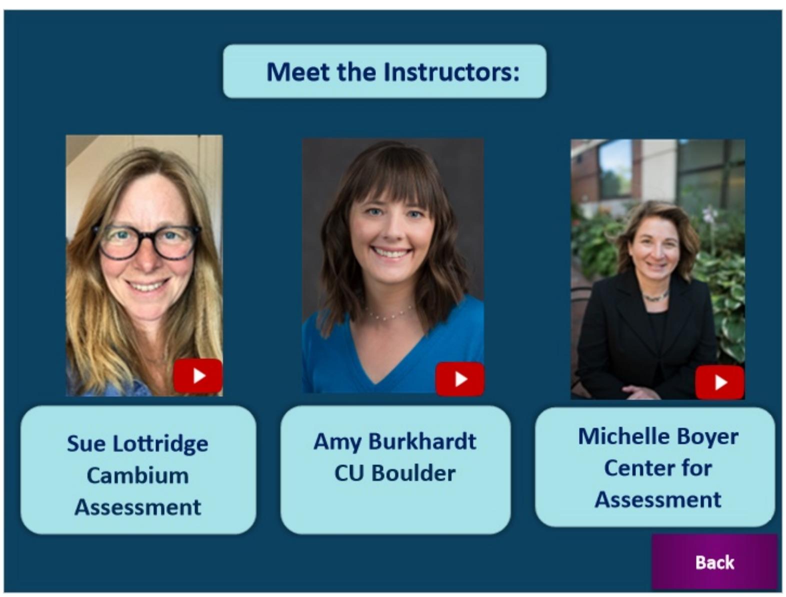
DM18 SLIDES (Automated Scoring, Version 1.1)

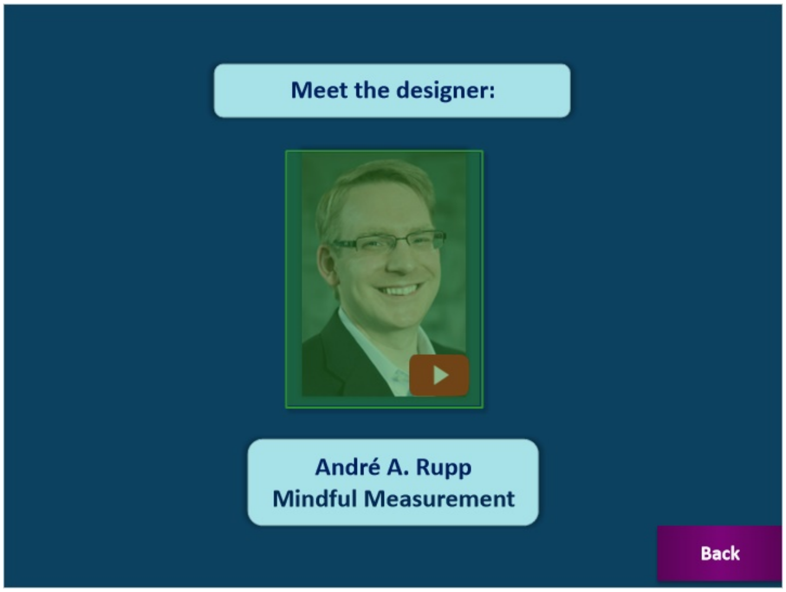
1. Module Overview

1.1 Module Cover (START)

1.2 Instructors



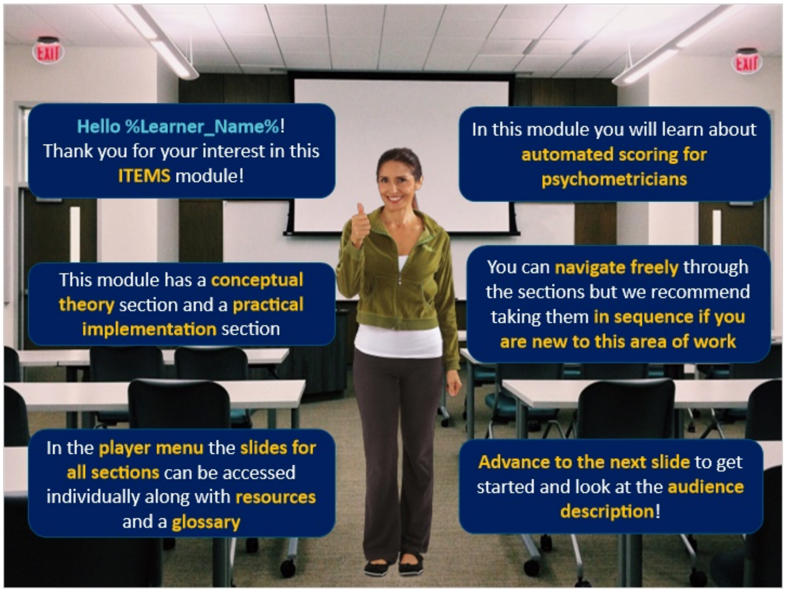
1.3 Designers



Andre V1 (Slide Layer)

1.4 Welcome

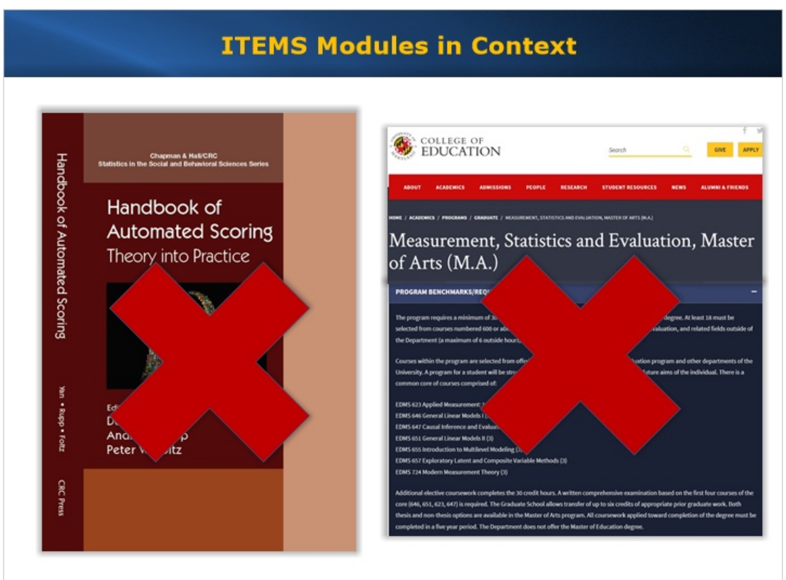
1.5 Overview



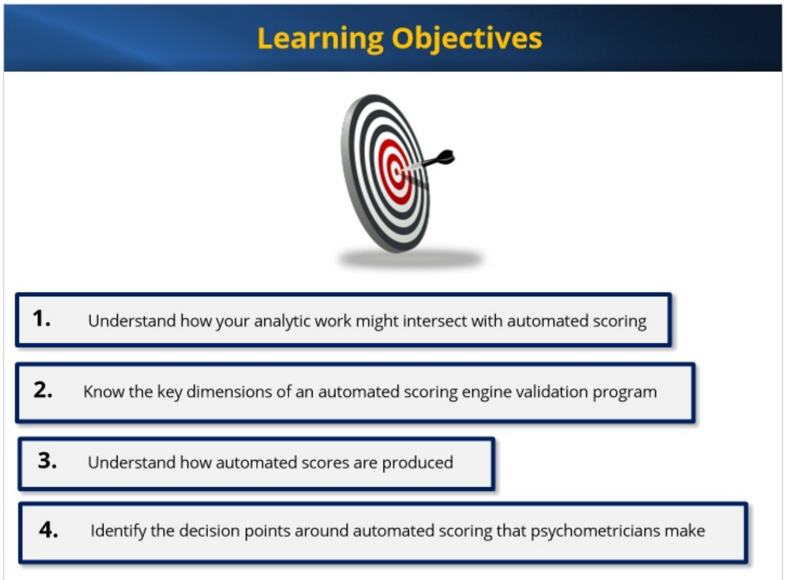
1.6 Target Audience

1.7 Expecations (I)

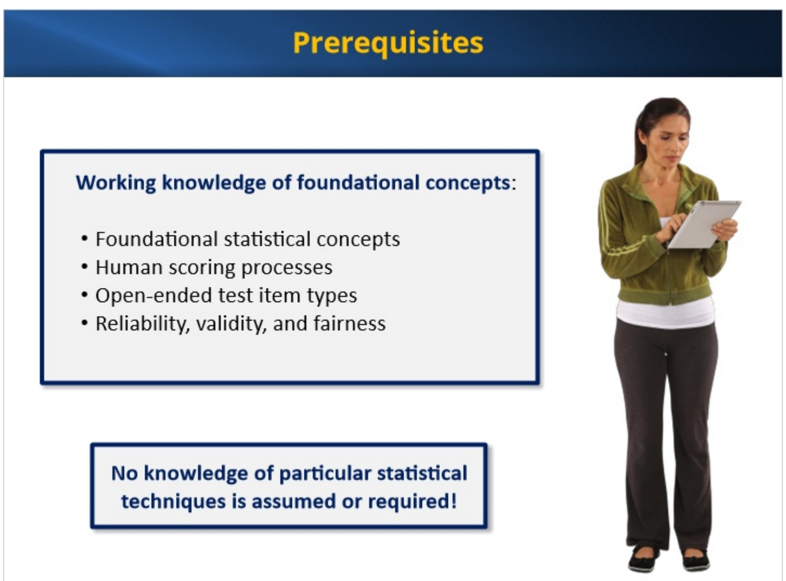
1.8 Expectations (II)



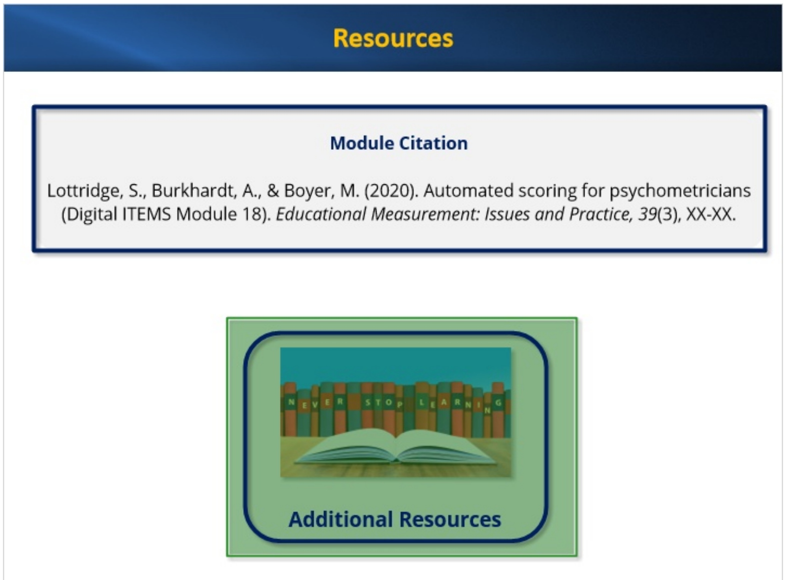
1.9 Learning Objectives



1.10 Prerequisites

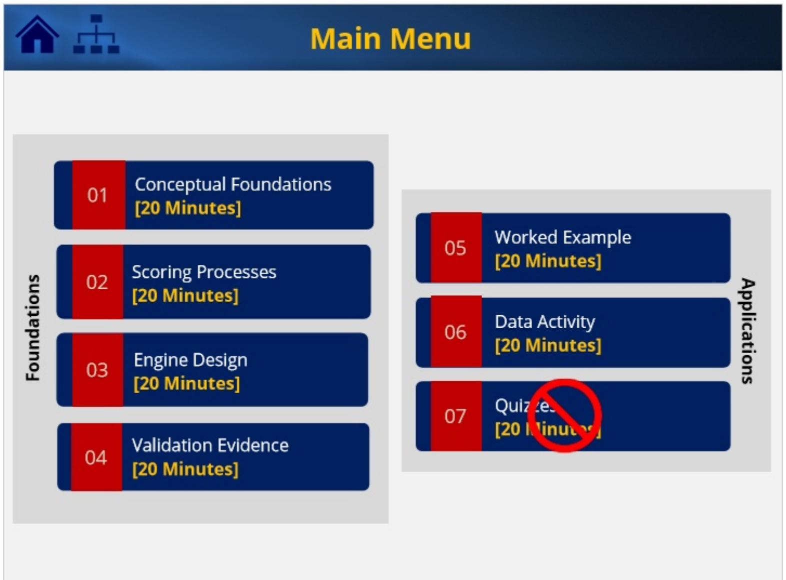


1.11 Resources



References (Slide Layer)

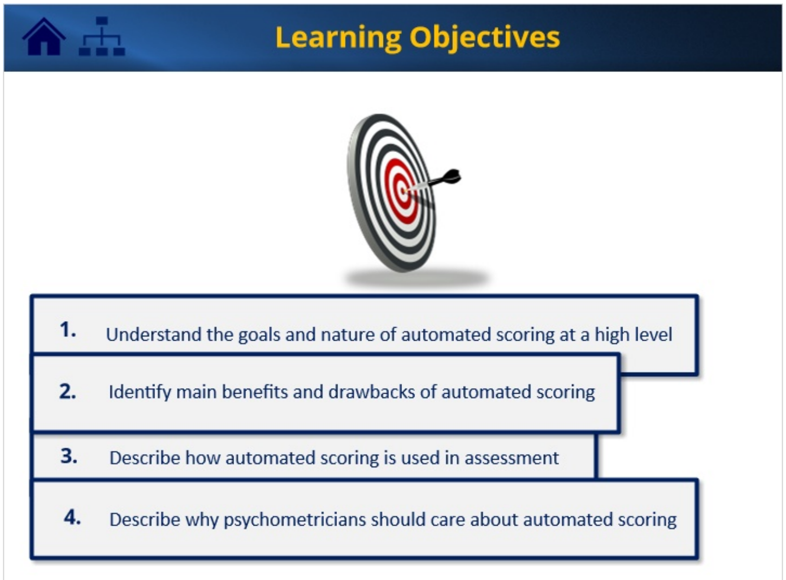
1.12 Main Menu



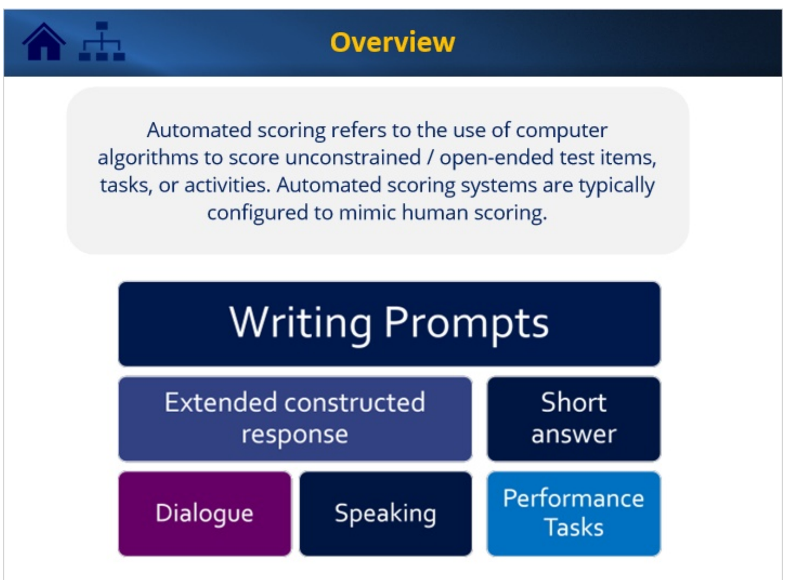
2. Section 1: Conceptual Foundations

2.1 Cover: Section 1

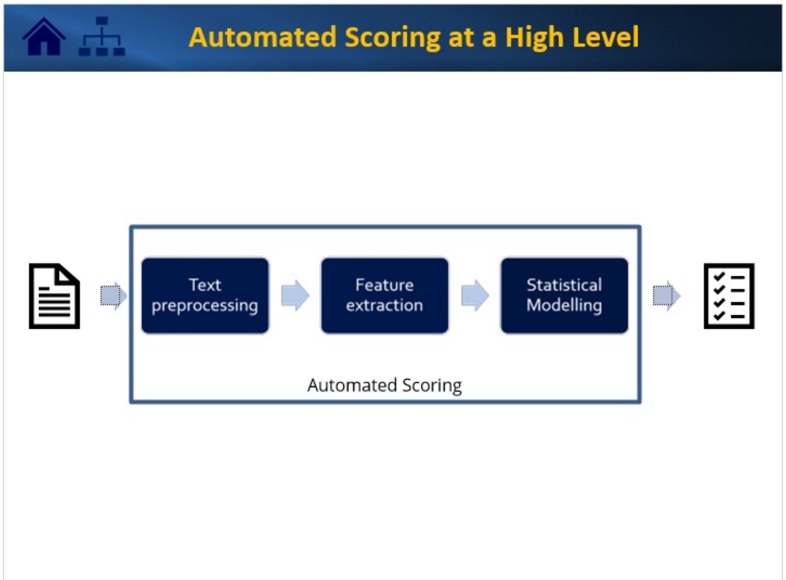
2.2 Learning Objectives



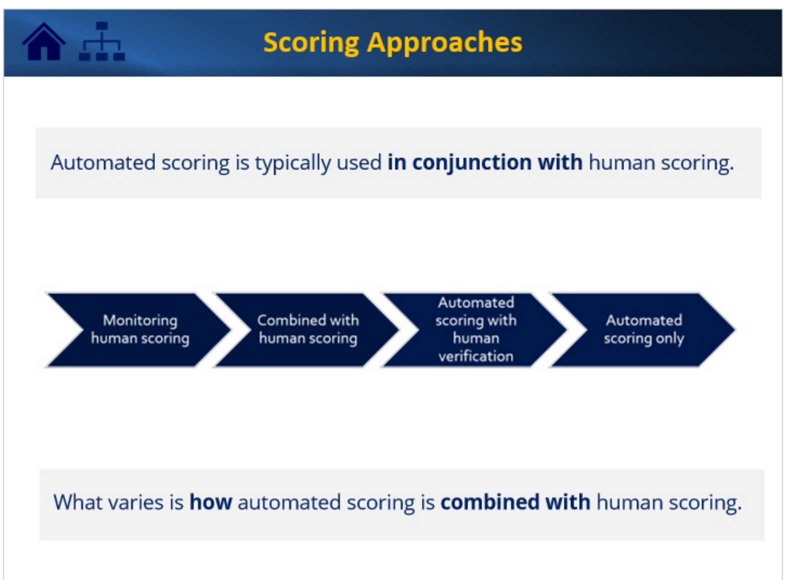
2.3 Overview



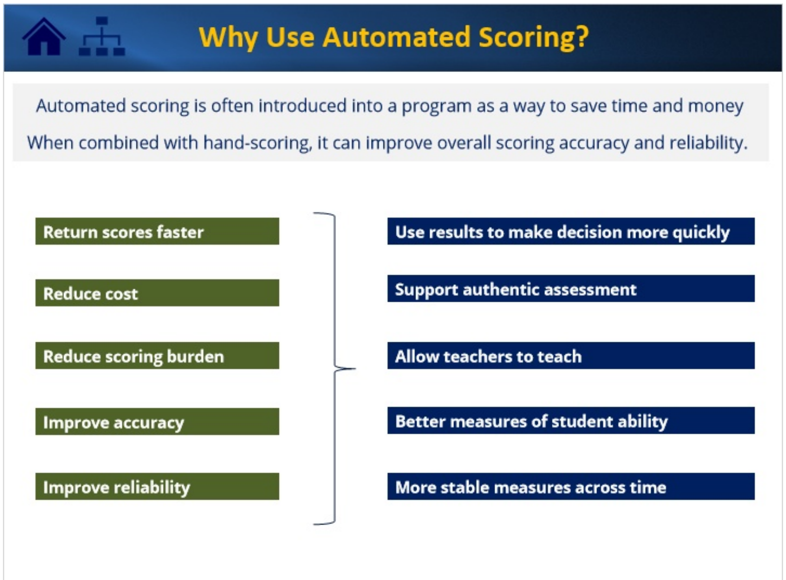
2.4 AS at a High Level



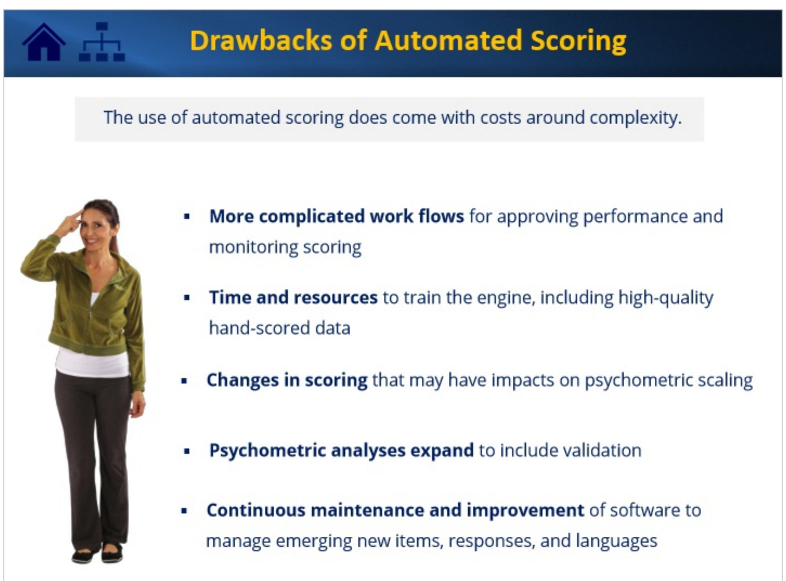
2.5 Scoring Approaches



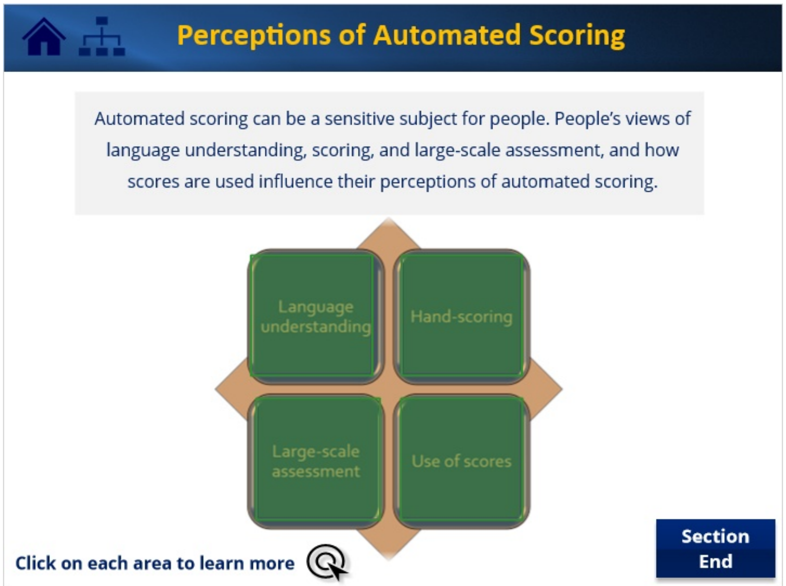
2.6 Reasons for Use



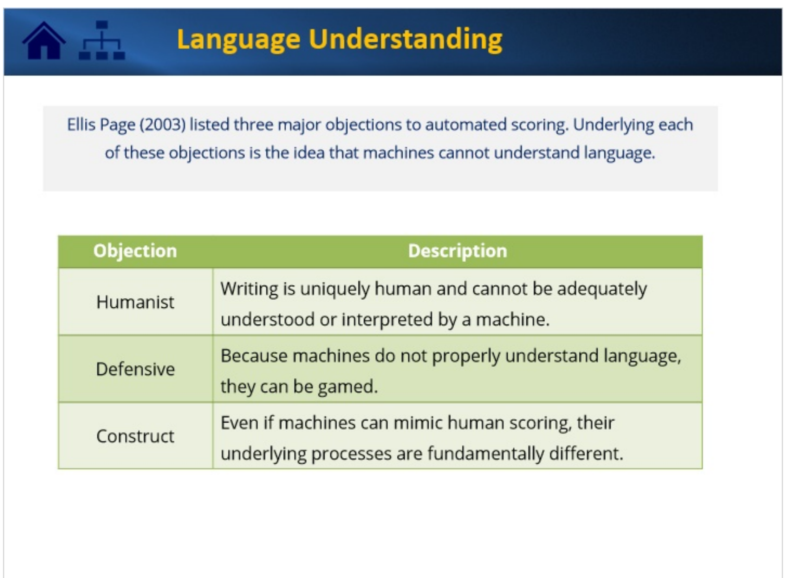
2.7 Drawbacks



2.8 Topic Selection



2.9 Understanding



Objection	Description
Humanist	Writing is uniquely human and cannot be adequately understood or interpreted by a machine.
Defensive	Because machines do not properly understand language, they can be gamed.
Construct	Even if machines can mimic human scoring, their underlying processes are fundamentally different.

2.10 Hand-scoring

Hand-Scoring

Perceptions of automated scoring are closely tied into perceptions of hand-scoring quality. We should know the limitations of hand-scoring.

Element	Description
Raters	The recruitment, qualification, and monitoring of raters requires considerable effort to achieve high-quality scoring.
Accuracy	Error exists in hand-scoring, as evidenced by exact agreement rates and other measures of reliability.
Bias & drift	Raters can exhibit severity/leniency bias, and scores can drift over time due to fatigue and context effects

The worked example provides an illustration about the need for human score quality and that humans don't score perfectly.

2.11 Large-Scale Assessment

Large-Scale Assessment

Understandably, views of automated scoring are linked with views on large-scale assessment because automated scoring is often – although not always – used in a large-scale assessment context.

Element	Description
Psychometrics	The psychometric processes underlying the scores reported to teachers and students are often opaque to most stakeholders and the link between item-level scores and test-level scores is not clear.
Curriculum & instruction	Stakeholders want to use the results of large-scale assessment to improve teaching and learning, and course curriculum is often modified to better align with large-scale assessment scores.

2.12 Use

Use of Scores

How automated scores are used can impact perceptions of automated scoring.

Element	Description
Immediate feedback	Scores that are returned immediately have generally been seen as useful because they can be applied to teaching and learning quickly.
Replacement for hand-scoring	When automated scoring replaces hand-scoring, it can raise questions around whether the engine is prioritizing different elements of writing over what hand-scorers may prioritize, thereby potentially impacting instruction
Combined with hand-scoring	Stakeholders are generally more receptive to scoring approaches in which automated scoring is used alongside hand-scoring. This is because a hybrid approach retains the link to the 'human' elements around scoring.

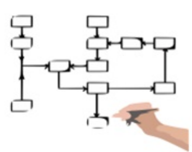
2.13 Psychometric Relevance

Why Should Psychometricians Care about AS?

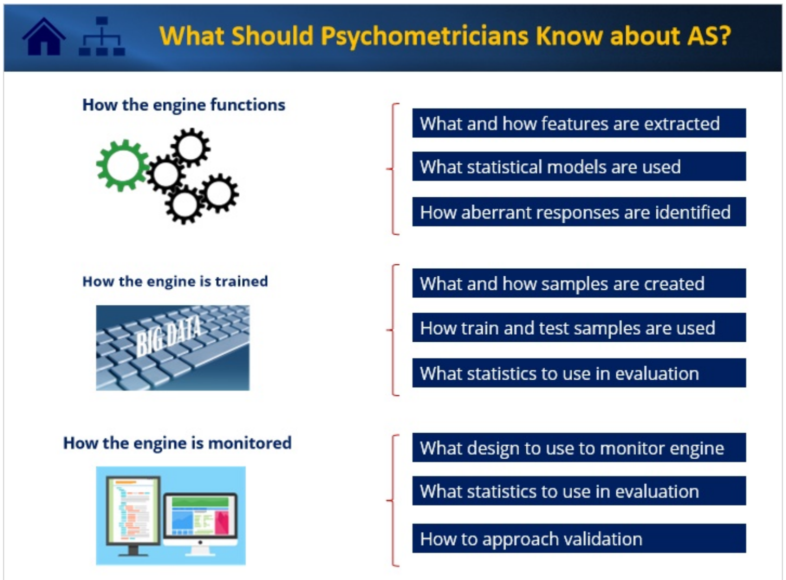
The use of automated scoring is growing in large-scale assessment programs

Scoring of open-ended items is a psychometric concern outlined in the *Standards*

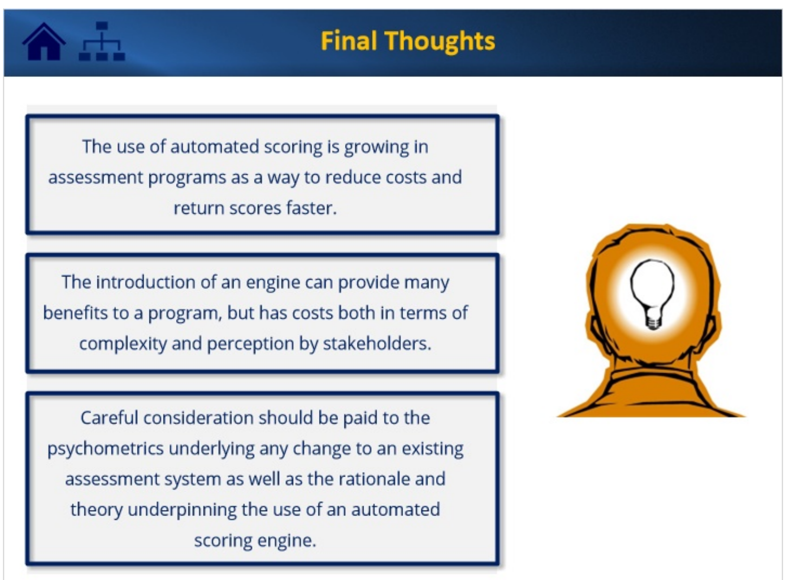
The analysis of automated scoring results is (or will be) part of the psychometric workflow



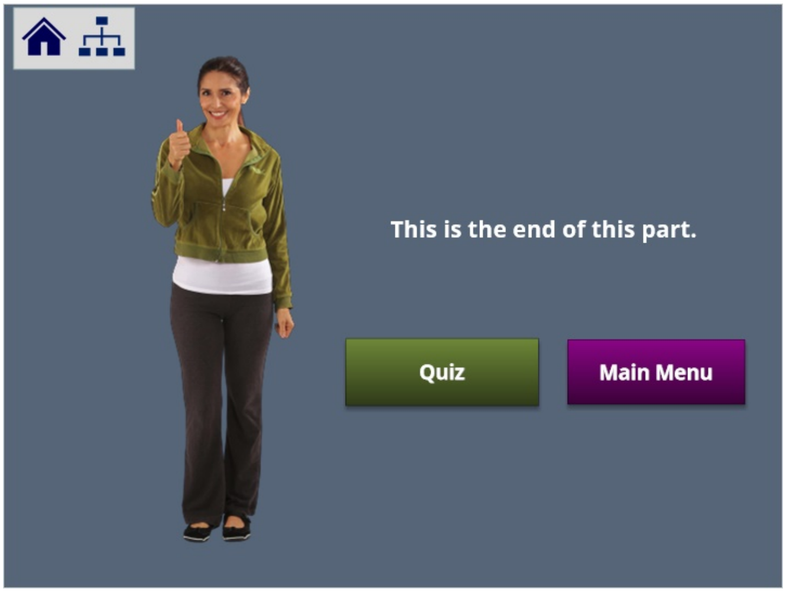
2.14 What Psychometricians Should Know



2.15 Final Thoughts

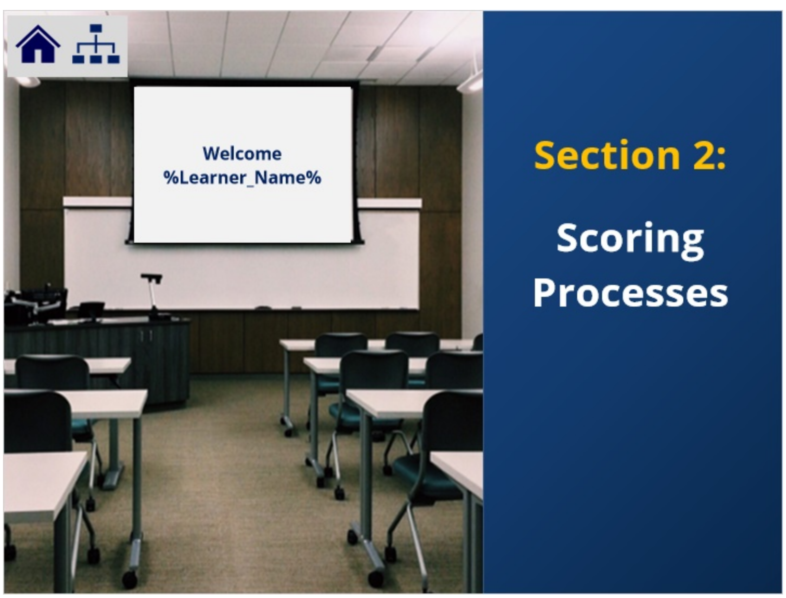


2.16 Bookend: Section 1

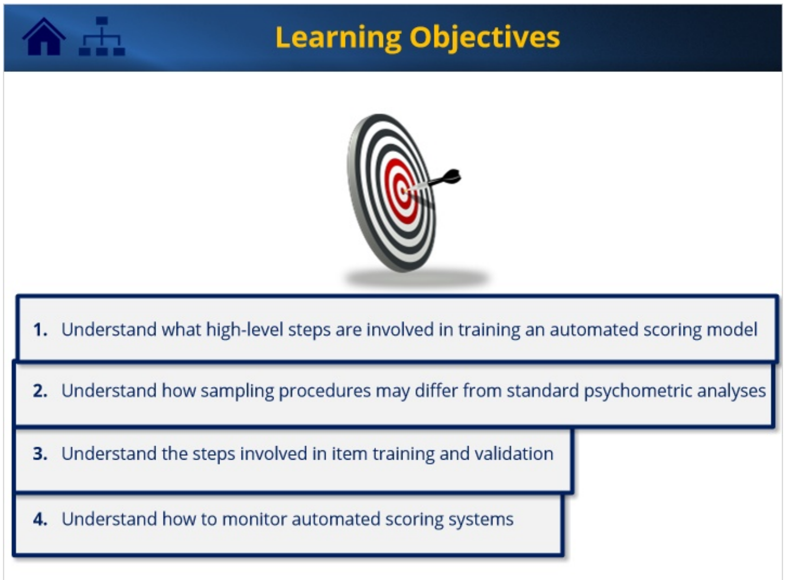


3. Section 2: Scoring Processes

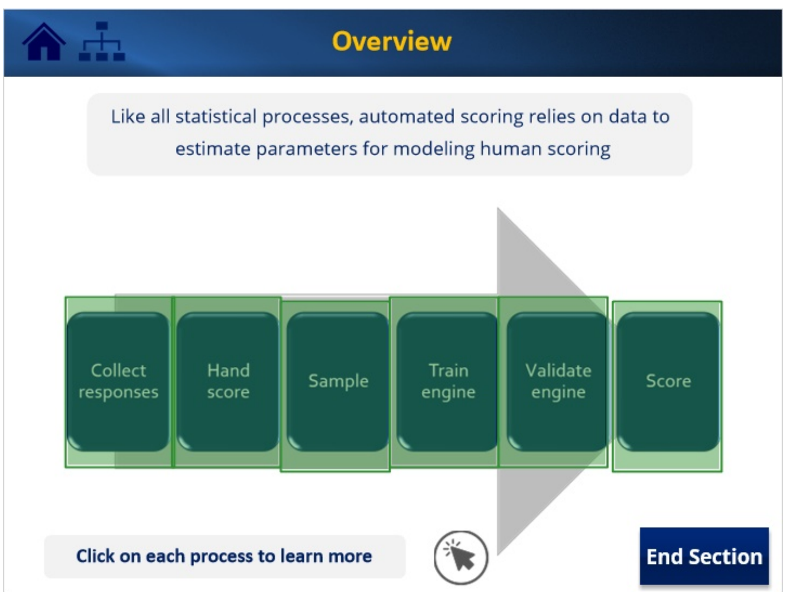
3.1 Cover: Section 2



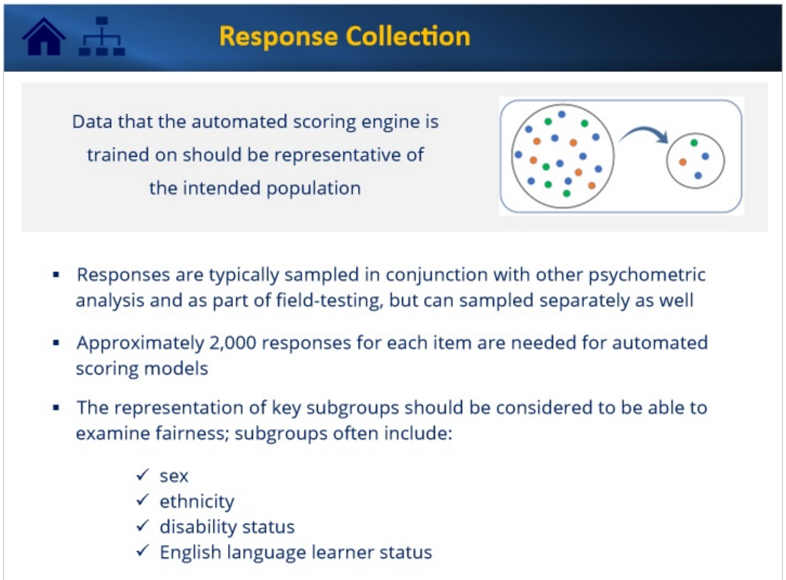
3.2 Learning Objectives



3.3 Topic Selection

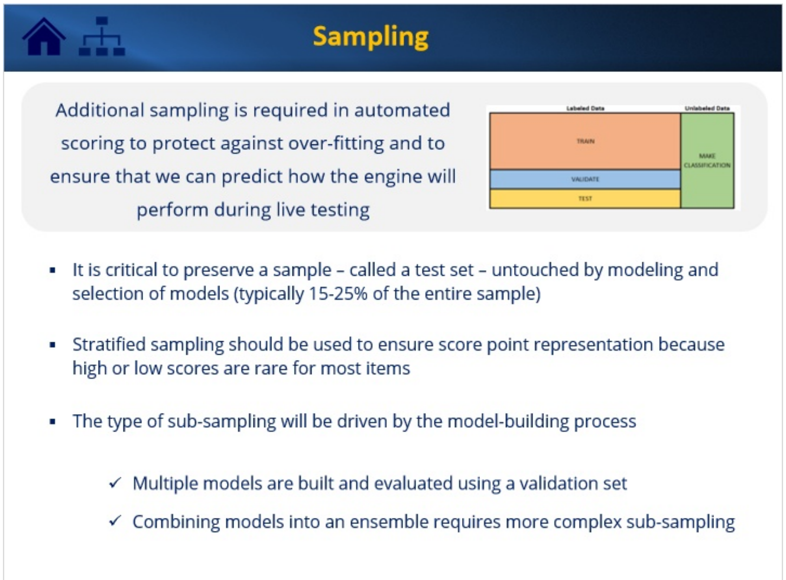


3.4 Response Collection

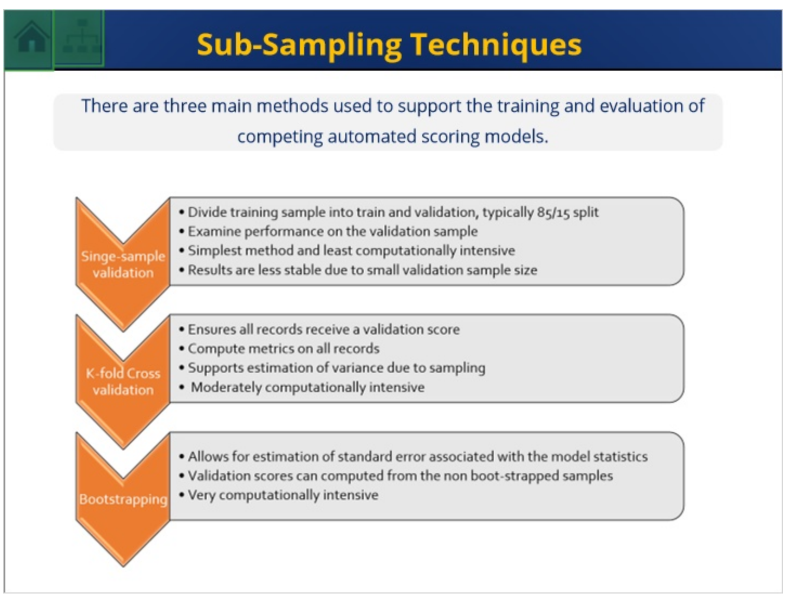


3.5 Hand-scoring

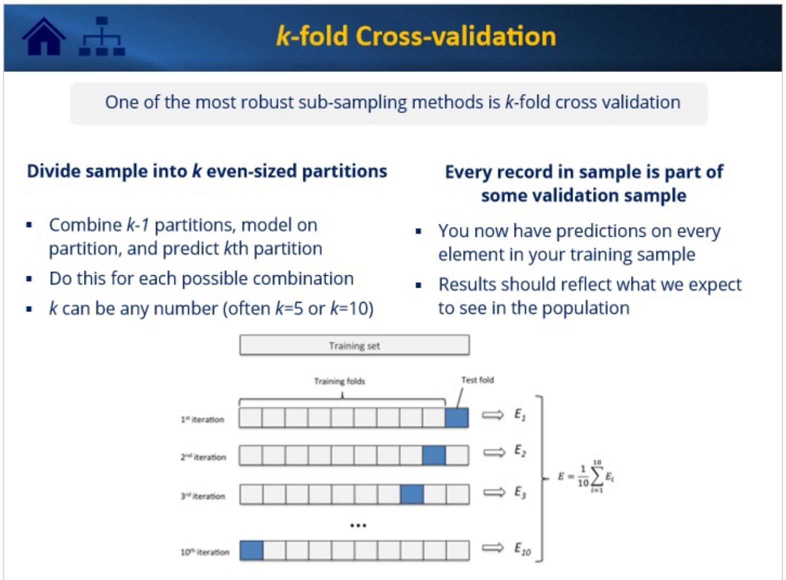
3.6 Sampling



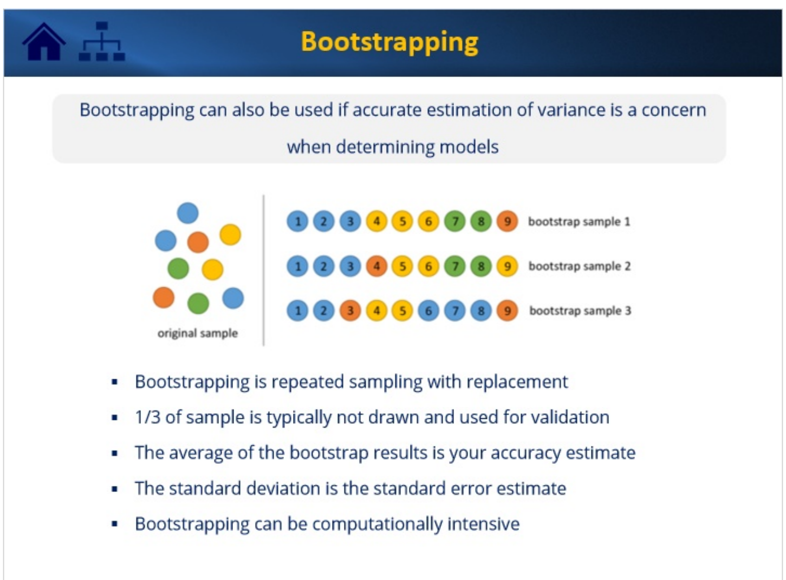
3.7 Sub-sampling



3.8 Cross-validation



3.9 Bootstrapping



3.10 Evaluation Criteria

Evaluation Criteria

Criteria used to evaluate engine performance can be absolute and or relative to human scoring

Measures	Example Absolute criteria	Example Relative criteria
Quadratic Weighted Kappa	Engine-human QWK must be greater than .7	Engine-human QWK should be within .1 of human-human
Exact Agreement	Engine-human Exact Agreement must be greater than 70%	Engine-human Exact Agreement should be within 5.25% of human-human
Standardized Mean Difference (Cohen's d)		Engine-human *d* must be no greater than .15 in magnitude
Standard deviation ratio		The ratio of engine score standard deviation to the human score standard deviation should be no less than .90

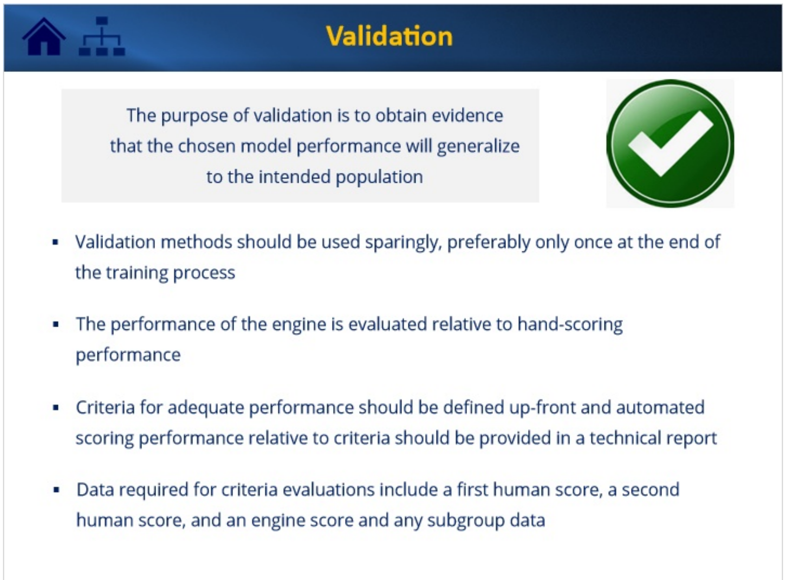
3.11 Training

Training

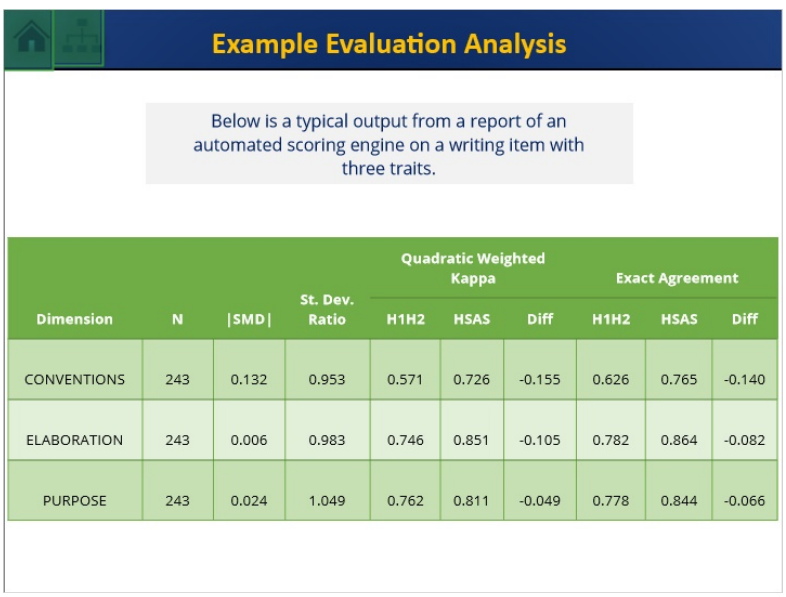
The goal of training is to build competing models and select the best-performing model from among them

- Competing models represent different selections around:
  - ✓ text preprocessing approaches
  - ✓ predictive features
  - ✓ statistical models
- Each model should reflect a theoretical approach that is aligned to the scoring rubric
- Model performance is evaluated relative to hand-scoring performance
- Model selection should consider errors in measurement in addition to thresholds
- Model selection should be done on a validation sample, not the test sample
- If models are combined, separate samples should be used to estimate parameters

3.12 Validation

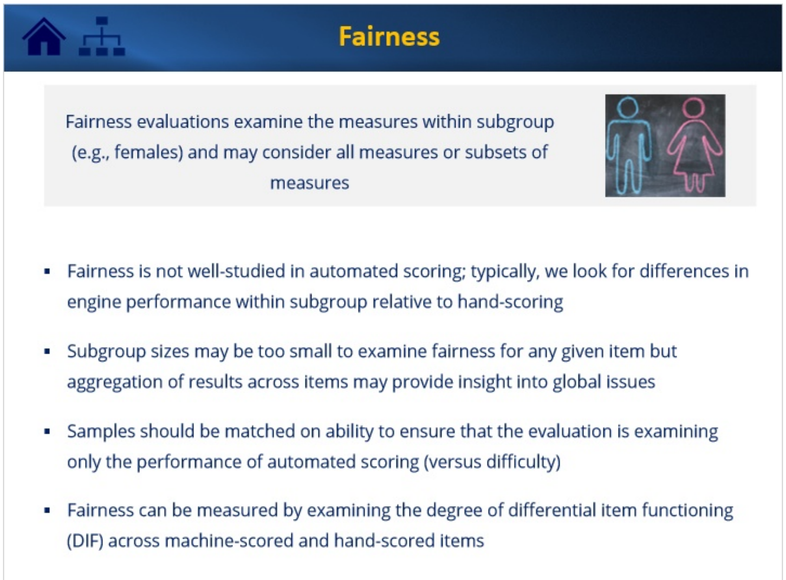


3.13 Example

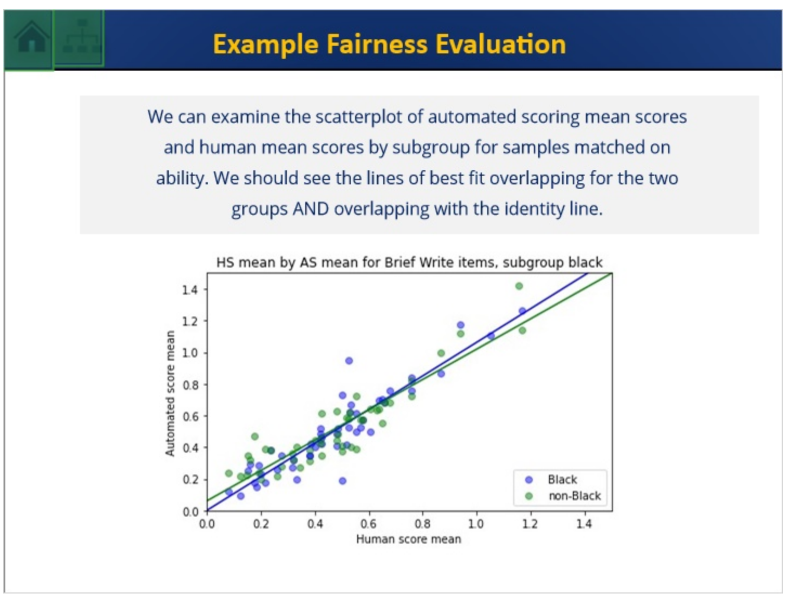


Dimension	N	|SMD|	St. Dev. Ratio	Quadratic Weighted Kappa			Exact Agreement		
				H1H2	HSAS	Diff	H1H2	HSAS	Diff
CONVENTIONS	243	0.132	0.953	0.571	0.726	-0.155	0.626	0.765	-0.140
ELABORATION	243	0.006	0.983	0.746	0.851	-0.105	0.782	0.864	-0.082
PURPOSE	243	0.024	1.049	0.762	0.811	-0.049	0.778	0.844	-0.066

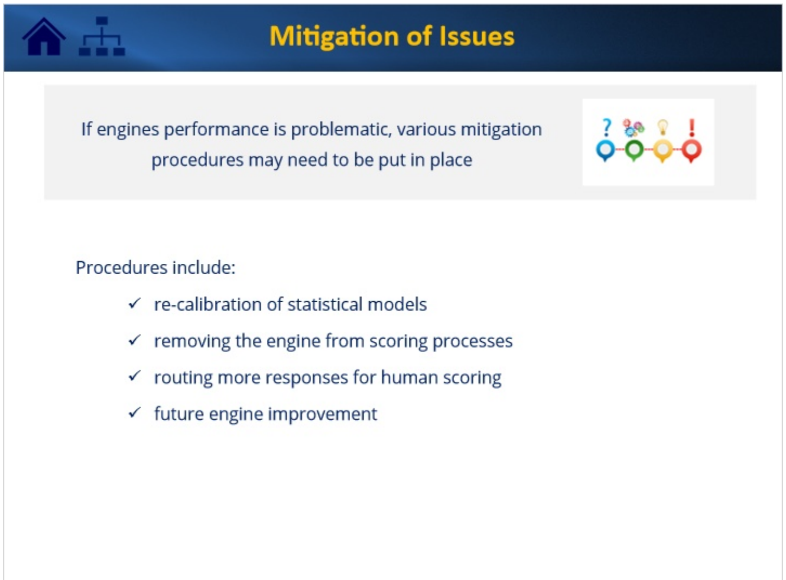
3.14 Fairness



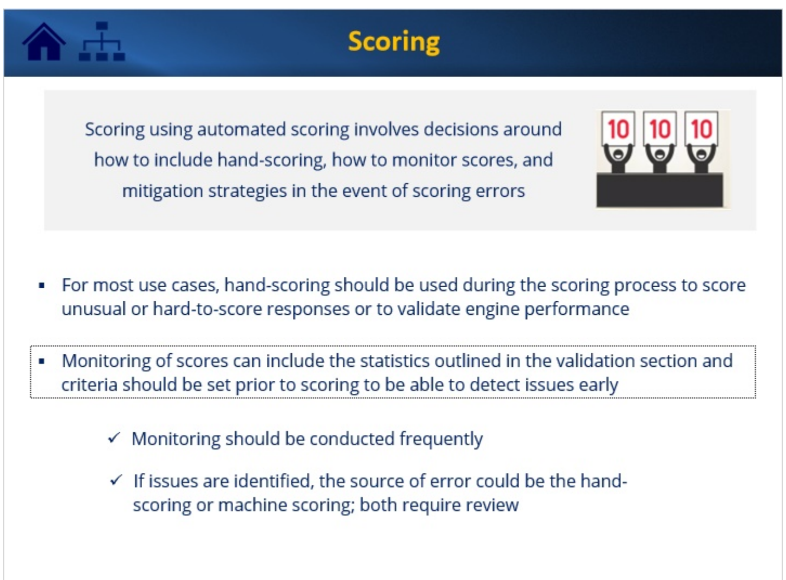
3.15 Fairness Example



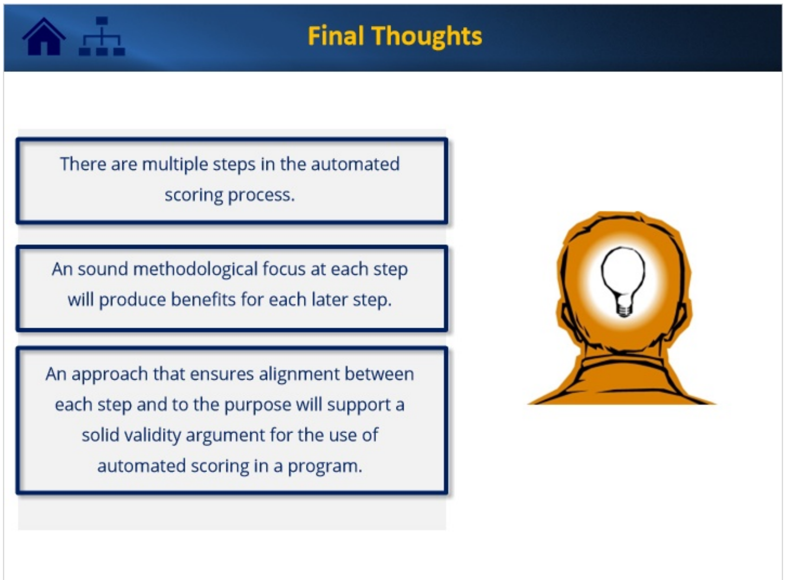
3.16 Mitigation of Issues



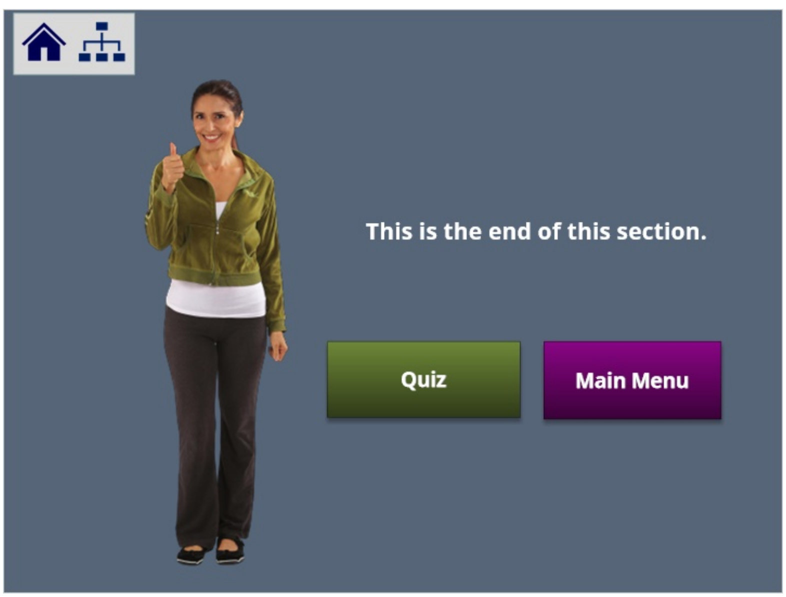
3.17 Scoring



3.18 Final Thoughts



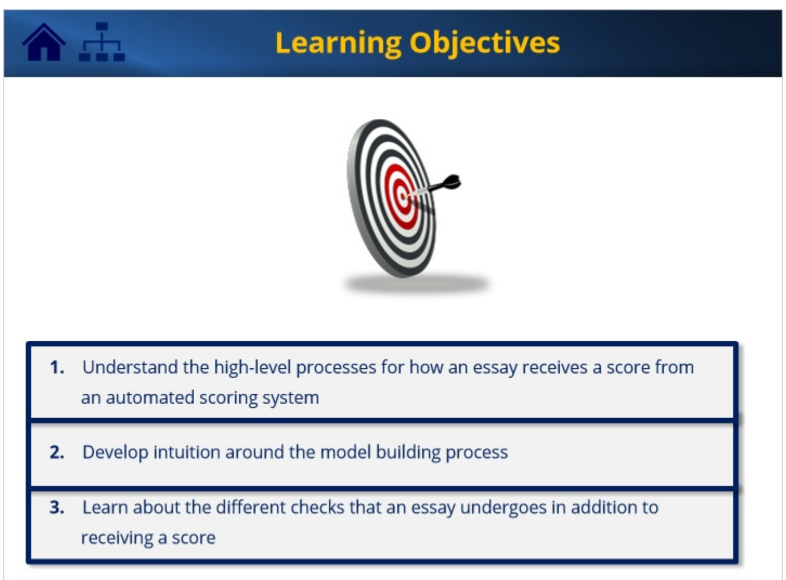
3.19 Bookend: Section 2



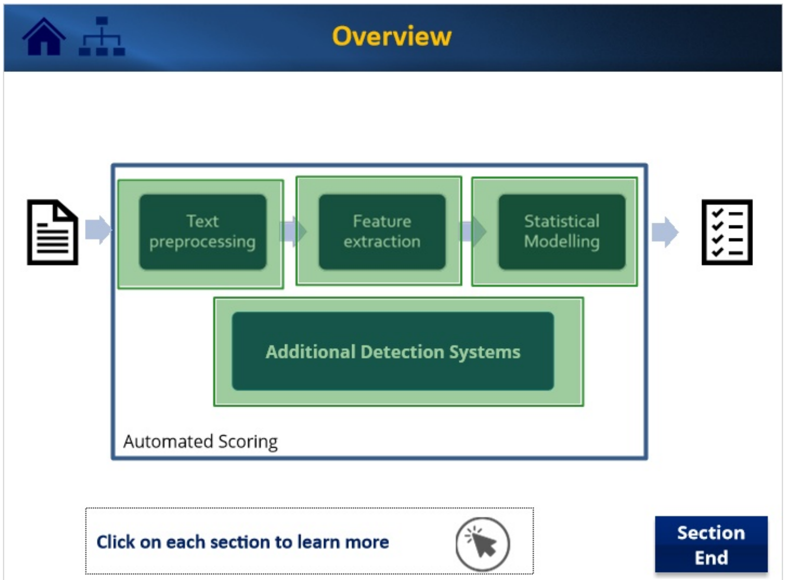
4. Section 3: Engine Design

4.1 Cover: Section 3

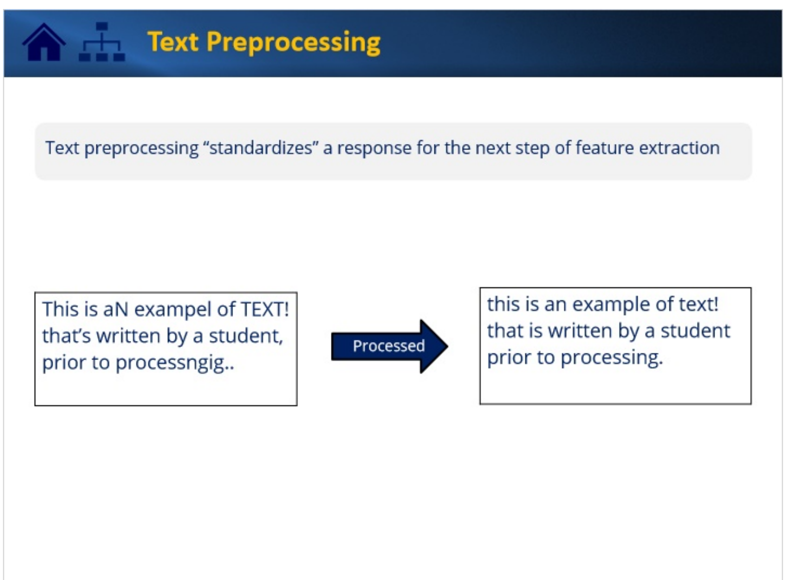
4.2 Learning Objectives



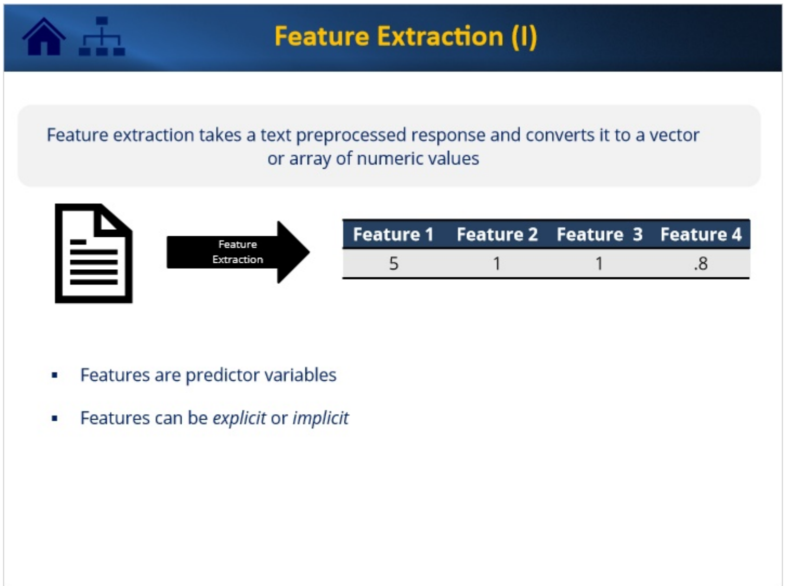
4.3 Topic Selection



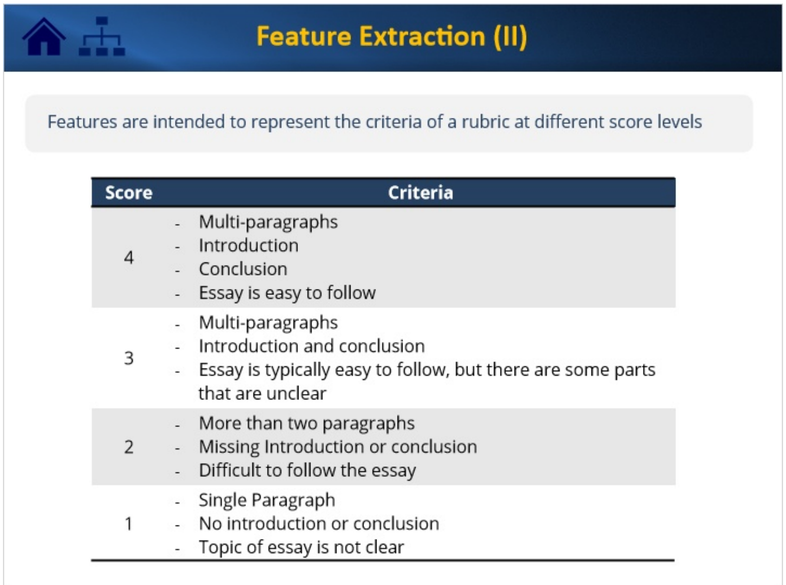
4.4 Preprocessing



4.5 Feature Extraction (I)



4.6 Feature Extraction (II)



4.7 Feature Extraction (III)

Feature Extraction (III)

Each response receives an array of values after the feature extraction phase and a data set is produced as below

Student	Human Score	# of Paragraphs	Introduction	Conclusion	Coherence
1	4	5	1	1	.8
2	3	4	0	1	.6
3	2	3	1	0	.3
4	1	1	0	0	.1
...	...	...	...	...	...

Specified Model:

$$\widehat{Score_i} = \beta_0 + \beta_1(Paragraphs_i) + \beta_2(Introduction_i) + \beta_3(Conclusion_i) + \beta_4(Coherence_i)$$

4.8 Statistical Modeling

Score Prediction

The feature values are entered into a predictive model to produce scores

$$\widehat{Score_i} = 0.5 + .05(Paragraphs_i) + 1(Introduction_i) + .75(Conclusion_i) + 2(Coherence_i)$$

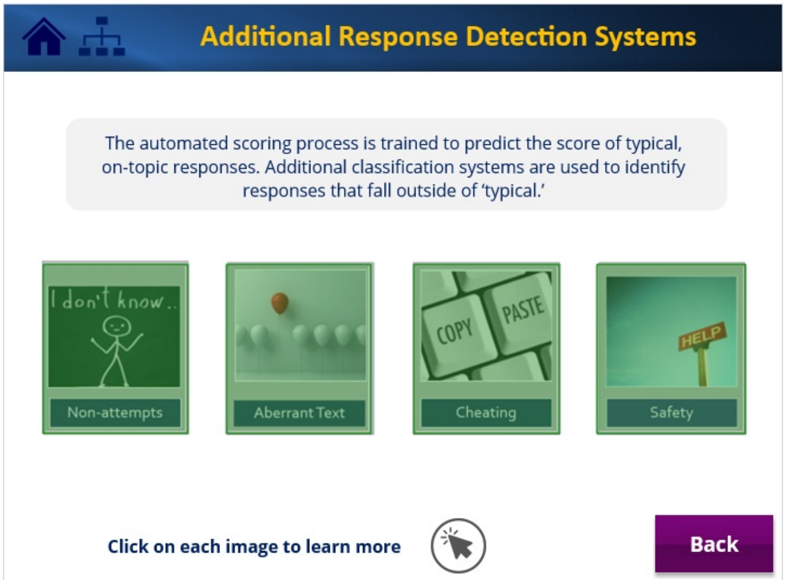
$$\widehat{Score_1} = 0.5 + .05(5) + 1(1) + .75(1) + 2(.8) = 4.1 \approx 4$$

$$\widehat{Score_2} = 0.5 + .05(4) + 1(0) + .75(1) + 2(.6) = 2.65 \approx 3$$

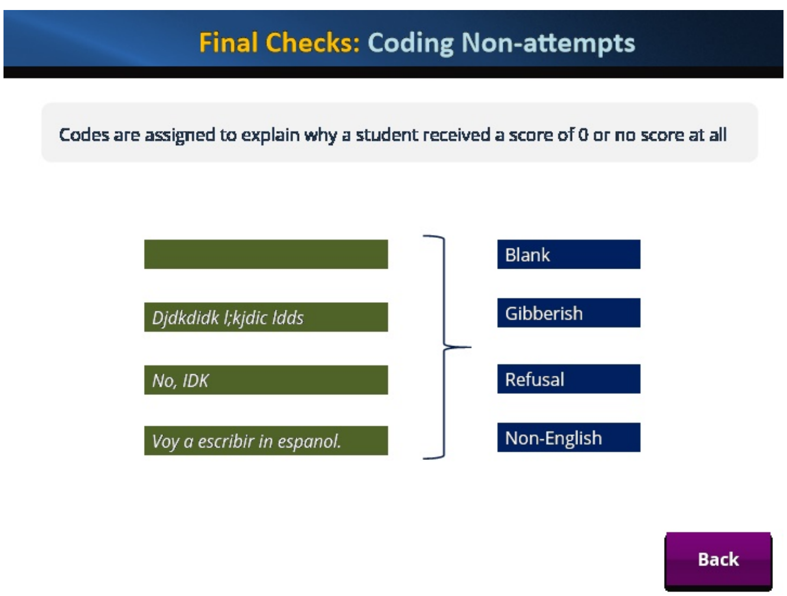
$$\widehat{Score_3} = 0.5 + .05(3) + 1(0) + .75(0) + 2(.3) = 2.25 \approx 2$$

$$\widehat{Score_4} = 0.5 + .05(1) + 1(0) + .75(0) + 2(.1) = 0.75 \approx 1$$

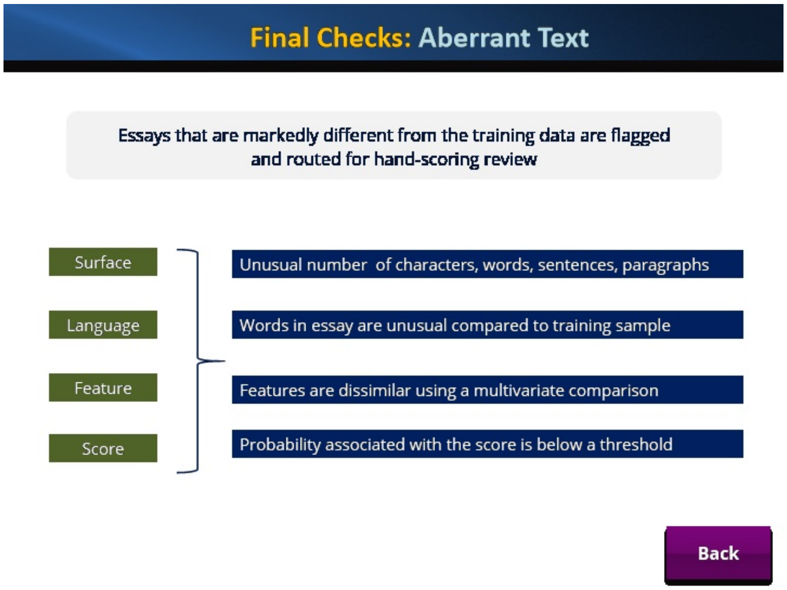
4.9 Final Checks



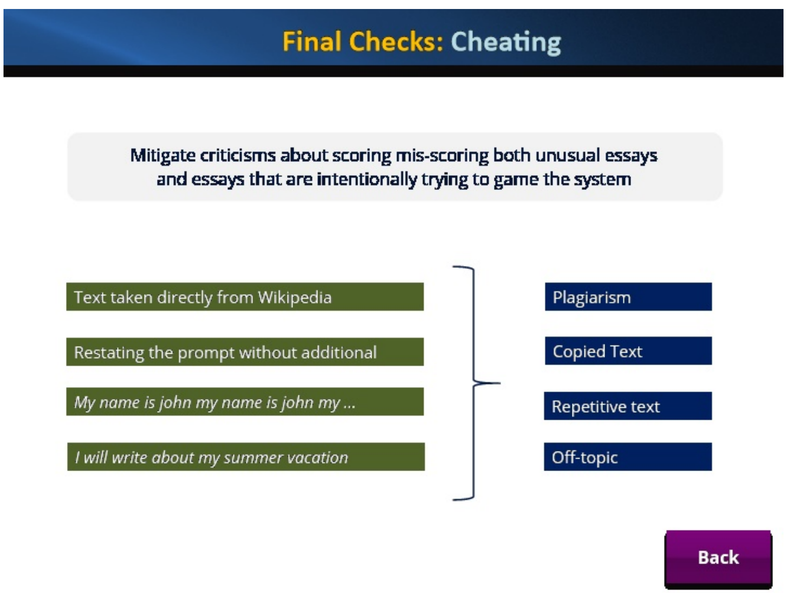
Nonattempts (Slide Layer)



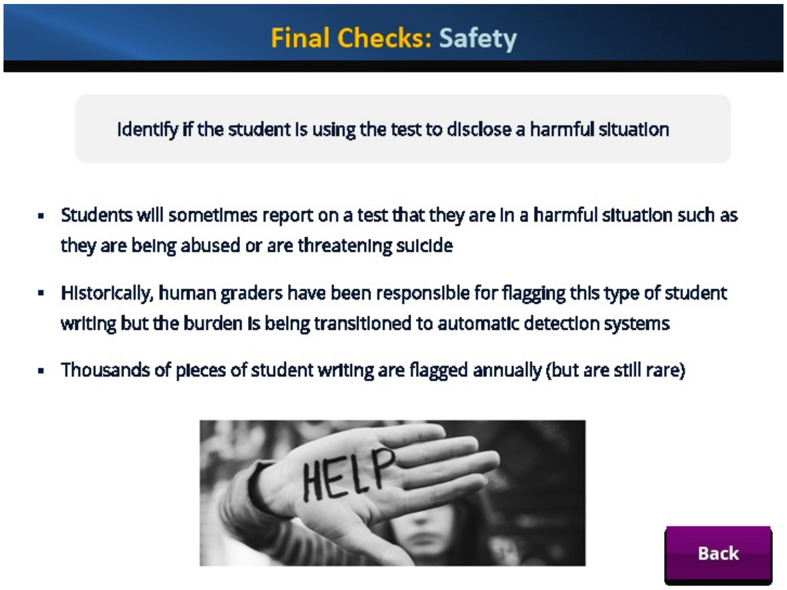
Aberrant Text (Slide Layer)



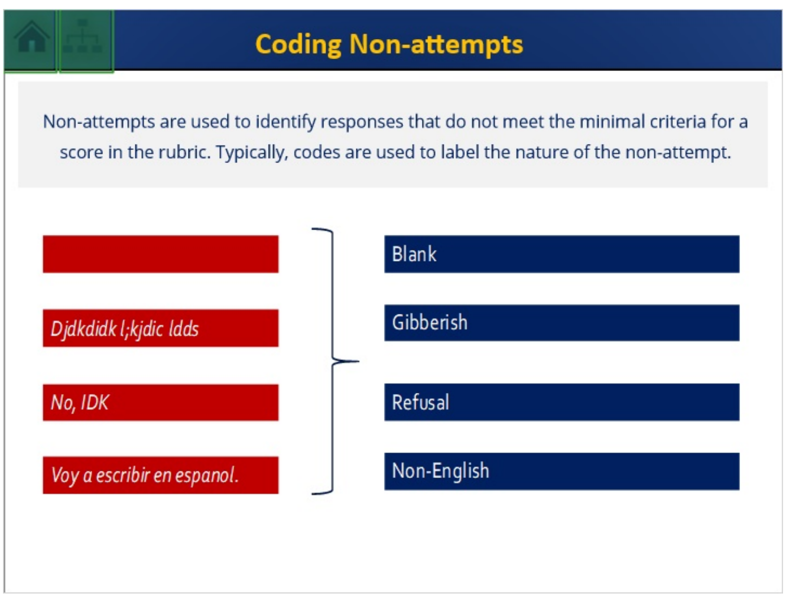
Cheating (Slide Layer)



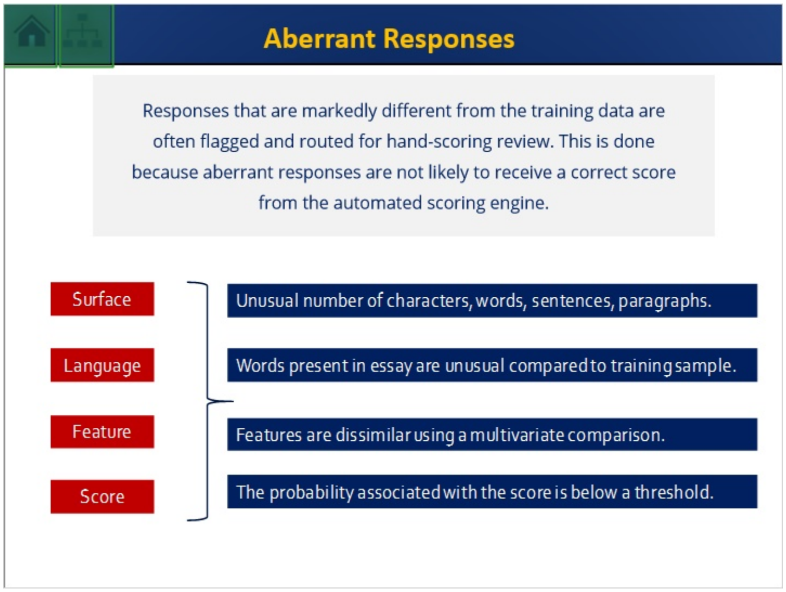
Safety (Slide Layer)



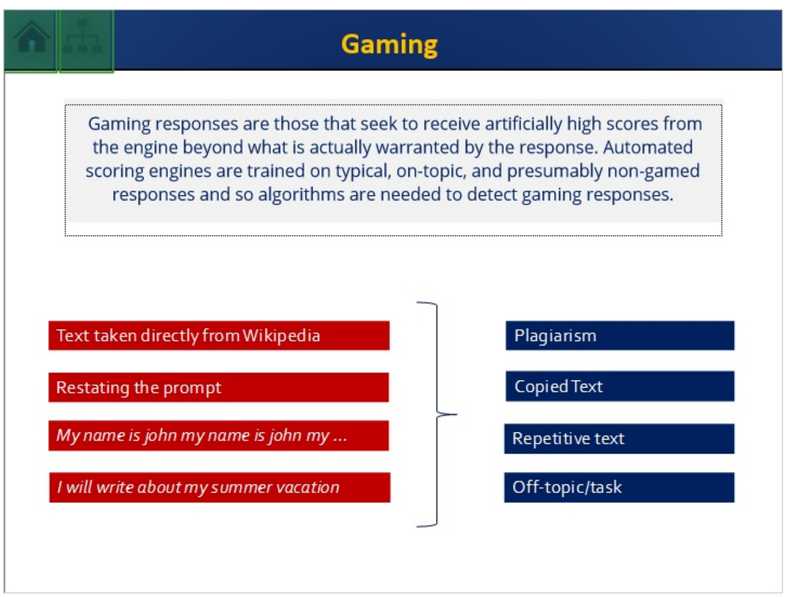
4.10 Non Attempts



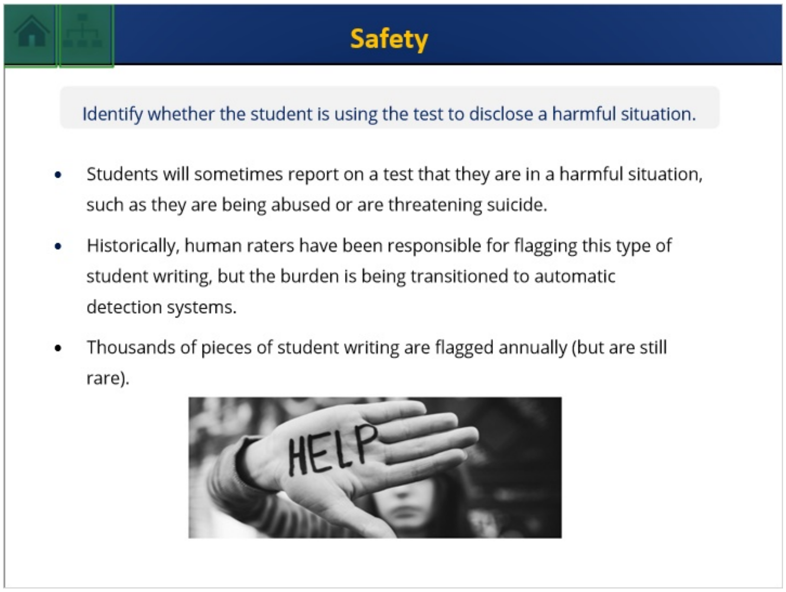
4.11 Aberrant Responses



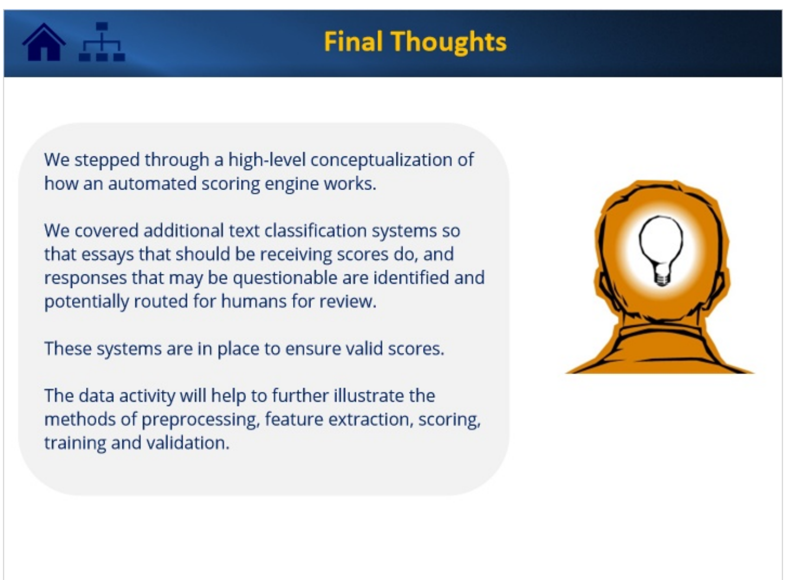
4.12 Gaming Responses



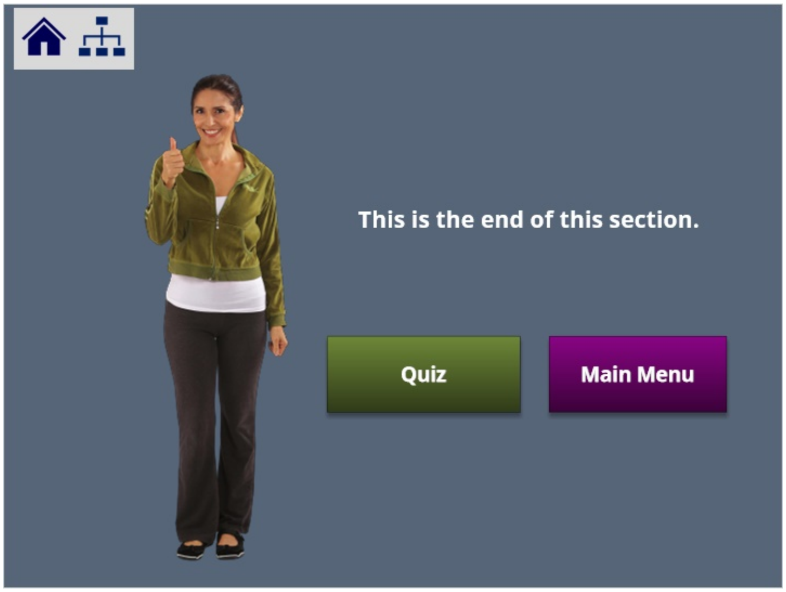
4.13 Safety



4.14 Final thoughts



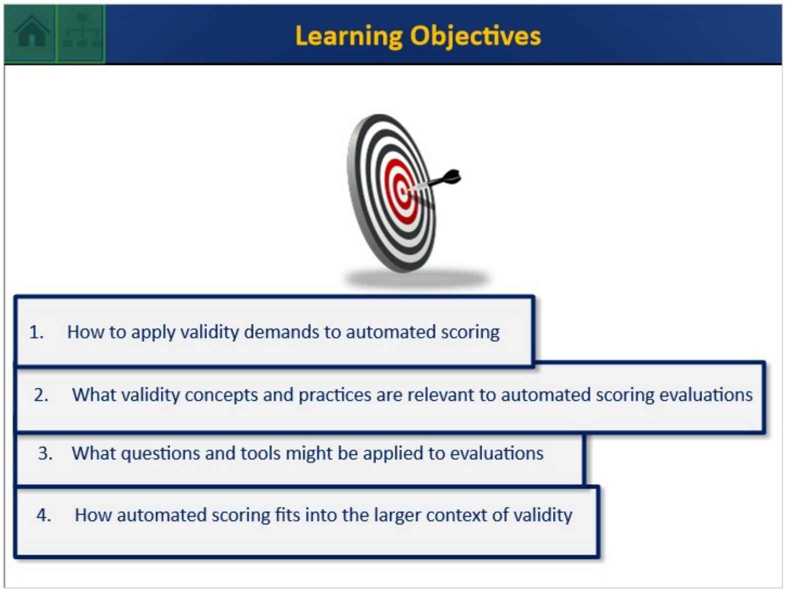
4.15 Bookend: Section 3



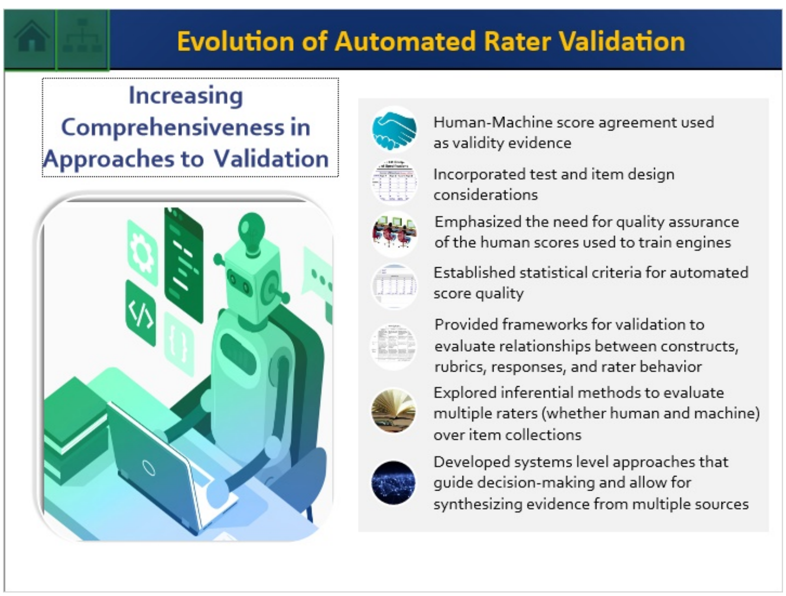
5. Section 4: Validation

5.1 Cover: Section 4

5.2 Learning Objectives



5.3 Evolution



5.4 Validation

Score Validity and Sources of Evidence

Validation in automated scoring sits in the larger context of test score validity, where the AERA, APA, & NCME *Standards for Educational and Psychological Testing* (*Standards*, 2014) defines validity as:

"…the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests."

As a reminder, the *Standards* outline five sources of validity evidence which serve as a practical basis for thinking about how to collect and interpret evidence supporting the use of test scores.

1. Content validity
2. Internal structure
3. Response processes
4. Relationships with other variables
5. Consequences of use

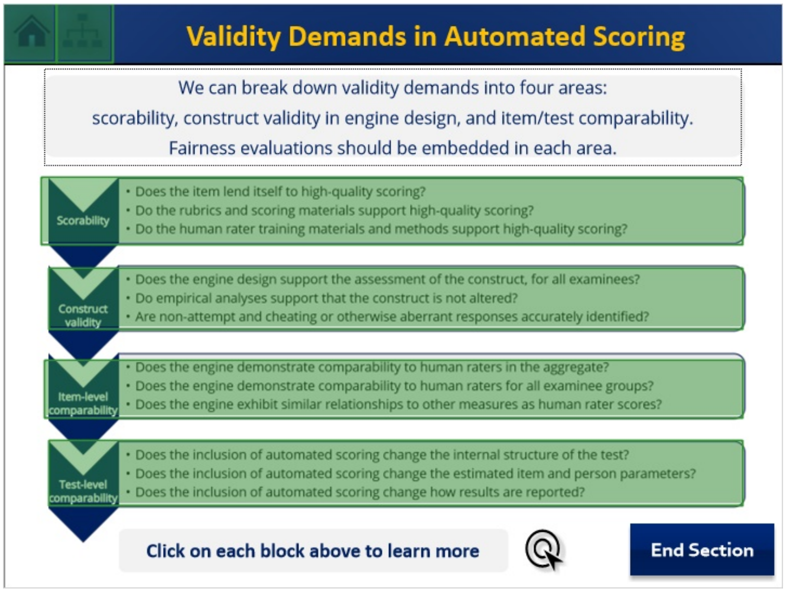
5.5 Validation Process

Validation Process

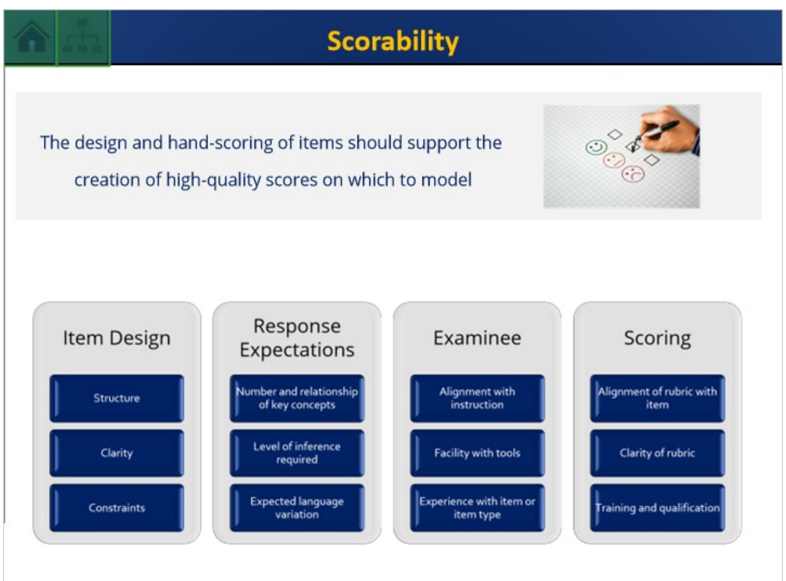
Validation – as applied to scoring – is an ongoing process in which evidence is collected over time to examine the degree to which automated scoring supports the use and interpretation of assessment results

- Automated scores are interchangeable with human rater scores
- Validation of automated scoring is similar to validation of human scoring
- Analyses depend upon:
  - ✓ the ways in which scores are used in the program for operational reporting
  - ✓ the ways in which human and automated scores are combined

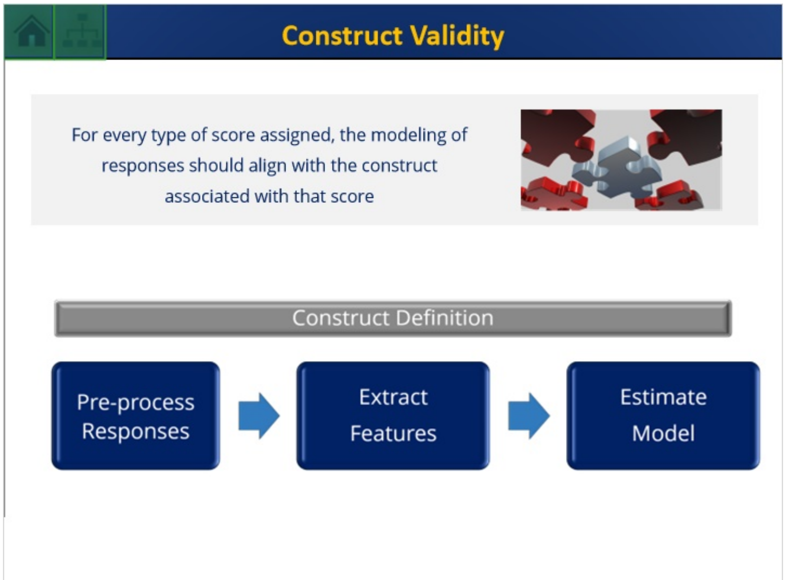
5.6 Topic Selection



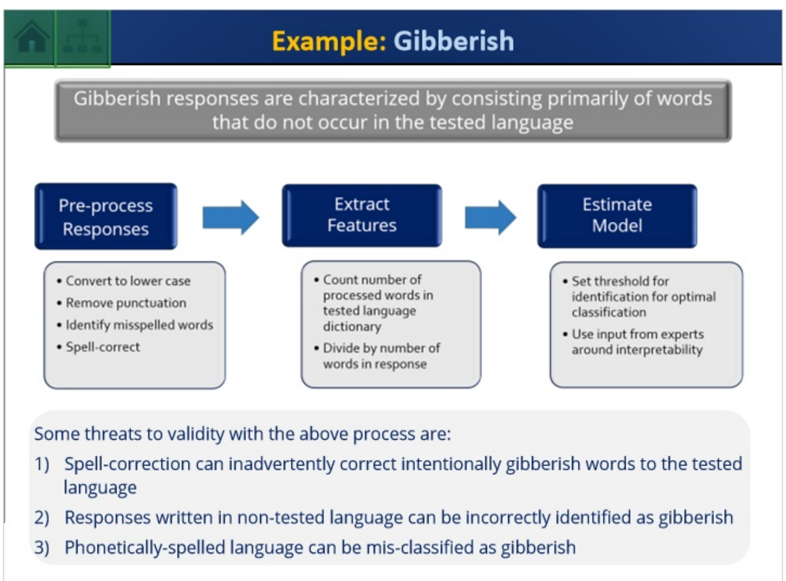
5.7 Scorabilty



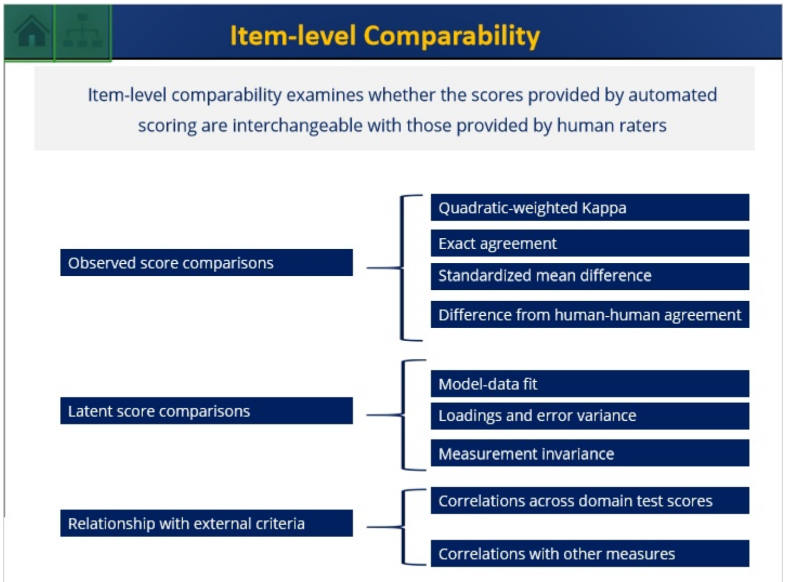
5.8 Construct Validity



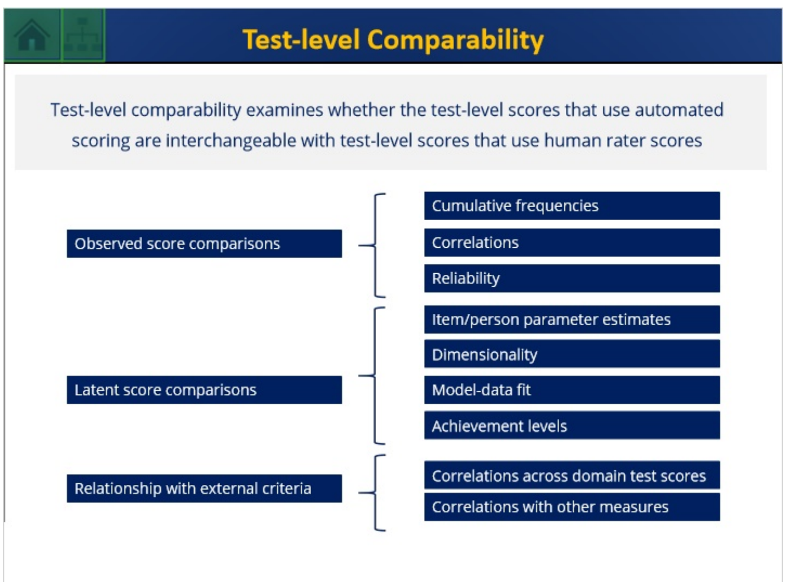
5.9 Example



5.10 Item-level Comparability

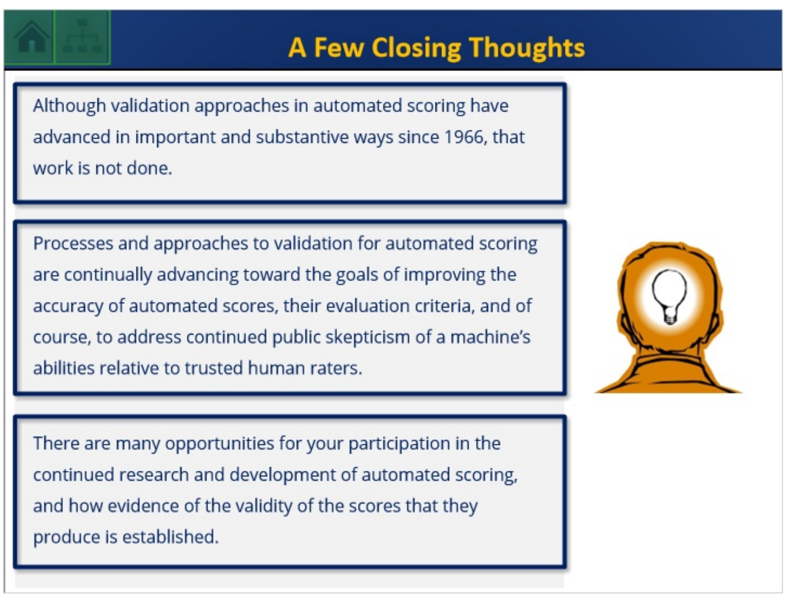


5.11 Test-level Comparability

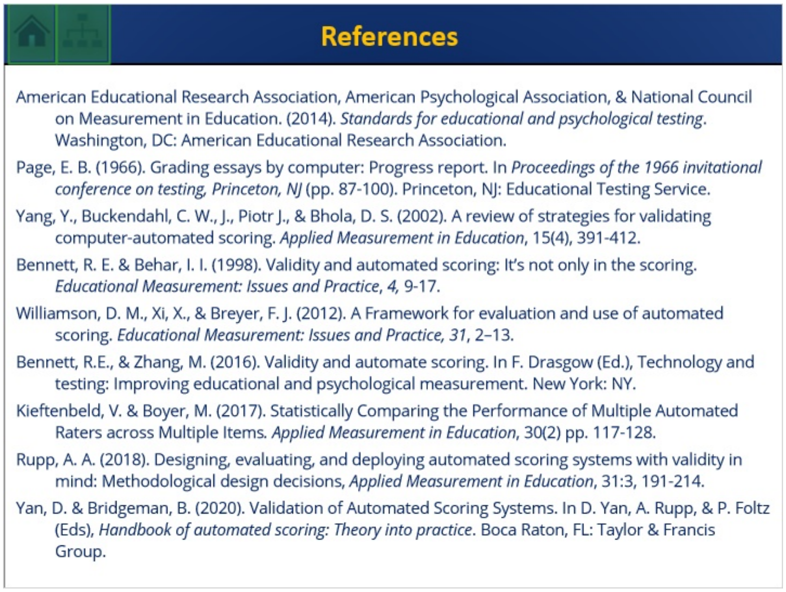


5.12 Standards

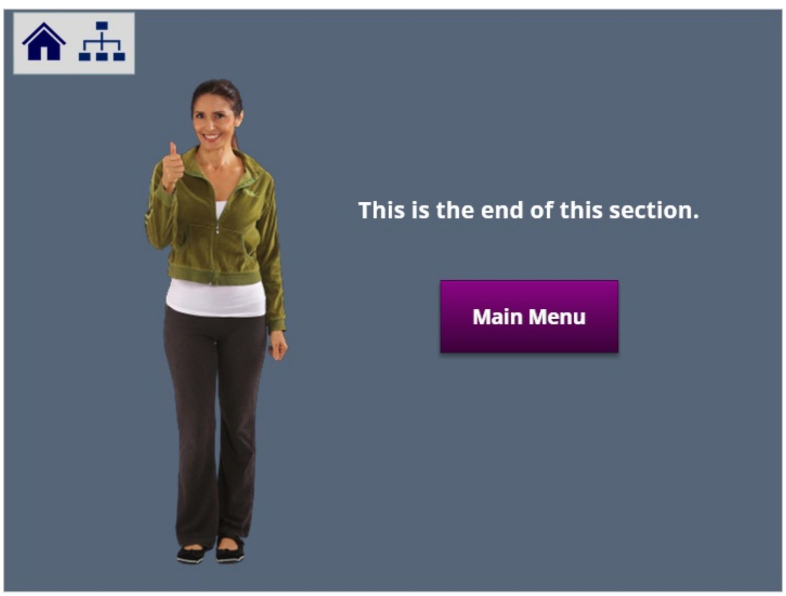
5.13 Closing Thoughts



5.14 References



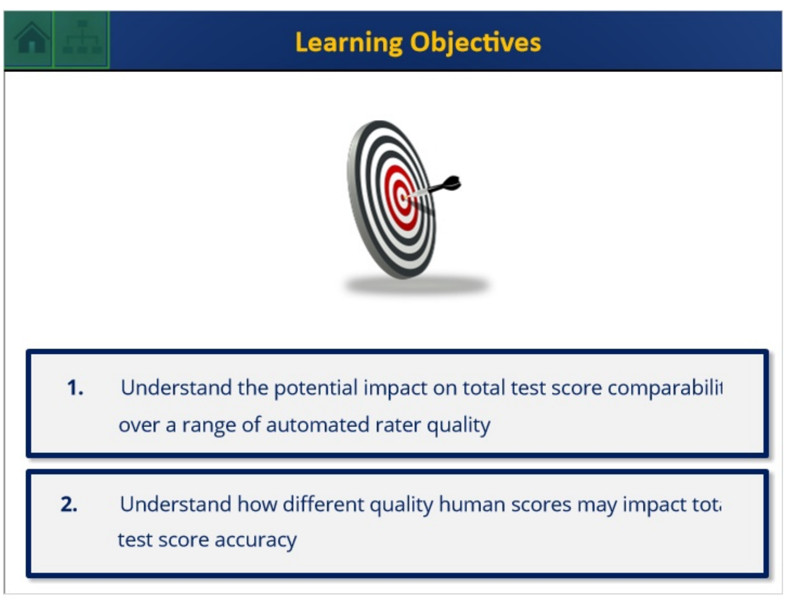
5.15 Bookend: Section 4



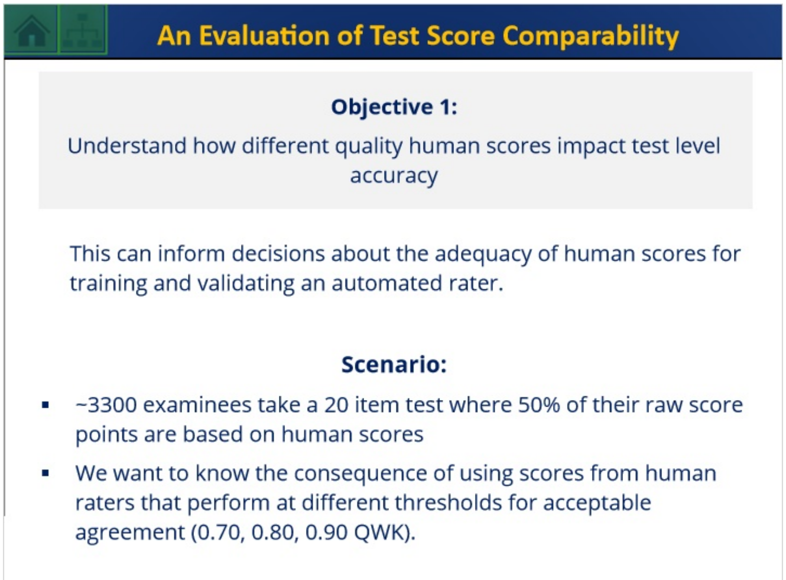
6. Section 5: Worked Example

6.1 Worked Example

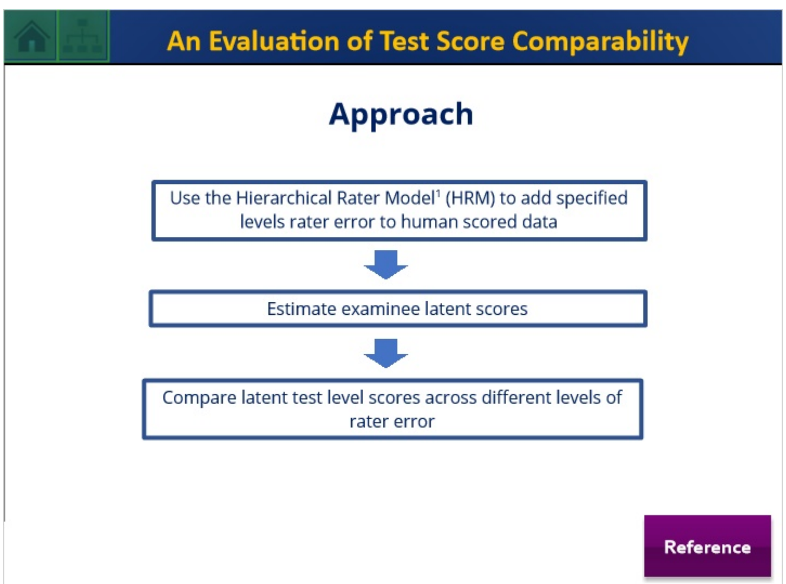
6.2 Learning Objectives



6.3 Objective 1



6.4 Approach



Reference (Slide Layer)

Journal of Educational and Behavioral Statistics
Winter 2002, Vol. 27, No. 4, pp. 341–384

The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data

Richard J. Patz
CTB/McGraw-Hill

Brian W. Junker
Carnegie Mellon University

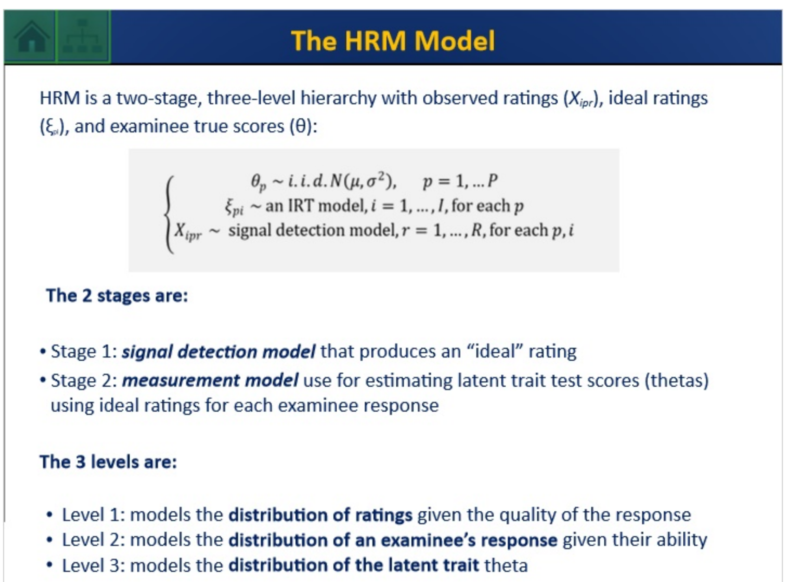
Matthew S. Johnson
Baruch College

Louis T. Mariano
RAND

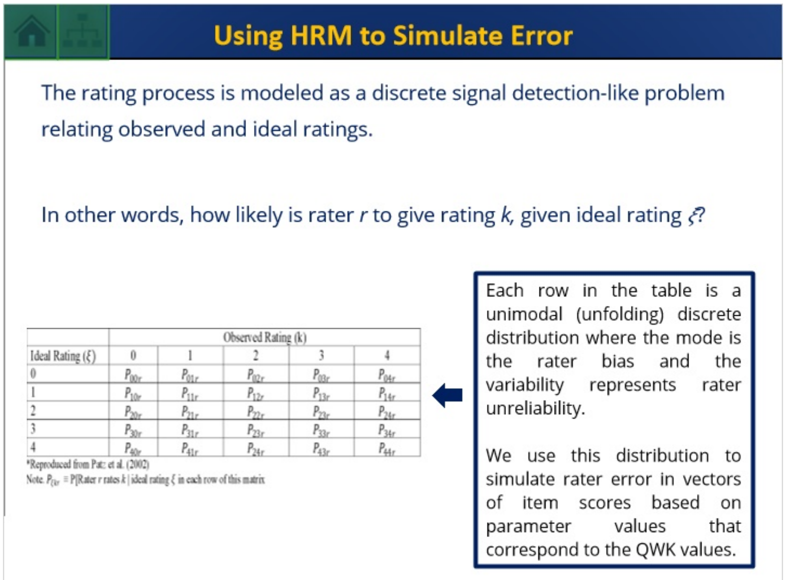
Open-ended or "constructed" student responses to test items have become a stock component of standardized educational assessments. Digital imaging of examinee work now enables a distributed rating process to be flexibly managed, and allocation designs that involve as many as six or more ratings for a subset of responses are now feasible. In this article we develop Patz's (1996) hierarchical rater model (HRM) for polytomous item response data scored by multiple raters, and show how it can be used to scale examinees and items, to model aspects of consensus among raters, and to model individual rater severity and consistency effects. The HRM treats examinee responses to open-ended items as unobsered discrete varibles, and it explicitly models the "proficiency" of raters in assigning accurate scores as well as the proficiency of examinees in providing correct responses. We show how the HRM "fits in" to the generalizability theory framework that has been the traditional tool of analysis for rated item response data, and give some relationships between the HRM, the design effects correction of Bock, Brennan and Muraki (1999), and the rater bundle model of Wilson and Hoskens (2002). Using simulated and real data, we compare the HRM to the conventional IRT Facets model for rating data (e.g., Linacre, 1989; Engelhard, 1994, 1996), and we explore ways that information from HRM analyses may improved the quality of the rating process.

Keywords: *generalizability, hierarchical Bayes modeling, item response theory, latent response model, Markov chain Monte Carlo, Multiple ratings, rater consensus, rater consistency, rater severity*

6.5 HRM

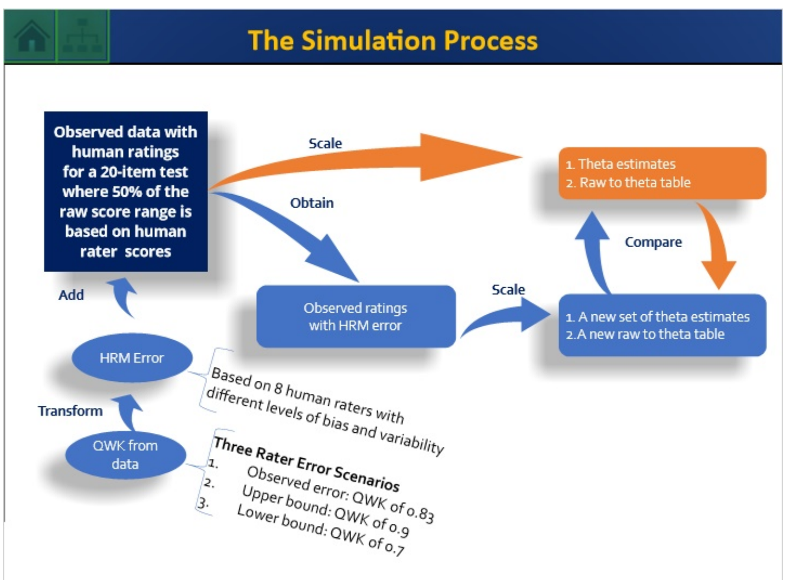


6.6 Signal Detection

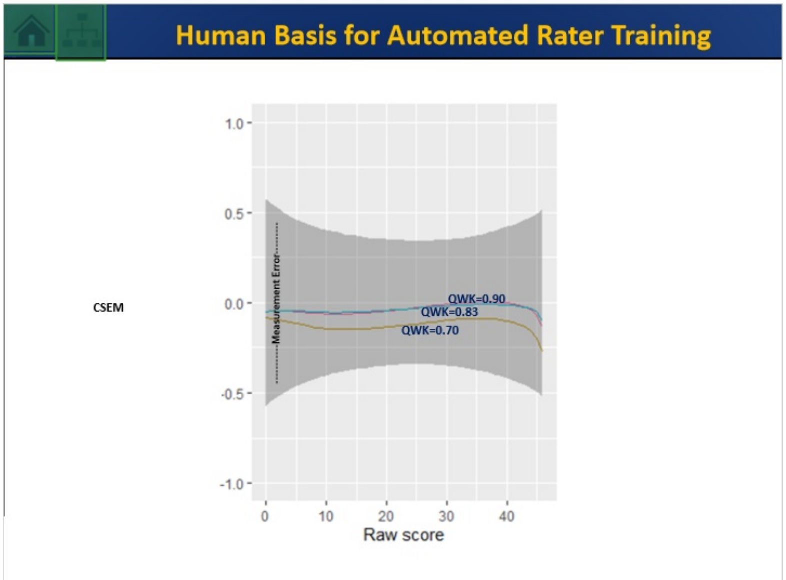


	Observed Rating (k)				
Ideal Rating (ξ)	0	1	2	3	4
0	P_{00r}	P_{01r}	P_{02r}	P_{03r}	P_{04r}
1	P_{10r}	P_{11r}	P_{12r}	P_{13r}	P_{14r}
2	P_{20r}	P_{21r}	P_{22r}	P_{23r}	P_{24r}
3	P_{30r}	P_{31r}	P_{23r}	P_{33r}	P_{34r}
4	P_{40r}	P_{41r}	P_{24r}	P_{43r}	P_{44r}

6.7 Simulation Process



6.8 Objectve1 Results



6.9 Objective1 Summary

What Do These Results Tell Us?

- There is a range of quality that can be present in the human scores used to train automated raters…

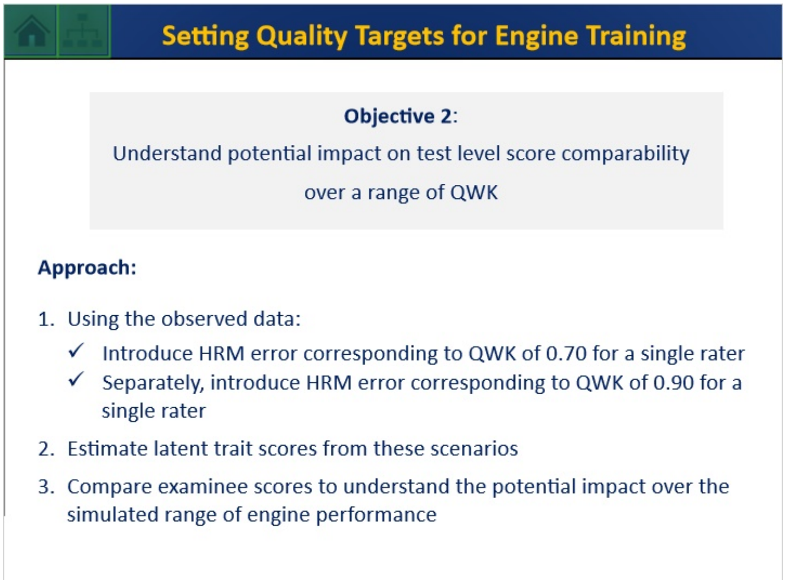
  …and we know from literature that automated raters are likely to reflect the level of quality of the human rater scores used to train and validate the engine

- The impact of different levels of rater error can be consequential for examinees…

  …so we may decide that those consequences are acceptable or they are not acceptable

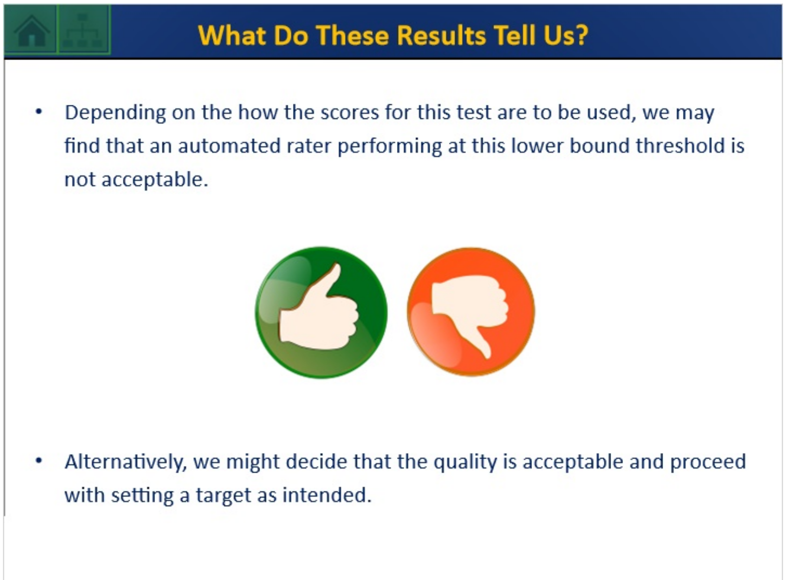
Now let's take this approach one step further and use it to target a range of acceptable score accuracy during automated rater training and validation!

6.10 Objective 2

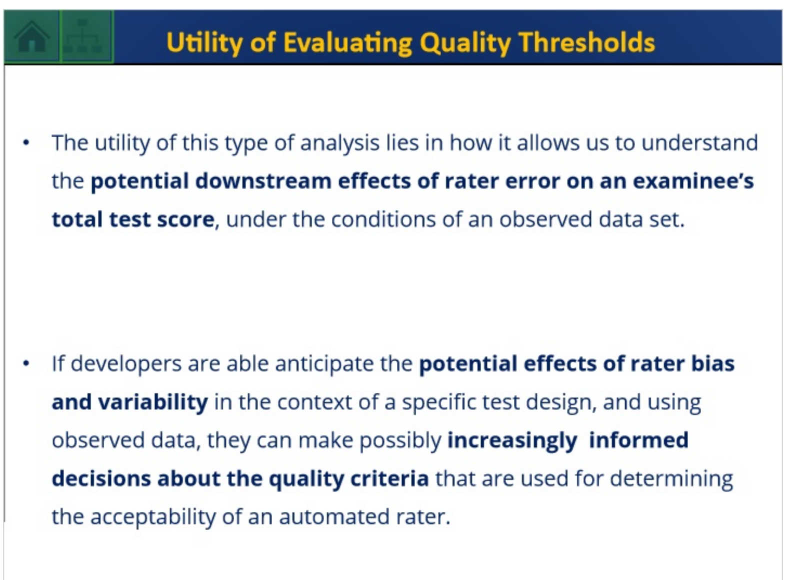


6.11 Objective 2 Results

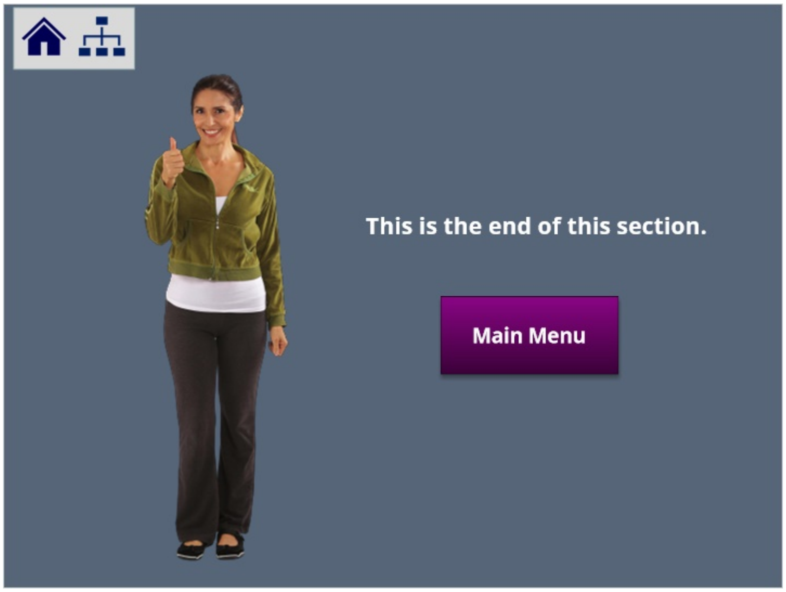
6.12 Objective 2 Summary



6.13 Overall summary



6.14 Bookend: Section 2

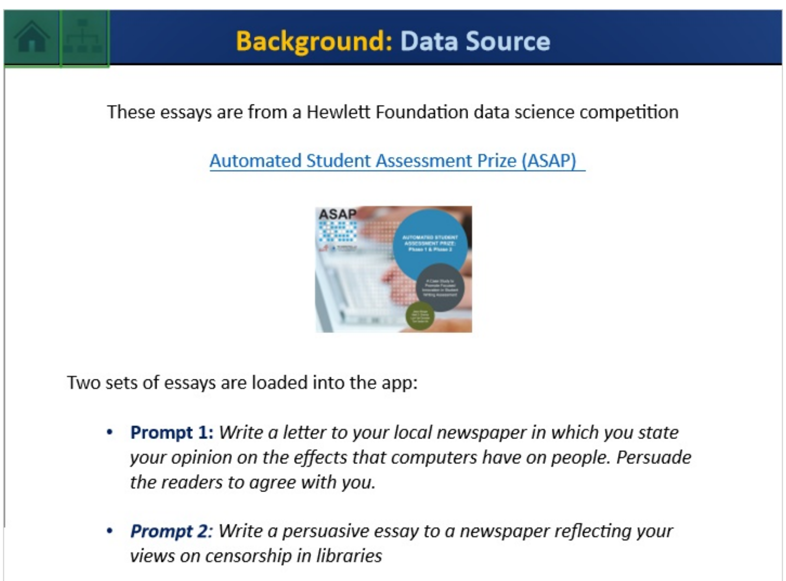


7. Section 6: Data Activity

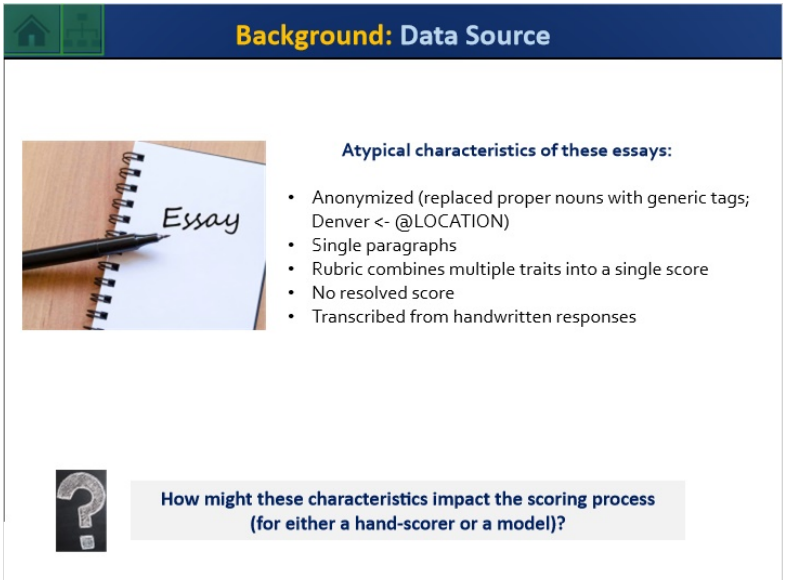
7.1 Cover: Section 6

7.2 Learning Objectives

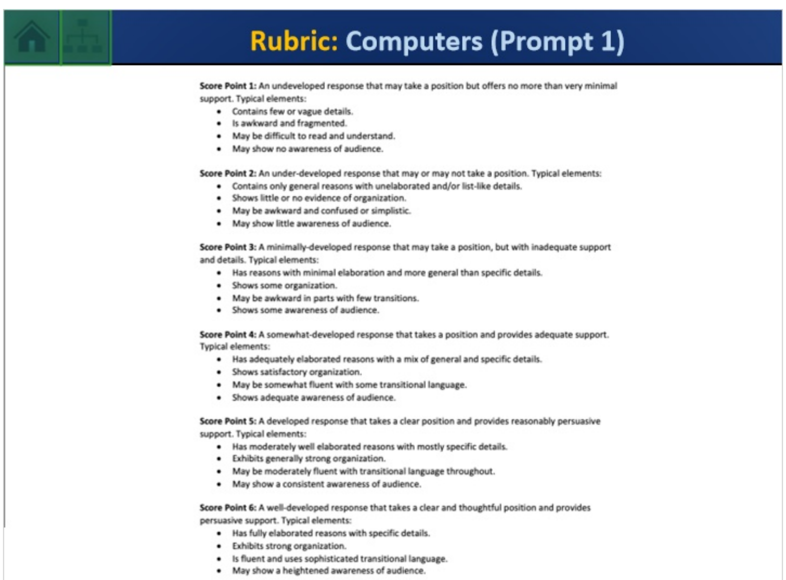
7.3 ASAP items



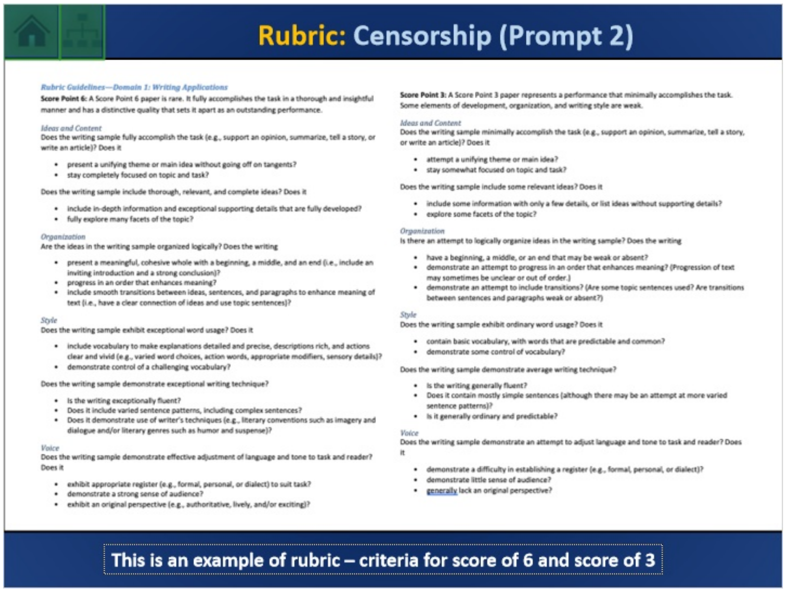
7.4 ASAP essays



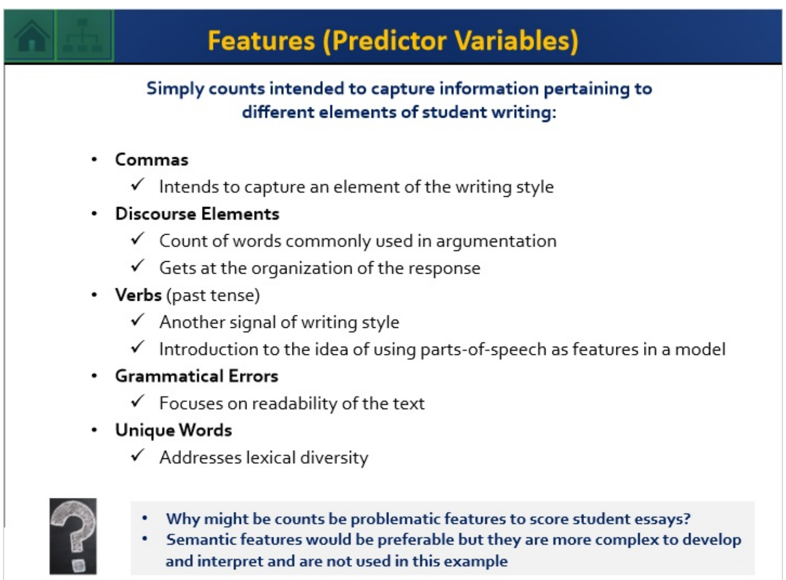
7.5 Prompt 1 Rubric



7.6 Prompt 2 Rubric

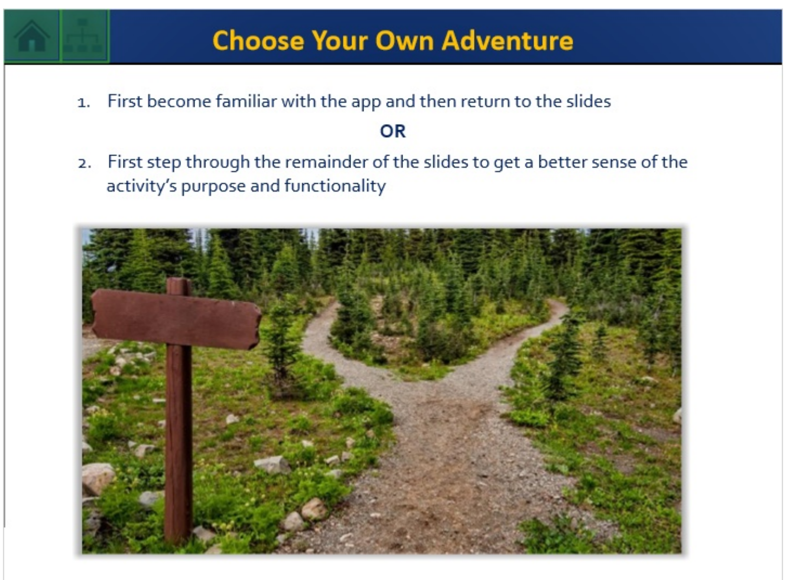


7.7 Features

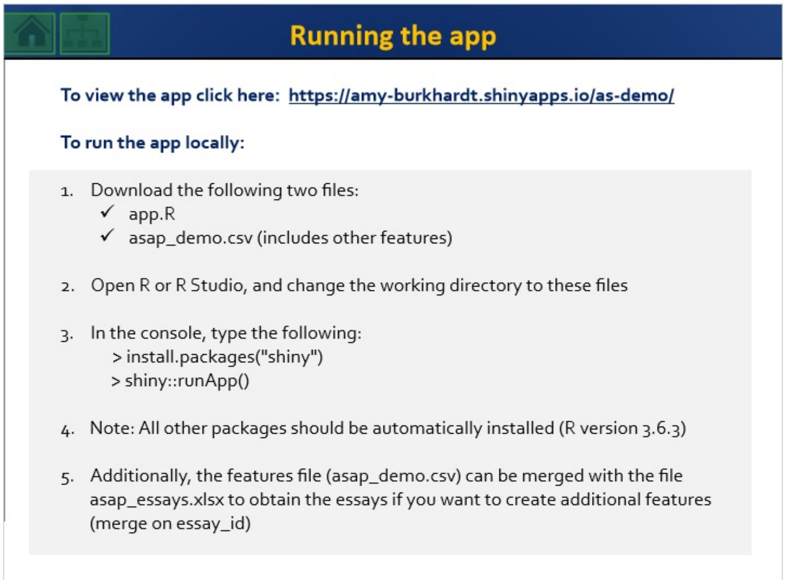


7.8 Reminder

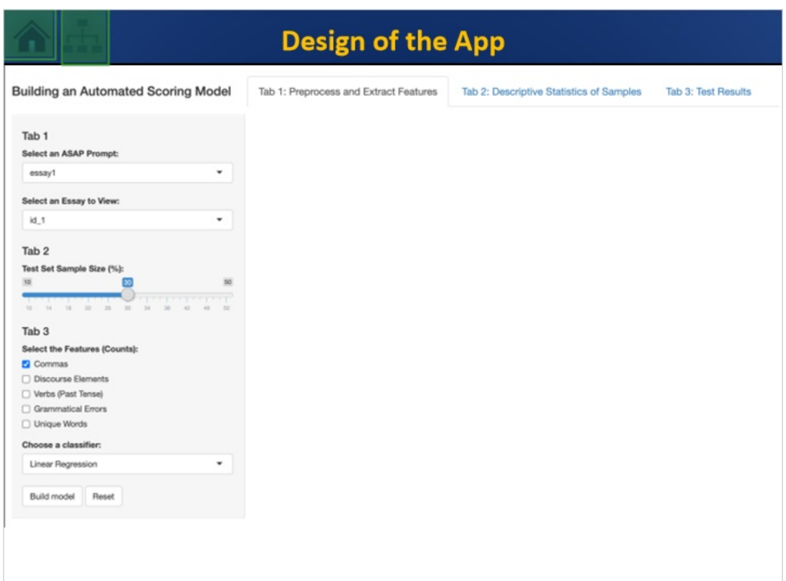
7.9 Choose adventure



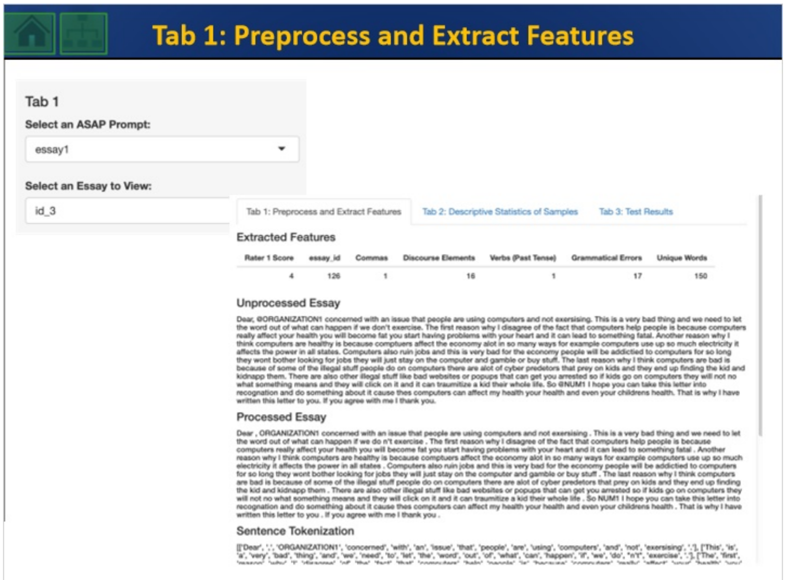
7.10 Running the App



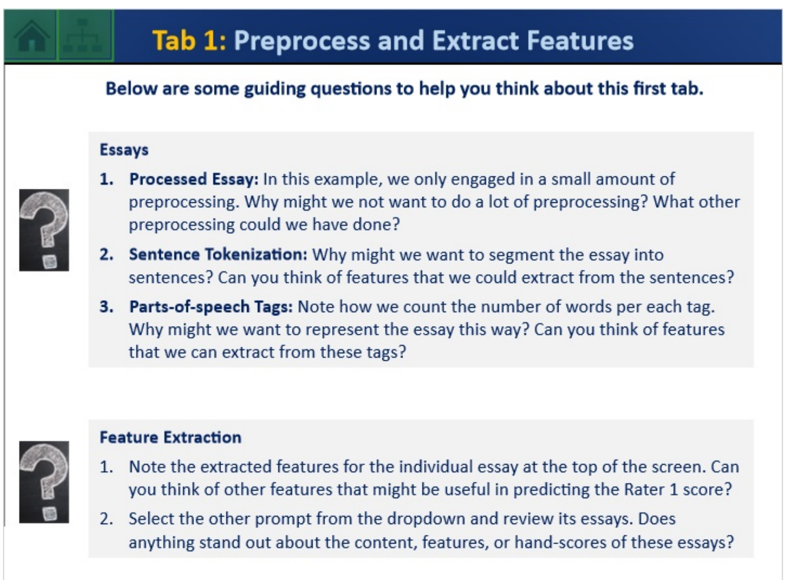
7.11 App Design



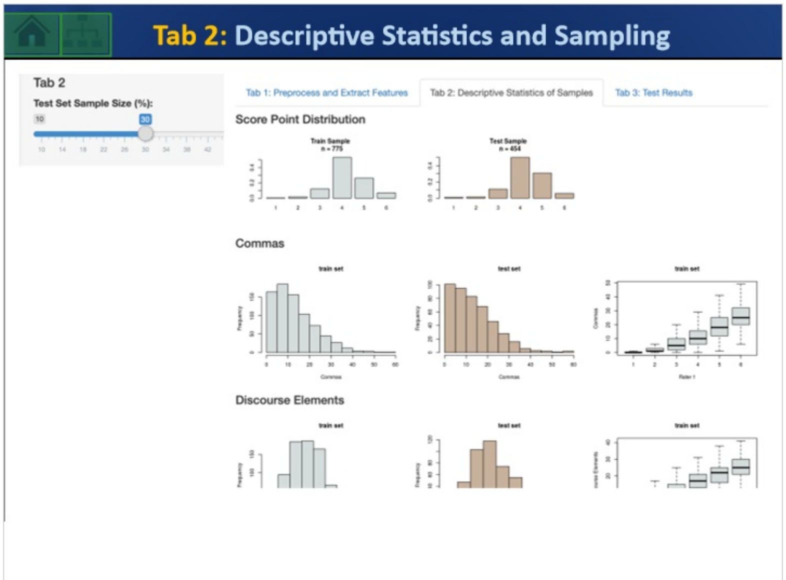
7.12 Tab 1



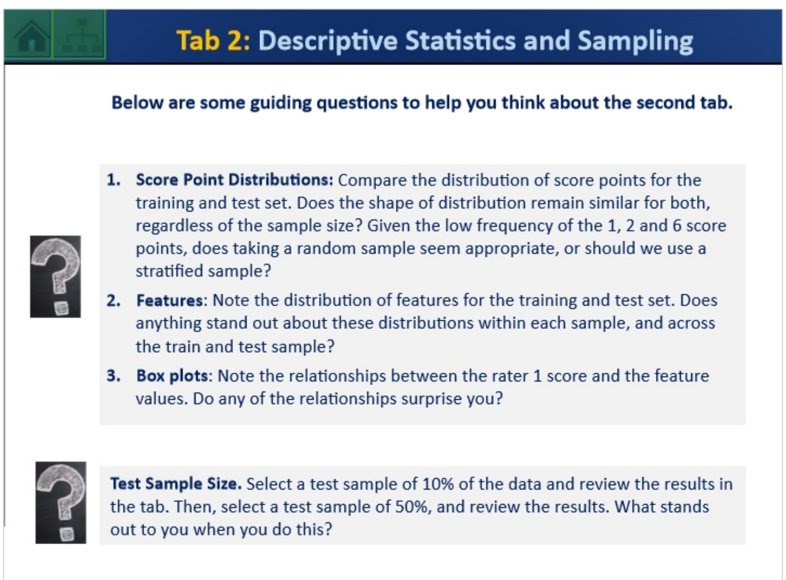
7.13 Tab 1 Questions



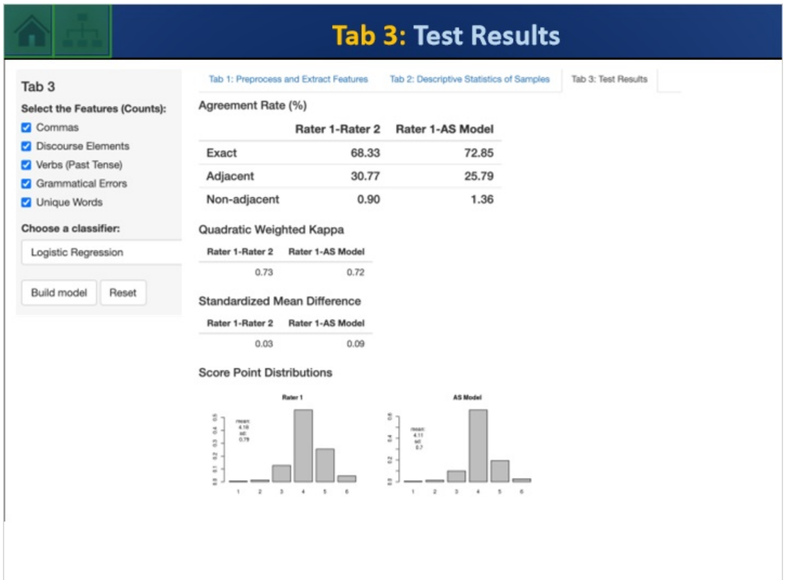
7.14 Tab 2



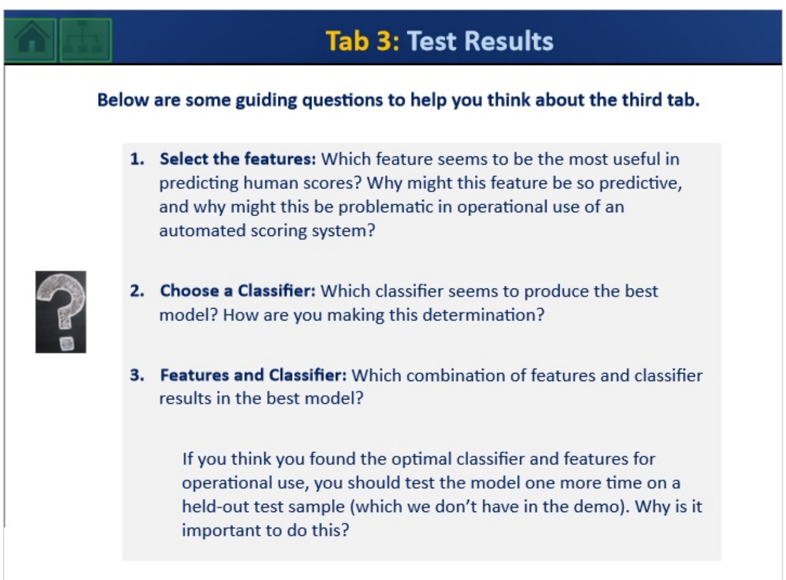
7.15 Tab 2 Questions



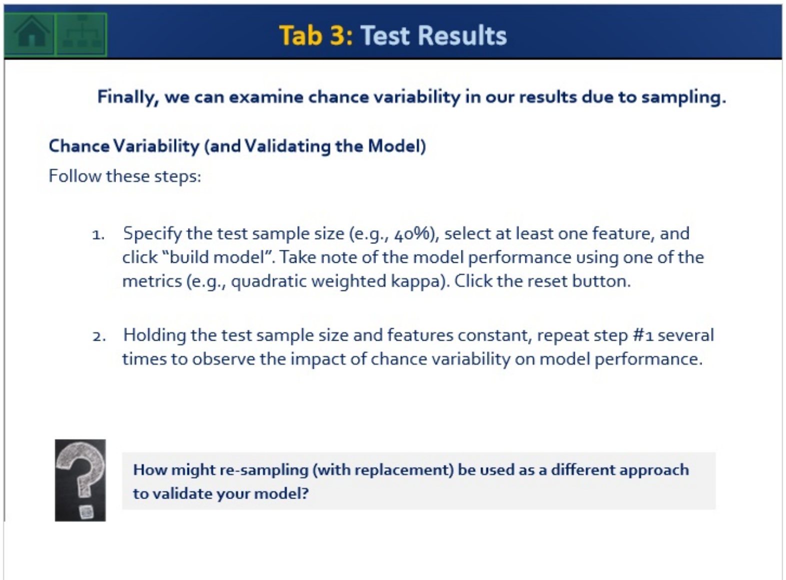
7.16 Tab 3



7.17 Tab 3 Questions



7.18 Tab 3 Questions (2)



7.19 Bookend: Section 2

7.20 Module Cover (END)