population is all U.S. adults in 2013. The sample is the 1,002 U.S. adults surveyed.

Scoring Guidelines

Essentially correct (E) if the response satisfies the following two components:

1. Identifies the population in context
2. Identifies the sample with sample size in context

Partially correct (P) if the response satisfies only one of the two components.

Incorrect (I) if the response does not meet the criteria for E or P.

Additional Note:

- Indicating the population is “all adults” is not sufficient to satisfy component 1.

How is a Rubric Developed?

- > Draft (use existing examples when possible)
- > Expert review
- > Edit
- > Practice using student responses
- > Edit

How is a Rubric Implemented?

- > In a classroom setting:
 - If it is only one person scoring, simply follow the guidelines defined in the rubric.
 - If multiple people are scoring different student responses, they should work collaboratively to ensure consistency.
 - Share the rubric with your students. This will help students:
 - understand the requirements,
 - understand where they lost credit, and
 - learn from their mistakes.
- > In a research setting:
 - Separately, at least 2 graders score the responses
 - Compare the scores for the 2 graders
 - If scores do not match, graders discuss until they reach a consensus

Example: Furball Farm Cat Rescue

> Stem:

Furball Farm Cat Rescue is an animal shelter in Central New Jersey that takes in stray cats and kittens picked up from animal control and the local community, feeds them, gives them shelter and medical care as needed, and attempts to get them adopted. They have space and resources to care for exactly 150 cats at any one time. Any more than that will result in overcrowding and can be dangerous for the cats. Naturally, they rely on cats getting adopted regularly so they continue to have room to safely admit the new cats that arrive.

Example: Furball Farm Cat Rescue

- > Topic: Inference
- > Learning Outcomes:
 - Understand how to construct a one-sample confidence interval for a population mean
 - Be able to interpret a one-sample confidence interval for a population mean
- > Original Question-Part a:

Over the last 52 weeks, the average number of cats and kittens adopted in any given week is 17.5, with a standard deviation of 8.4. Construct and interpret a 95% confidence interval for the true mean number of cats adopted per week at Furball Farm. You may assume each week's number of adoptions are independent of other weeks.

Example: Furball Farm Cat Rescue

> Final Question-Part a:

Over the previous 5 years, 20 weeks were randomly selected. In those 20 weeks, the mean number of cats and kittens adopted per week was 17.5, with a standard deviation of 8.4. The distribution of the number of cats and kittens adopted per week was approximately normal. Construct and interpret a 95% confidence interval for the true mean number of cats adopted per week at Furball Farm. Assume all necessary conditions for inference have been met.

Example: Furball Farm Cat Rescue

> Model Solution:

The appropriate procedure is a one-sample t -interval for the population mean number of cats and kittens adopted per week.

The necessary conditions for inference were stated to be met.

Therefore, a 95% confidence interval for the population mean is given by

$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 17.5 \pm 2.093 \frac{8.4}{\sqrt{20}}$ with $df = n - 1 = 20 - 1 = 19$, which is 17.5 ± 3.931 , and the interval is (13.569, 21.431).

We can be 95% confident that the mean number of cats and kittens adopted per week in all weeks in the 5 past years is between 13.569 and 21.431.

Example: Furball Farm Cat Rescue

- **Essentially correct (E)** if the response satisfies the following four components:
 1. Identifies the appropriate procedure as a one-sample t -interval by name or formula or by the calculations of the correct confidence interval endpoint values
 2. Reports the values for a correct interval consistent with the procedure stated in component 1
 3. Indicates 95% confidence and interprets the interval using words such as “we are 95% confident” or “with 95% confidence” and provides interval endpoint values consistent with calculations in component 2
 4. Conveys inference about a population mean in the proper context, i.e., “the true mean number of cats and kittens adopted per week,” “the population mean number of cats and kittens adopted per week,” or “the mean number of cats and kittens adopted per week in all weeks in the 5 past years”

Example: Furball Farm Cat Rescue

- **Partially correct (P)** if the response satisfies either component 3 or component 4 and at least one of the other three components.
- **Incorrect (I)** the response does not meet the criteria for E or P.

Example: Furball Farm Cat Rescue

> Additional Notes:

- If the response includes supporting work for calculating the confidence interval that displays a correct formula with correct values inserted for \bar{x} , s , n , and t , then component 2 is satisfied even if values for the endpoints of the confidence interval are not displayed or calculated incorrectly.
- Clear indication of an inference to the sample of 20 weeks, rather than the population of all weeks in the past 5 years, does not satisfy component 4.

Example: Furball Farm Cat Rescue

- > Topic: Inference
- > Learning Outcome:
 - Understand the purpose of a one-sample confidence interval for a population mean
- > Final Question-Part b:

The shelter is currently full. Further, the shelter director has gotten a call that local animal control has 10 cats to drop off at Furball Farm at the end of the week. Assume that these are the only cats to be dropped off this week. Based on this information and your answer to part (a) above, can the shelter director conclude that because 10 is below their confidence interval, they do not need to be concerned (at least at the 95% level of confidence) that the addition of 10 new cats will put the shelter above capacity after the week's adoptions?

Example: Furball Farm Cat Rescue

> Model Solution:

The purpose of a confidence interval is to estimate a parameter. In this case, the interval is estimating the mean number of cats and kittens adopted per week in all weeks in the 5 past years.

A confidence interval cannot estimate an individual case, such as the number of cats dropped off in one week. Therefore, the shelter director cannot conclude that they do not need to be concerned (at least at the 95% level of confidence) that the addition of 10 new cats will put the shelter above capacity after the week's adoptions.

Example: Furball Farm Cat Rescue

- > **Essentially correct (E)** if the response satisfies the following three components:
 1. Indicates the purpose of a confidence interval is to estimate a parameter
 2. Indicates a confidence interval cannot estimate an individual case
 3. States the shelter director cannot make the conclusion in context

- > **Partially correct (P)** if the response satisfies only two of the three components.

- > **Incorrect (I)** the response does not meet the criteria for E or P.

Lessons Learned

- > Not easy!
- > Good assessments take a long time to create
- > Working with others is necessary
- > Interviews with students can be very informative
- > Just because others have used a particular assessment does not mean it is appropriate for you.



Advice

- > Consider your
 - Purpose
 - Goals
 - Learning outcomes
- > Find a colleague to collaborate with
 - Brainstorm ideas
 - Take each other's exams



Questions?

> Thank you!

Contact

Laura Ziegler

Teaching Professor

lziegler@iastate.edu

<https://www.stat.iastate.edu/>

References and Resources

- > Abell, M., Braddy, L., Ensley, D., Ludwig, L., & Soto, H. (2018). *MAA instructional practices guide*. Mathematical Association of America. <https://maa.org/programs-and-communities/curriculum%20resources/instructional-practices-guide>
- > American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME] (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- > *AP Statistics Exam questions*. AP Statistics Exam Questions – AP Central | College Board. (n.d.). <https://apcentral.collegeboard.org/courses/ap-statistics/exam/past-exam-questions>
- > Chance, B. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3). doi: 10.1080/10691898.2002.11910677
- > Cobb, G. W. (1993). Reconsidering Statistics Education: A National Science Foundation Conference. *Journal of Statistics Education*, 1(1). doi: 10.1080/10691898.1993.11910454

References and Resources

- > GAISE College Report ASA Revision Committee, “Guidelines for Assessment and Instruction in Statistics Education College Report 2016,” <http://www.amstat.org/education/gaise>.
- > Garfield, J. (1994). Beyond testing and grading: Using assessment to improve student learning. *Journal of Statistics Education*, 2(1). doi: 10.1080/10691898.1994.11910462
- > Garfield, J., delMas, R., & Chance, B. (2002). The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project. NSF CCLI grant ASA- 0206571. [Online: <https://app.gen.umn.edu/artist/>]
- > Garfield, J., & Gall, I. (1999). Assessment and Statistics Education: Current challenges and directions. *International Statistical Review*, 67(1), 1-12. doi: 10.1111/j.1751-5823.1999.tb00377.x