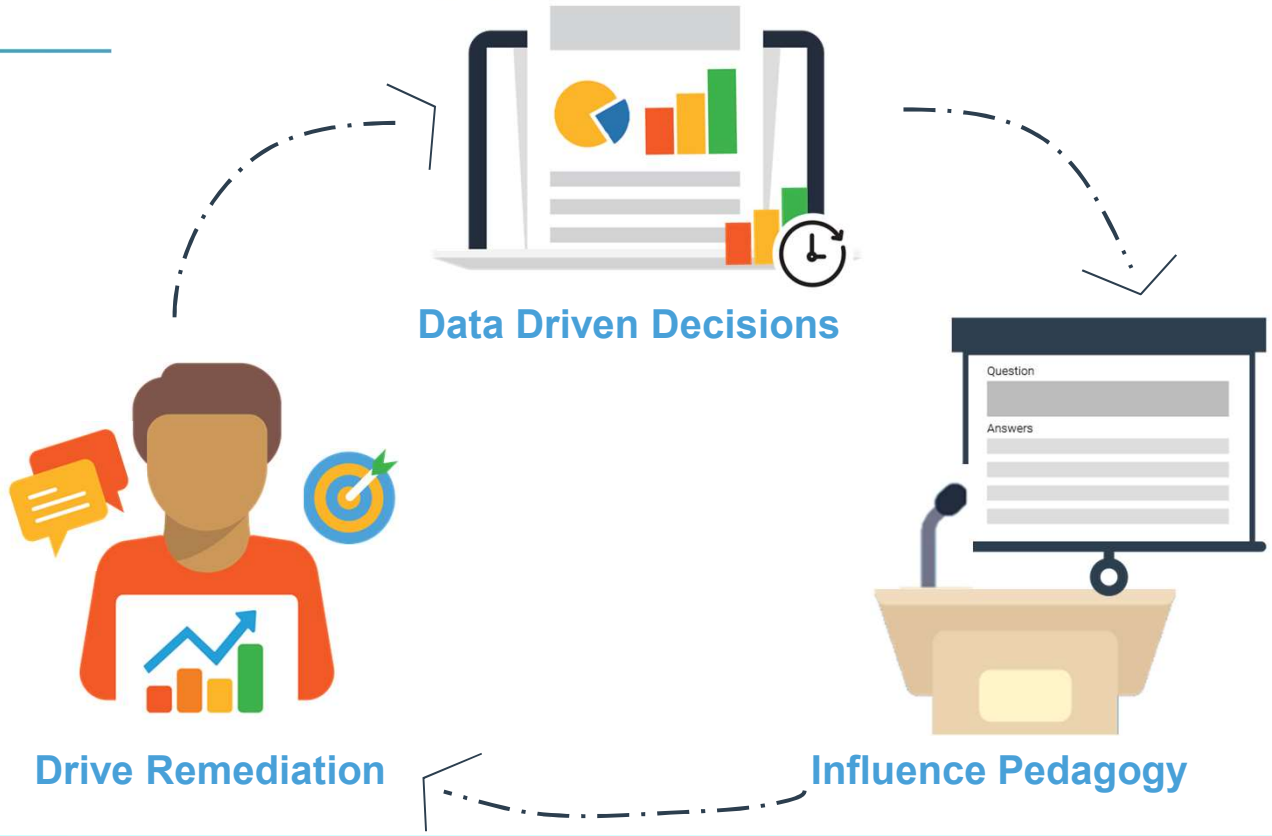
Psychometrics 101:

Using Psychometric Data to Improve Assessments and Student Learning

Psychometrics 101



Context Matters

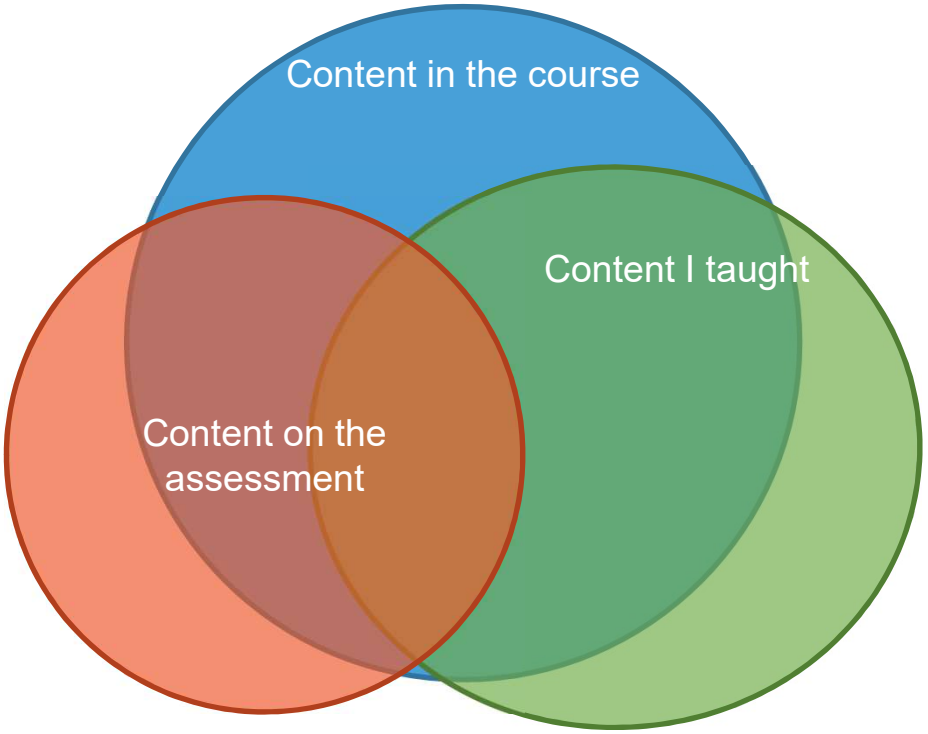
Always have the end in mind

- What are we trying to achieve?
  - Creating high cognitive questions?
  - Stimulate high order of thinking?
  - Part of assessment?
  - How does item analysis feature in?
  - Quality control?
  - Part of teaching/learning?
  - Low stake/high stake
  - Curriculum mapping

Validity

- Are you actually testing what you intended to measure?
- Clinical subjects:
  - Am I testing recall? (how well can the student memorize)
  - Which skill am I trying to test?
  - Am I testing problem solving skills?
  - Am I testing understanding/applying the context?
  - Competency? Mastery?
- Test relevant material
- But cover a wide variety
- More than one assessment!!!!

Is the assessment valid?



Reference: 1, 2

Validity

- Really a matter of degree and not “all or nothing”
- Content validity
  - Have at least moderate to high levels of internal consistency
  - Blueprinting
- Construct validity
  - Measures what we want to measure
- Multiple measures (triangulation)

Reliability

- Test-retest reliability
  - Will the scores be stable over time?
- Parallel forms reliability
  - Are the two forms of the test or measure equivalent
- Internal consistency reliability
  - How well does each item measure the content or construct under consideration?
- Inter-rater reliability
  - Are the raters consistent in their ratings?

Threats to Reliability

- Group homogeneity
- Time limits
- Test or measure length
- Scoring errors
- Item quality

How is reliability different from validity?

- Reliable but not valid

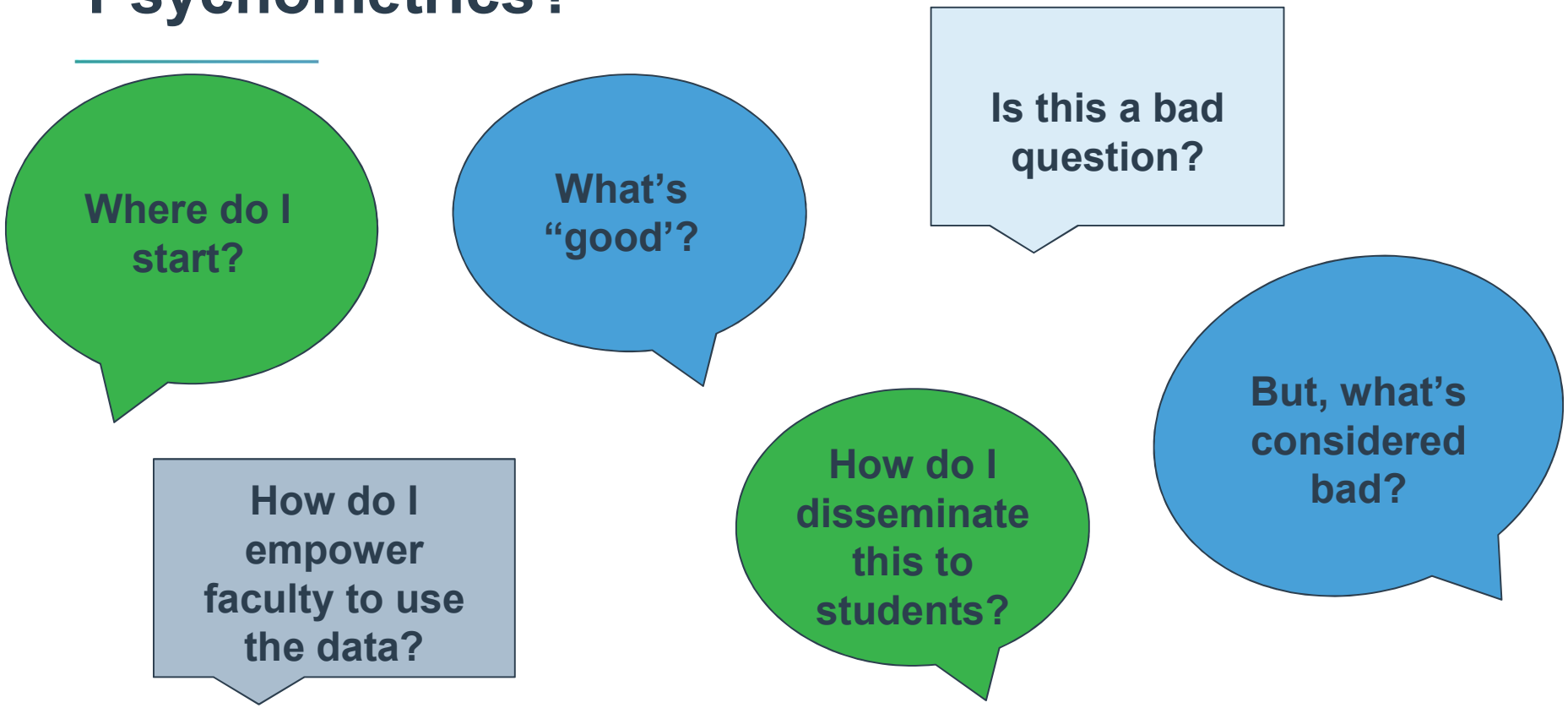
- Reliable and valid

Formative and summative assessment

- Assessment for learning,
  not learning for assessment
- Use frequent assessments
  - Remediation

Types of Psychometrics

Psychometrics?



Types of Statistics

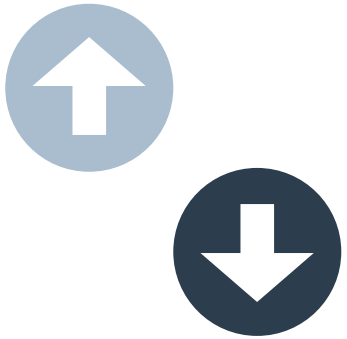
- Item Difficulty Index (P-Value)
- Upper Difficulty Index (Upper 27%)
- Lower Difficulty Index (Lower 27%)
- Discrimination Index
- Point-Bi-serial Correlation Coefficient (PBI)
- Kuder-Richardson Formula (KR-20)
- Cronbach’s Alpha

Types of Statistics: Item Difficulty/p Value

- **What is it?**
- Percentage of students who got the item correct

- **What's acceptable?**
- The lower the decimal the higher the difficulty
    - Mastery: Close to 1.00 as possible
    - Discrimination: Lower levels acceptable

Types of Statistics: Upper/Lower 27%



What is it?

- Quartile representation of highest and lowest scorers

What's acceptable?

- **Upper 27%**
- Closer to 1.00 = higher scorers performed well on the assessment
- **< 60%** of your top performers are getting a question correct, a further analysis is needed
- **Lower 27%**
- Closer to 1.00 = item was not intended to discriminate
- This value should not exceed upper 27%

Types of Statistics: Discrimination Index

What is it?

- Measures whether the item discriminates between highest (upper 27%) and lowest (lowest 27%) scorers

What's acceptable?

Good discrimination	**≥ 0.3**
Fair Discrimination *(may need to review)*	0.10 - 0.29
No discrimination – upper 27% and lower 27% answered with the same accuracy	= 0
Negative (why? Completely remove or revise item)	**≤ 0**

Types of Statistics: Point-Biserial (PBI)

What is it?

- Indicates whether doing well on that specific item correlated with doing well on the exam overall.

What's acceptable?

Very good	**0.4 - 0.7**
Good	0.3 - 0.39
Fair	0.2 - 0.29
Poor *(consider revisions)*	**< 0.2**

Types of Statistics: KR-20

What is it?

- Measures internal reliability of the assessment. How likely are students to perform the same taking this assessment again?

- **What's acceptable?**
  - Instructor-made assessment: >0.5 is acceptable
  - Items from a database/licensure assessments: expected to consistently score

Instructor created assessments	> 0.5 is acceptable
Items from a database, licensure exams	0.6 - 0.8

Types of Statistics: Cronbach's Alpha

What is it?

- Measures reliability or internal consistencies for assessments with items that have dichotomous items (opportunity for partial credit)

What's acceptable?

Excellent	**≥ 0.9 – ≥ 0.8**
Good	0.79 - ≥ 0.6
Poor	0.59 - ≥ 0.5
Consider Revisions	**0.5 >**

Keep in mind,

When evaluating any statistic context matters!

Extraneous Variables

“Statistics alone cannot tell the *whole* story..”

5 Factors to always consider when evaluating item performance:

1. Cheating
2. Return on investment
3. Conflicting content or faculty
4. Author intent
5. Content delivery method

What other factors may skew your data?

Interpreting Item Analysis

Scenario 1

Item #	Correct Responses			Disc. Index	Point Biserial	Correct Answer	Response Frequencies (*Indicates correct answer)				
	Diff. (p)	Upper	Lower				A	B	C	D	E
7	0.66	82.00%	46.15%	0.36	0.28	D	7	17	28	120 *	9
						% Selected	3.87	9.39	15.47	66.30	4.97
						Point Biserial (rpb)	-0.11	-0.19	-0.12	0.28	-0.07
						Disc. Index	-0.04	-0.19	-0.09	0.36	-0.04
						Upper 27%	0.00	0.00	0.12	0.82	0.06
						Lower 27%	0.04	0.19	0.21	0.46	0.10

What data point do we look at first?

Scenario 2

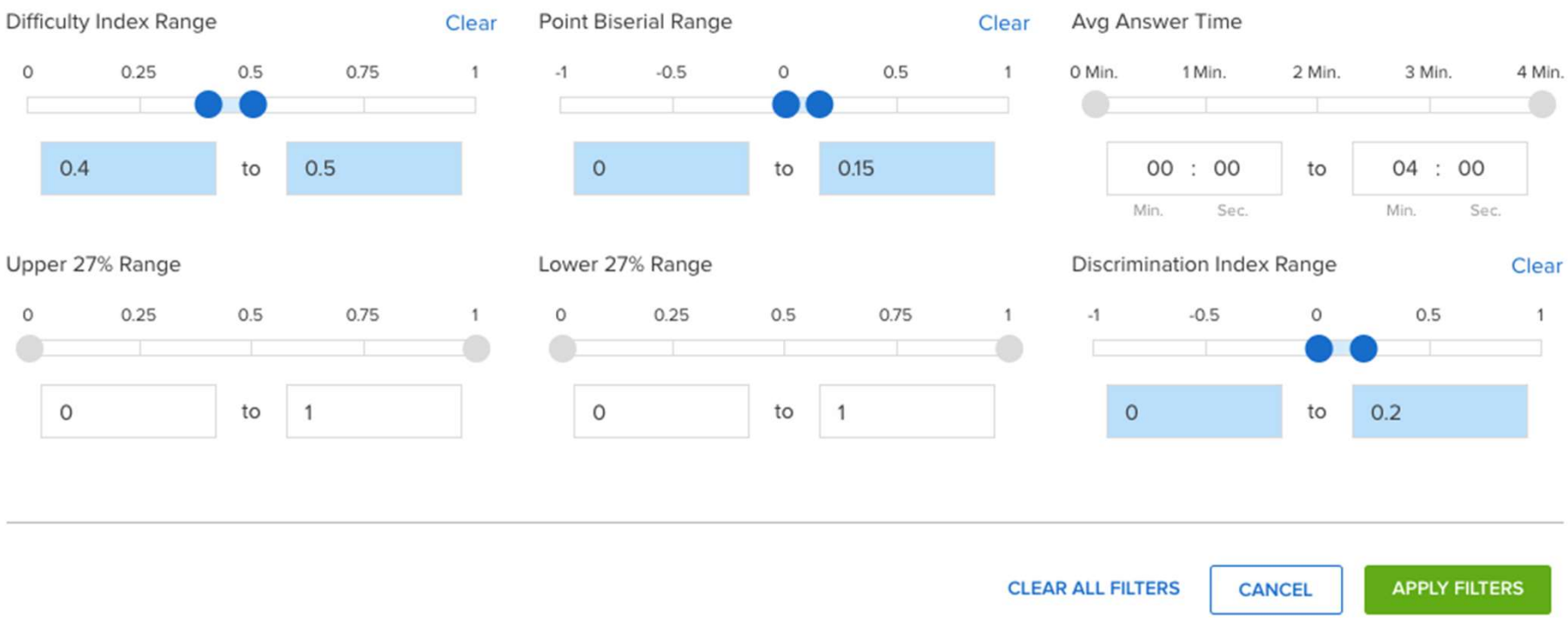
Item #	Correct Responses			Disc. Index	Point Biserial	Correct Answer	Response Frequencies (*Indicates correct answer)				
	Diff. (p)	Upper	Lower				A	B	C	D	E
22	0.36	52.00%	26.92%	0.25	0.22	D	35	34	21	66 *	25
						% Selected	19.34	18.78	11.60	36.46	13.81
						Point Biserial (rpb)	-0.09	0.04	-0.20	0.22	-0.06
						Disc. Index	-0.15	0.07	-0.15	0.25	-0.02
						Upper 27%	0.10	0.24	0.04	0.52	0.10
						Lower 27%	0.25	0.17	0.19	0.27	0.12

Scenario 3

Item #	Correct Responses			Disc. Index	Point Biserial	Correct Answer	Response Frequencies (*Indicates correct answer)				
	Diff. (p)	Upper	Lower				A	B	C	D	E
24	0.52	64.00%	42.31%	0.22	0.18	C	61	21	94 *	5	0
						% Selected	33.70	11.60	51.93	2.76	0.00
						Point Biserial (rpb)	-0.10	-0.19	0.18	0.12	0.00
						Disc. Index	-0.12	-0.13	0.22	0.04	0.00
						Upper 27%	0.26	0.04	0.64	0.06	0.00
						Lower 27%	0.38	0.17	0.42	0.02	0.00

What statistic strikes you right away about this item?

Scenario 4 - Borderline Questions



Scenario 4 - Borderline Questions

Performance History										
Rev	Attempted	Right	Wrong	Diff. (p)	Point Biserial	A	B	C	D	E
1	240	144	96	0.6	0.14	36	195	49	47	159
2	280	84	196	0.3	0.13	48	106	96	52	140

Scenario 4 - Borderline Questions

Exam Name	Date	Correct	Pt. Biserial	Upper 27%	Lower 27%	Disc. Index
Exam Session 1	1/28/19	59%	0.27	56%	43%	13%
Exam Session 2	1/28/19	15%	-0.19	19%	16%	3%
Exam Session 3	1/28/19	61%	0.11	50%	35%	15%
Exam Session 4	1/28/19	21%	-0.1	28%	27%	1%

Post Assessment Decisions

Step 1: Identifying Items for Evaluation

- **Reflect on overall impressions**
  - Big picture
  - Number of red flags
- **Identify items that warrant further evaluation:**
  - Items with difficulty < 40-50%
  - Items with PBI < 0.2
  - Distractors with positive PBI
    - High class % choosing same wrong answer

Step 2: Individual Item Evaluation

- **Evaluate:**
  - Compare with lecture material, homework, and readings
  - Read the question from the student's perspective
  - Self-reflection
  - Determine if items require adjustments

Keep in mind, context matters! Could there be variables contributing to poor performing items?

Step 3: Item Adjustment

- **Adjust:**
  - Departmental standards
  - Evidence/personal choice
  - Faculty error
    - Give full credit, throw out, bonus
  - Exam curve
- **No adjustment:**
  - Identify at-risk students
  - Reinforce most missed concepts in class

What are department policies you adhere to when adjusting items?

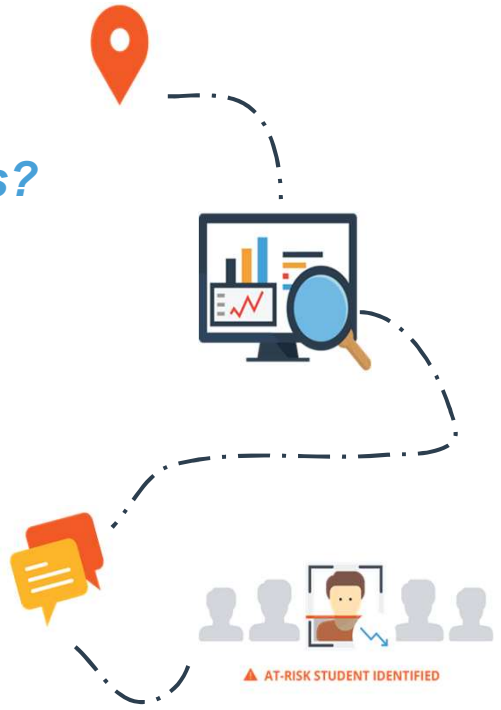
Step 4: Post Evaluation and Adjustment

- Post/share adjusted scores with students
- Review adjusted items with class
- Categorize adjusted items for review prior to the next assessment
  - Items that require revisions
  - Items that surprised you
  - Modify levels of Bloom's as necessary

How do you use categories to identify items for review within the bank?

How do you review/revise items?

- *Are items reviewed collaboratively?*
- *What criteria do you evaluate ahead of adjusting items?*
- *What policies are instituted at a department level?*
- *How long do items remain active in your item bank?*

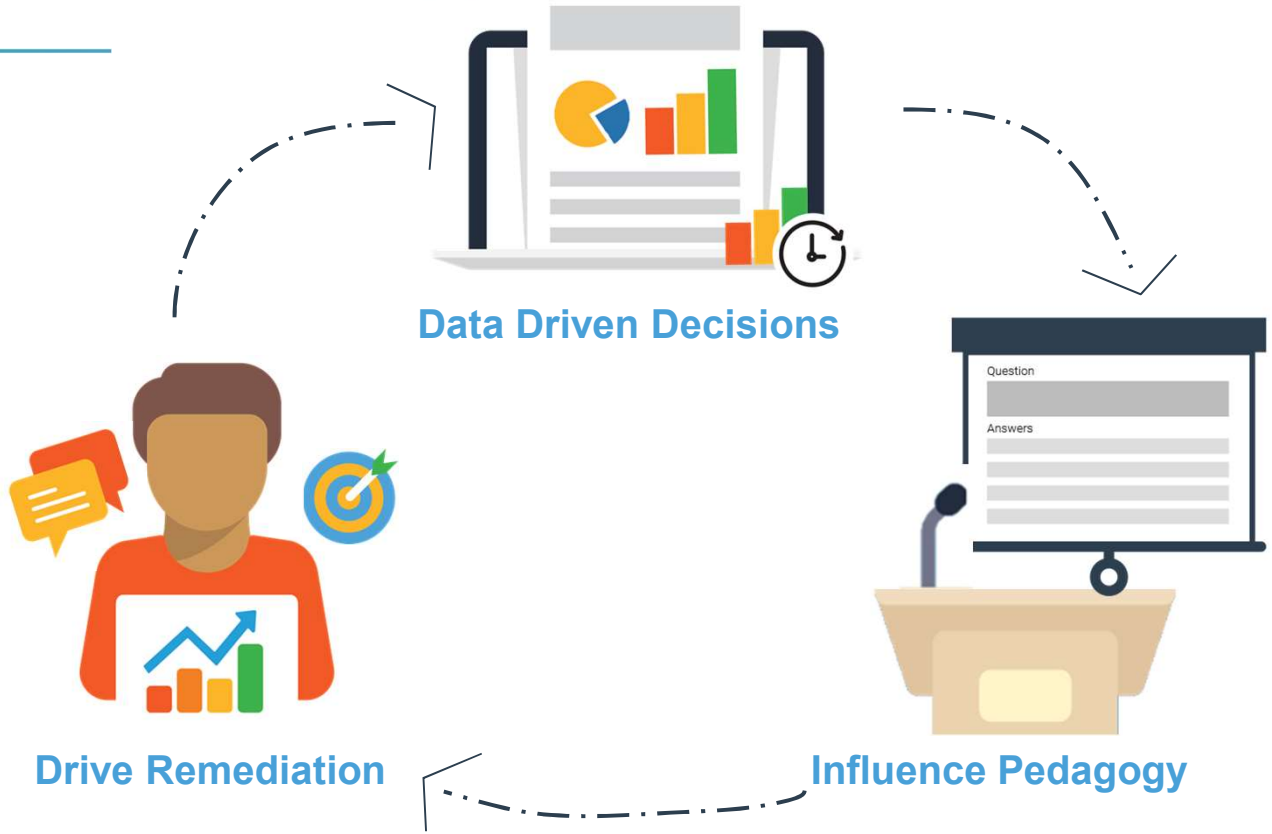


Using Assessment Data

Objectives

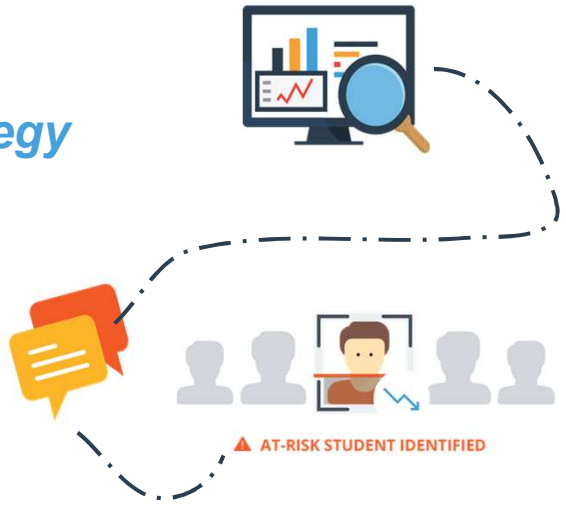
1. Recognize different strategies that can be used to inform your teaching
2. Provide meaningful feedback to students for self-assessment
3. Evaluate effectiveness of teaching methods
4. Apply use of exam data for student remediation

Psychometrics 201



Things to think about…

- *If you are a faculty, how would you describe your pedagogy?*
- *Has there ever been a teaching method or strategy you've want to try but haven't?*
- *How does data influence your pedagogy?*



Weaving Pedagogy into the Portal

Strategies to inform teaching:

- Question data and analytics
- Categories
- Examplify for content delivery
- Rationale

Categories

- **How do you currently tag your items?**
  - Accreditation?
  - Program/Institution Standards?
  - Content standards/learning objectives?
  - Topics?
  - Certification exam areas?

Categories

What about using categories that influence your pedagogy?

- **Do you tag:**
  - Lecture dates?
  - Remediation item?
  - Writer/Faculty?
  - Instructional mode/method?
  - Item use time?

Categories

Tagging items to the method used to teach the content can help inform on how effective those methods are.

- Team-based learning
- Self-paced
- Lecture
- Flipped classroom

Questions?