AN ASSESSMENT TO EVERY PURPOSE, UNDER HEAVEN

MARIANNE PERIE, UNIVERSITY OF KANSAS

DECEMBER 15, 2017

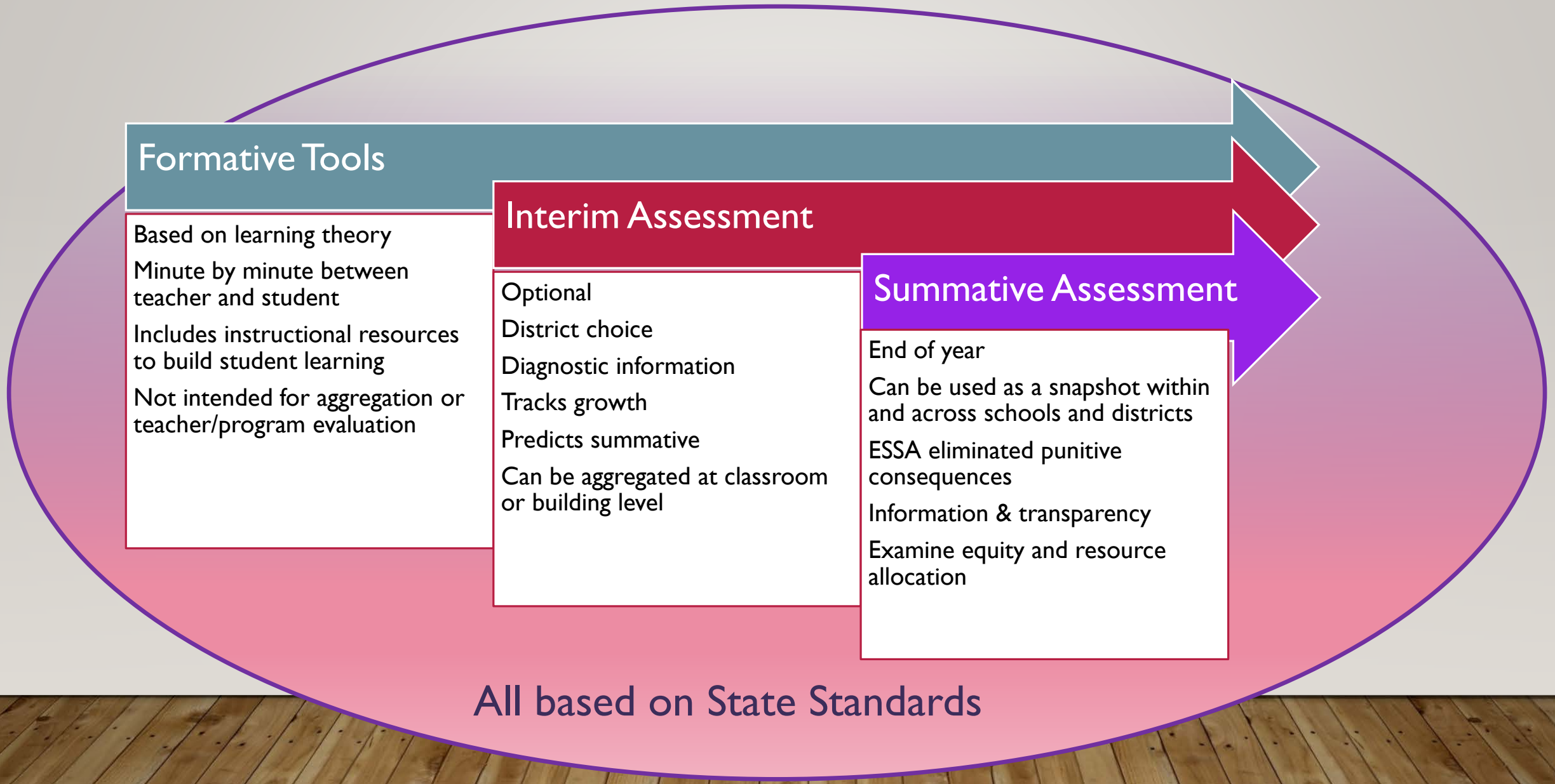
To everything (turn, turn, turn)
There is a season (turn, turn, turn)
And a time to every purpose, under heaven

—the Byrds

AN ASSESSMENT FOR EVERY PURPOSE

- What is an interim assessment?
  - How does it differ from formative assessment
- How does it fit into a balanced assessment system?
- What are the necessary components?
- How should one build or evaluate an interim assessment?

BALANCED ASSESSMENT SYSTEM



SUMMATIVE

- Typically statewide and used for accountability purposes
- Largest grain size of all assessments
- Aligned to state content standards
- Incudes targets of performance
- Requires sufficient items for reliability
- Can include selected response or open ended items but hand-scoring increases costs

FORMATIVE

- Black and Wiliam wrote seminal piece in 1998 describing how formative assessment can improve student learning within a classroom.
- *"An assessment is formative to the extent that information from the assessment is used, during the instructional segment in which the assessment occurred, to adjust instruction with the intent of better meeting the needs of the students assessed."*
- Key components
  - Integrate seamlessly with curriculum
  - Provide timely, corrective feedback
  - Include both diagnostic and prescriptive feedback (i.e., what should teachers do next)
- Although they are optional, teachers really use them every day to assess student learning. Resources exist to formalize the process.

FORMATIVE AND INTERIM ARE NOT THE SAME

- Back when we wrote Perie, Marion & Gong (2009), companies were building multiple-choice tests to be delivered during the school year on computer and calling them "formative.
  - When these tests were called "formative" the research on formative assessment was cited to support their use
  - Our paper was intended to clarify this type of assessment as being distinct from both formative and summative.
- Collected benchmark, diagnostic, and wrongly-named formative under one umbrella, called interim Some assessments may help form instruction without meeting all the criteria for formative assessment
- Interim assessment some requirements for formative by
  - Providing *qualitative* insights about understandings and misconceptions not just a numeric score
  - Giving timely feedback on what to do besides re-teaching every missed item

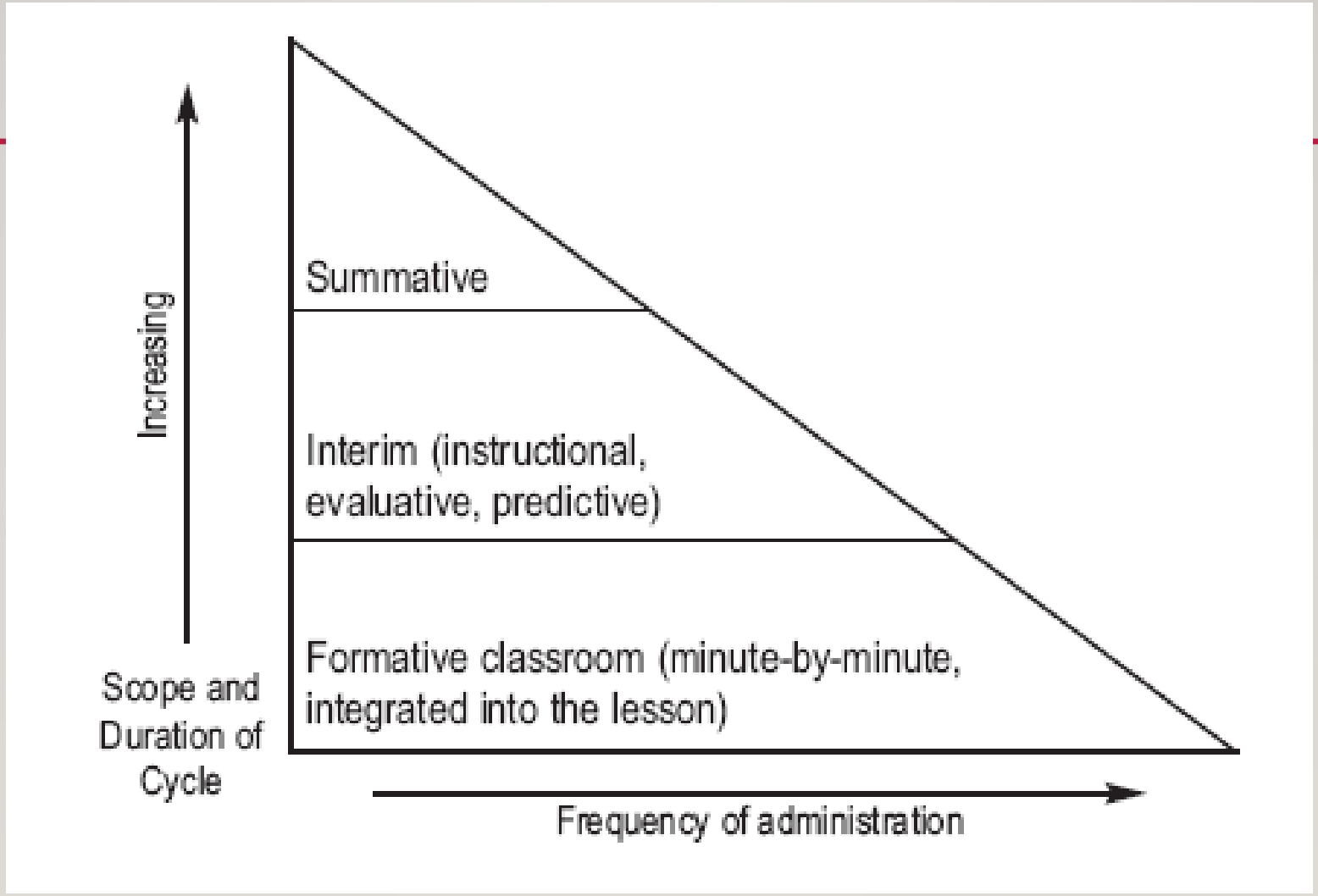
INTERIM

- Allows teachers to check student progress throughout the year
  - Instructionally, gives teachers and students information they can use to identify gaps in knowledge or misconceptions
  - Can be used to measure growth within a school year
  - Often linked to summative assessment for predictive purposes
- Interim tools are used at the discretion of schools and districts, and teachers can employ them to check students' progress at mastering specific concepts at strategic points during the school year.

INTERIM (CONTINUED)

- Typically used at the district level
- Needs to provide valid and reliable results that are easy to interpret
- Must include a rich representation of content with items linked to content standards and specific teaching units
- Should be an extension of learning not a "time-out" from learning
- Item types similar to summative will help with predictive purpose, but more open-ended probes and performance tasks assist with instructional purpose

PERIE, MARION, & GONG (2009)



BEFORE WE GO TOO FAR…

- …Let's address the elephant in the room: Overtesting
- Parents and teachers have pushed back about the amount of testing, but where does that overtesting occur?
  - Typically, summative assessment is blamed.
  - However, only 8-10 hours per year is spent on summative assessment, typically.*
  - Formative assessment is part of instruction and not typically called out as a "test."
  - Interim assessment is where the bulk of testing time occurs.
- So, if we are going to use an interim assessment, let's be judicious and targeted.

**Based on a survey done in 2016 by the Council of Chief State School Officers*

DESIGNING OR SELECTING AN ASSESSMENT

- First, determine goal/purpose of the assessment
  - Provide information on student performance relative to some target
  - Sort students, classrooms, or schools
  - Identify achievement gaps among student groups
  - Provide instructional feedback
  - Evaluate instruction or instructor
  - Predict performance on a future activity
  - Track growth over time
- Be judicious: An assessment purporting to serve multiple purposes serves no purpose well.

CONSIDER THESE QUESTIONS…

- What do I want to learn from this assessment?
- Who will use the information gathered from this assessment?
- What action steps will be taken as a result of this assessment?
- What professional development or support structures should be in place to ensure the action steps are taken?

MATCH TEST CHARACTERISTICS TO THE PURPOSE

- Match item types to the purpose
  - Similar to summative?
  - Provide additional insight into student understanding?
- Length and frequency
  - Predictive should be closer to the time of summative
  - Growth may need to be at set intervals
  - Instructional requires more flexibility

RELIABILITY

- Reliability refers to the consistency of results. If you gave a student the same test 100 times (erasing their memory of the assessment in between administrations), how many times would they get the same score?
- The level of reliability needed is related to the use of the assessments
  - If the score will be used to make a judgment about an individual student, reliability needs to be high, e.g., college entrance or high school graduation. Typically requires reliability coefficients of 0.9 or higher (90 times out of 100).
  - If the score will be used to make a judgment about an aggregate of students (e.g., a school), reliability of the individual score can be lower because reliability is increased through the number of students (e.g., 0.8 or higher).

INCREASING RELIABILITY

- Increase reliability by increasing the measures
  - That means more items
  - Consider, for example, if I asked you to add 54+79. If you got it right, can I assume you have mastered adding two-digit numbers? If you get it wrong, can I assume you don't know how to add two-digit numbers? How many problems would you have to answer before I could comfortably say you've mastered the skill, you don't know it at all, or you have partial understanding?
- Also, increase reliability by testing the same construct multiple ways

VALIDITY

- Tests are not valid, score interpretations are.
- For a score to have a valid interpretation, it must accurately and reliably reflect a student's knowledge and skills.
- Can a test be reliable but not valid?
  - Yes, but this isn't good. Consider an archer who consistently shoots the arrow in the same spot but that spot is always a foot from the bullseye.
- Can a test be valid but not reliable?
  - No. That would be like saying if an arrow hit the bullseye once, the person is a master archer.

ITEM TYPES

- Multiple choice
- Multi-select multiple choice
- Technology enhanced
- Short constructed response
- Extended constructed response

What do we learn from each type and when would we use each?

SAMPLE MC ITEM

Consider the four diagrams shown below. In which of the following diagrams, is one quarter of the area shaded?

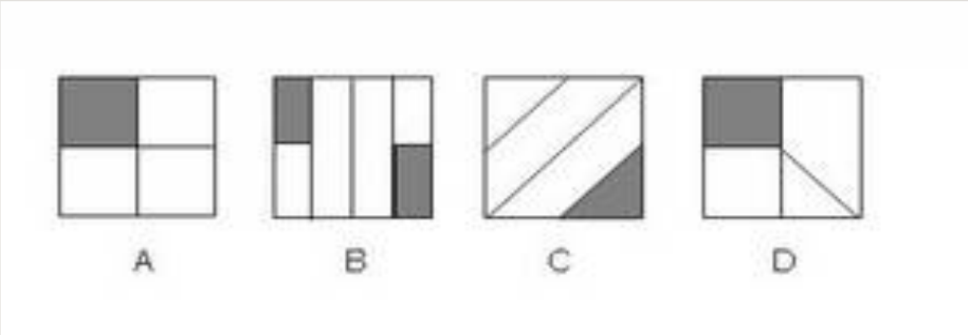


Diagram A is the obvious answer, but B is also correct. However, some students do not believe that one quarter of B is shaded because of a belief that the shaded parts have to be contiguous. Students who believe that one quarter of C is shaded have not understood that one region shaded out of four is not necessarily a quarter. Diagram D is perhaps the most interesting here. One quarter of this diagram is shaded, although the pieces are not all equal; students who rely too literally on the "equal areas" definition of fractions will say that D is not a correct response.

SAMPLE MSMC WITH SHORT CONSTRUCTED RESPONSE

Can It Reflect Light?

What types of objects or materials can reflect light? Put an X next to the things you think can reflect light.

____ water
____ gray rock
____ leaf
____ mirror
____ glass
____ sand
____ potato skin
____ wax paper
____ tomato soup
____ crumpled paper
____ shiny metal
____ dull metal
____ red apple
____ rough cardboard
____ the Moon
____ rusty nail
____ clouds
____ soil
____ wood
____ milk
____ bedsheet
____ brand new penny
____ old tarnished penny
____ smooth sheet of aluminum foil

Explain your thinking. Describe the “rule” or the reasoning you used to decide if something can reflect light.

_____
_____
_____
_____

SAMPLE SHORT CONSTRUCTED RESPONSE

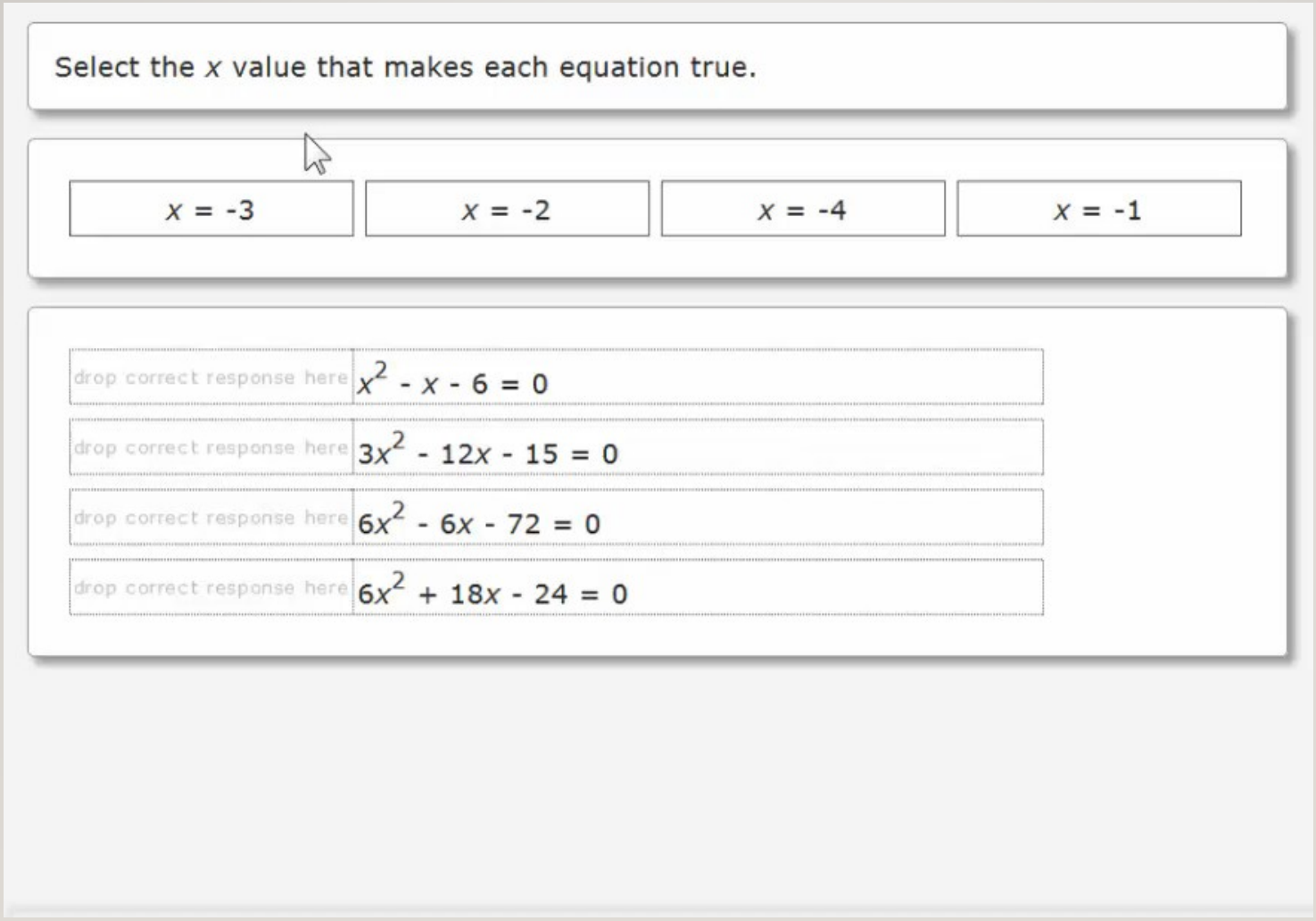
Mr. Ruiz is starting a marching band at his school. He first does research and finds the following data about other local marching bands.

	Band 1	Band 2	Band 3
Number of Brass Instrument Players	123	42	150
Number of Percussion Instrument Players	41	14	50

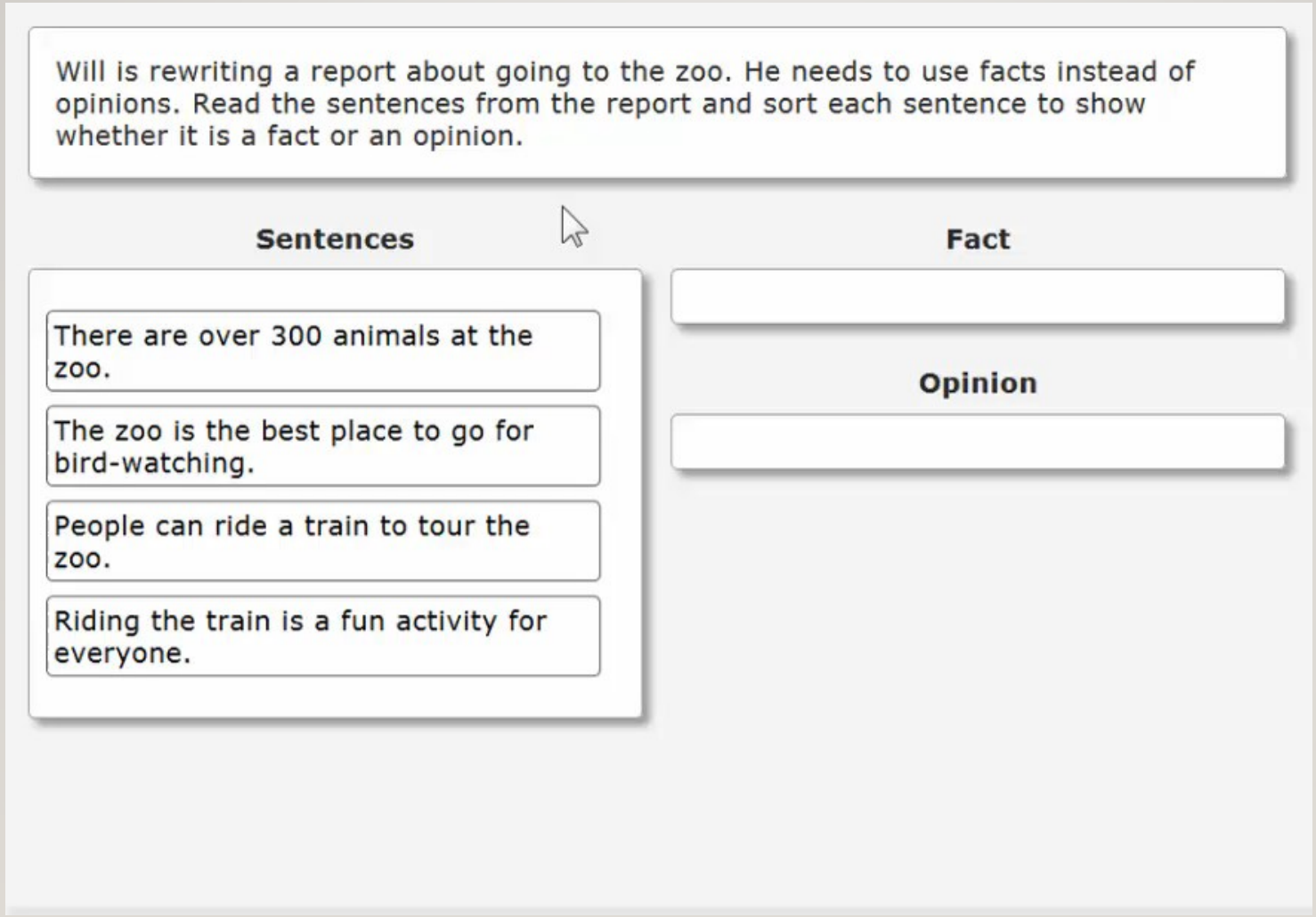
Enter your answer in the box.

Mr. Ruiz realizes there are ☐ brass instrument player(s) per percussion player

LABELING (ONE TO ONE): DRAG AND DROP



CATEGORIZATION: DRAG AND DROP



Graphing

Grant runs 3 miles each day. Show the graph of how far Grant runs over time.

EXTENDED CONSTRUCTED RESPONSE

You have read a website entry and an article, and viewed a video describing Amelia Earhart. All three include information that supports the claim that Earhart was a brave, courageous person.

The three titles are:
- "The Biography of Amelia Earhart"
- "Earhart's Final Resting Place Believed Found"
- "Amelia Earhart's Life and Disappearance" (video)

Consider the argument each author uses to demonstrate Earhart's bravery.

Write an essay that analyzes the strength of the arguments related to Earhart's bravery in at least two of the three supporting materials. Remember to use textual evidence to support your ideas.

CHOOSING AN ITEM TYPE

- What are you trying to measure?
  - Consider "identify" versus "create" or "interpret" versus "graph"
- What resources are available for hand scoring?
- Will the assessment be given solely on computer or split between computer and paper/pencil?

PUTTING IT ALL TOGETHER: DESIGNING AN ASSESSMENT

- Recommend backwards design
- Start with considering what you want the score report to include
  - Scale score and performance category
    - How many performance categories?
    - What are the distinctions?
  - Sub scores
    - How many?
    - What are the important reporting categories?

EVALUATING AN OFF-THE-SHELF ASSESSMENT

- Often, districts purchase, rather than build, and assessment. There are plenty to choose from, so how do you choose?
  - Match your purpose to the purpose the test was designed to fill.
  - Evaluate technical quality.
  - Price (I'm not going to touch this)

CHARACTERISTICS OF A GOOD INTERIM ASSESSMENT SYSTEM

- Provides valid and reliable results that are easy to interpret and provide information on next steps
- Includes a rich representation of content with items linked directly to the content standards and specific teaching units.
- Fits within the curriculum so that the test is an extension of the learning rather than a time-out from learning
- Three main elements
  - Reporting Elements
  - Assessment Design
  - Administration Guidelines

REPORTING ELEMENTS

- Type of data summary
  - Include normative reference
  - Compare against criterion reference
  - Aggregate across classroom/school/district
- Type of qualitative feedback
  - Information on correct/incorrect responses by content area
  - Feedback on what an incorrect answer implies
  - Suggestions for next steps

ADMINISTRATION GUIDELINES

- Flexibility in creating forms
- Administered within instruction or separate from instruction
- Flexibility in when/where the assessment is given
  - Computer-based
  - Web-based
  - Paper-and-pencil
- Turnaround time for results
- Adaptive or not

AN EXAMPLE

- Consider a district with the goals of
  - Implementing an early-warning system to identify which students, classrooms, and schools are not on track to perform well on the end-of-year assessment
  - Identifying areas of weakness both at the student level and aggregated to the classroom and school level for those not on track
  - Providing additional tools for improving performance on those areas identified as weak
  - Administering this test 3–4 times over the year to track student progress toward the goal.

EXAMPLE—CONTINUED

- Reporting criteria
  - Report "on-track to succeed"
  - Identify areas of weakness
  - Aggregate across classrooms, school, and the district
  - Disaggregate results by the same reporting categories used in the end-of-year reports (racial/ ethnic group, disability status, LEP)
  - Illustrate progress over time
    - How the progress relates to where they should be by the end of the year.
  - Provide information that can be used to determine next steps

EXAMPLE—CONTINUED

- Assessment design
  - Items map directly to the content standards and be similar in type to the items on the end-of-year test
    - May include further probes to help identify misunderstandings or schema problems
  - Items also link to teaching units and text books specific to that district.
    - Using items that link directly to instructional materials will help provide the connection between any weaknesses found and instructional interventions
  - Each test should only assess what's been taught to date
    - Do not give a series of parallel mini-summative assessments
    - May be some overlap between assessments

EXAMPLE—CONTINUED

- Administration requirements
  - Results should be available within a week to allow time for intervention
  - Either computer-based testing or a pencil-and-paper test would serve this district's purpose
  - Flexibility is not a requirement for this system
    - Standardization in the items administered would be necessary to aggregate results across the district

EXAMPLE—CONSIDERATIONS

- Could be developed by same vendor as summative
  - Access to same item bank
  - Ability to scale interim items with summative for better predictive value
- Requires good data up front to plot reasonable trajectories
  - Need to link student performance at various points in time to summative performance
  - Issue of students and teachers taking test seriously without rich results

BUYING A COMMERCIAL OFF-THE-SHELF PRODUCT

- What should you consider when buying a test that already has data and reporting built in?
- What information should you request from the vendor?
- Who reviews tests for quality and appropriateness?
- What criteria should they use?

YI, ET AL. 2010 RECOMMENDED SIX CRITERIA

- Test purpose and use
- Test development and documentation
- Administration and inclusion
- Test scores and reports
- Test utility
- Practicality and logistics

TEST PURPOSE AND USE

- Does it match the district's intended purpose and use?
- Does it purport to have too many uses?

TEST DEVELOPMENT AND DOCUMENTATION

- Vendors should supply technical documentation on how the tests were developed.
- Item level information:
  - Content standards: What standards were items built to? How do they match yours?
  - Item types: Do they match the information you are trying to get at?
  - Item difficulties: Reasonable range? How calculated?
  - Item reviews: Content? Bias/fairness? Accessibility?
- Form level information:
  - Balance of representation: What is the emphasis of each standard?
  - Difficulty: What is the range of item difficult and the average of all items?
  - Reliability: How did the calculate the statistic and is it reasonable (e.g., above 0.90)

ADMINISTRATION AND INCLUSION

- How does the test include students with disabilities and English learners?
- If online:
  - What paper version is available?
  - What accommodations are provided?
  - What are the technology requirements?

TEST SCORES AND REPORTS

- How is the overall score calculated?
- What subscores are provided?
- What types of reports are available? (student, classroom, school, district)
- What other information is provided? E.g., growth, normative, predictive
- How is information displayed?
- Is error band reported?

TEST UTILITY

- *"Utility represents the extent to which intended users find the test results meaningful and are able to use them to improve teaching and learning" (*Herman, J.L., & Baker, E.L. (2005). Making Benchmark Testing Work. *Assessment to Promote Learning, 63,* 48-54.)
- How does the test fit within the instructional time?
- How easy is it to administer and take?
- How fast are results provided?
- Is test interpretation time supported by the school/district?

HOW DO I KNOW I'M GETTING MY MONEY'S WORTH?

- Validating the evidence will be important to do over the next couple of years
  - If the test is used for predictive purposes, do a follow up study to determine that the predictive link is reasonably accurate and that the use of the test contributes to improving criterion (e.g., end of year) scores
  - If the test is used for instructional purposes, follow up with teachers to determine how the data were used and whether there was evidence of improved student learning for current students
  - If the test is used for evaluative purposes, gather data from other sources to triangulate results of interim assessment and follow up to monitor if evaluation decisions are supported

REMEMBER…

- It should fit within a balanced assessment system
  - Evaluate the same standards as the summative
  - Fit into classroom time as an extension of instruction, not a time out from instruction
  - Provide results that feed into the teachers' instructional planning

THANK YOU!

Marianne Perie

mperie@ku.edu

@DrMarianneP