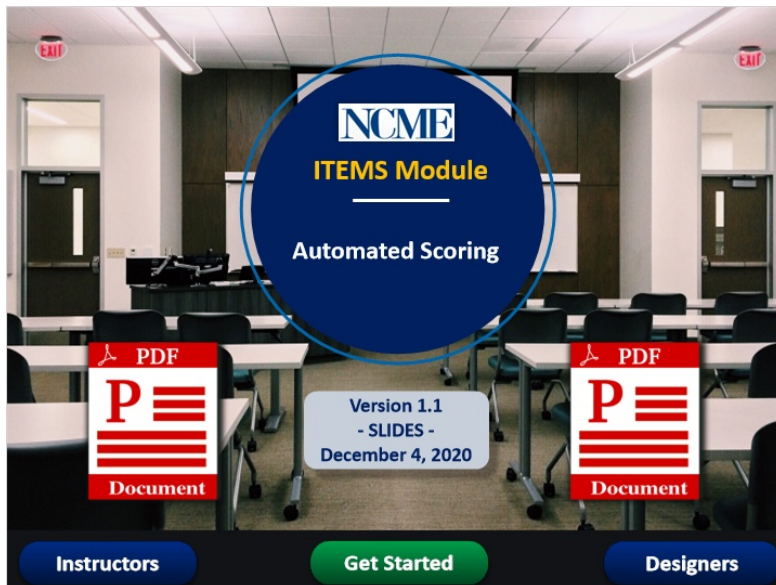


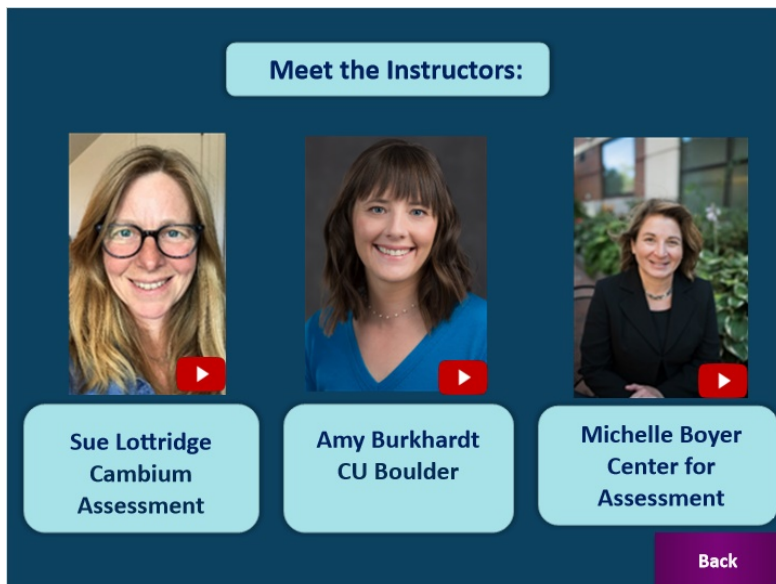
DM18 SLIDES (Automated Scoring, Version 1.1)

1. Module Overview

1.1 Module Cover (START)




1.2 Instructors



1.3 Designers

Meet the designer:



André A. Rupp
Mindful Measurement

Back

Andre V1 (Slide Layer)

1.4 Welcome



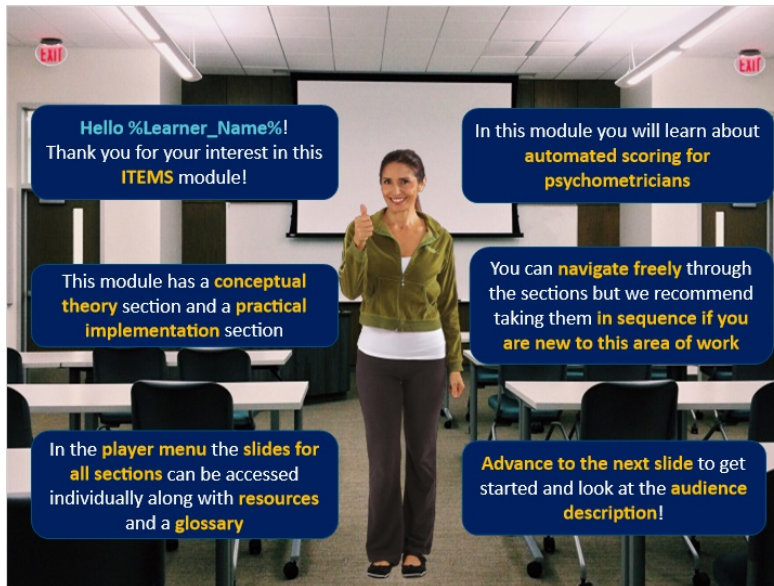
**Welcome to the
ITEMS Module!**

The woman to the left is Nora!

Along with the instructors
she will be guiding you
through the module content

Type your name below:

1.5 Overview



1.6 Target Audience

Target Audience

Anyone who would like a gentle conceptual introduction to this topic:

- graduate students / faculty in Master's, Ph.D., or certificate programs
- psychometricians and other measurement professionals
- data scientists / analysts
- research assistants or research scientists
- technical project directors
- assessment developers



However, we hope that you find the information in this module useful no matter what your official title or role in an organization is!

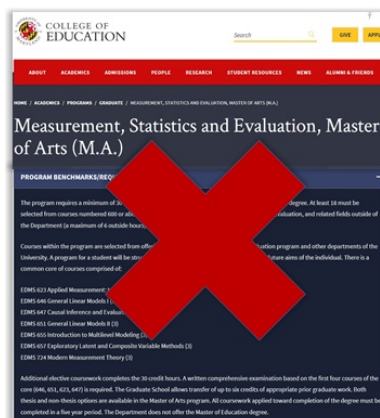
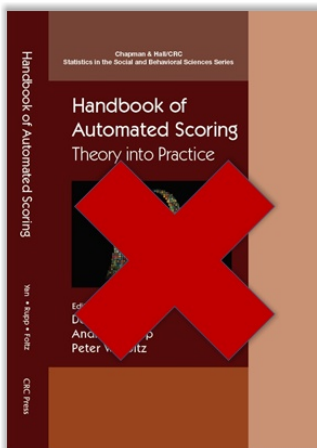
1.7 Expectations (I)



Let's discuss expectations....


1.8 Expectations (II)

ITEMS Modules in Context



1.9 Learning Objectives

Learning Objectives



1. Understand how your analytic work might intersect with automated scoring
2. Know the key dimensions of an automated scoring engine validation program
3. Understand how automated scores are produced
4. Identify the decision points around automated scoring that psychometricians make


1.10 Prerequisites

Prerequisites

Working knowledge of foundational concepts:

- Foundational statistical concepts
- Human scoring processes
- Open-ended test item types
- Reliability, validity, and fairness

No knowledge of particular statistical techniques is assumed or required!




1.11 Resources

Resources

Module Citation

Lottridge, S., Burkhardt, A., & Boyer, M. (2020). Automated scoring for psychometricians (Digital ITEMS Module 18). *Educational Measurement: Issues and Practice*, 39(3), XX-XX.

**Additional Resources**

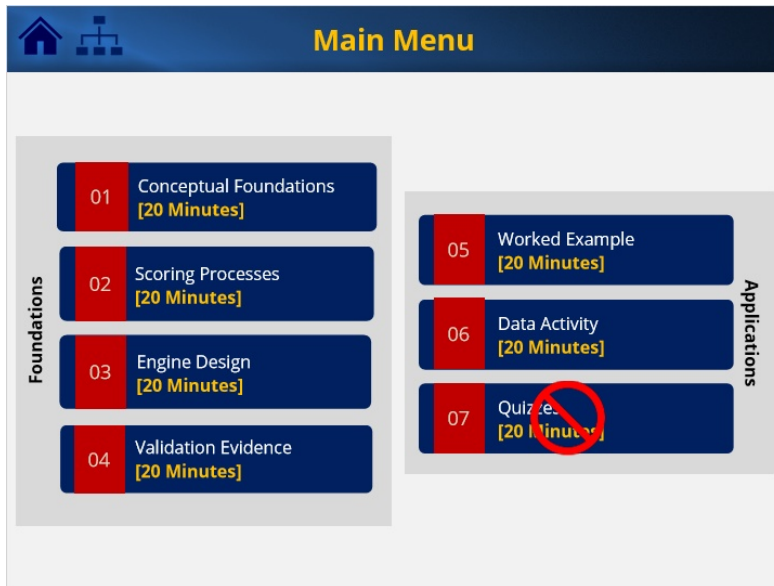
References (Slide Layer)

Resources



Back

1.12 Main Menu





2. Section 1: Conceptual Foundations


2.1 Cover: Section 1



2.2 Learning Objectives





Learning Objectives



1. Understand the goals and nature of automated scoring at a high level
2. Identify main benefits and drawbacks of automated scoring
3. Describe how automated scoring is used in assessment
4. Describe why psychometricians should care about automated scoring

2.3 Overview



Overview

Automated scoring refers to the use of computer algorithms to score unconstrained / open-ended test items, tasks, or activities. Automated scoring systems are typically configured to mimic human scoring.

Writing Prompts

Extended constructed response

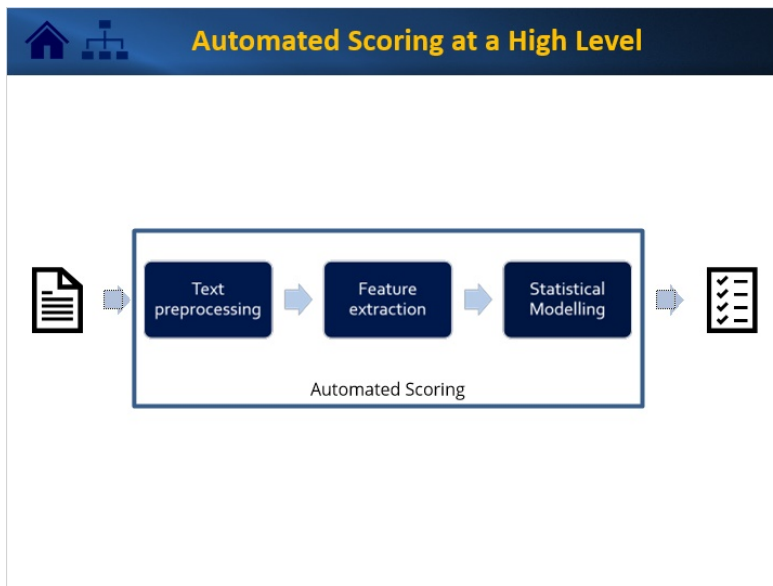
Short answer

Dialogue

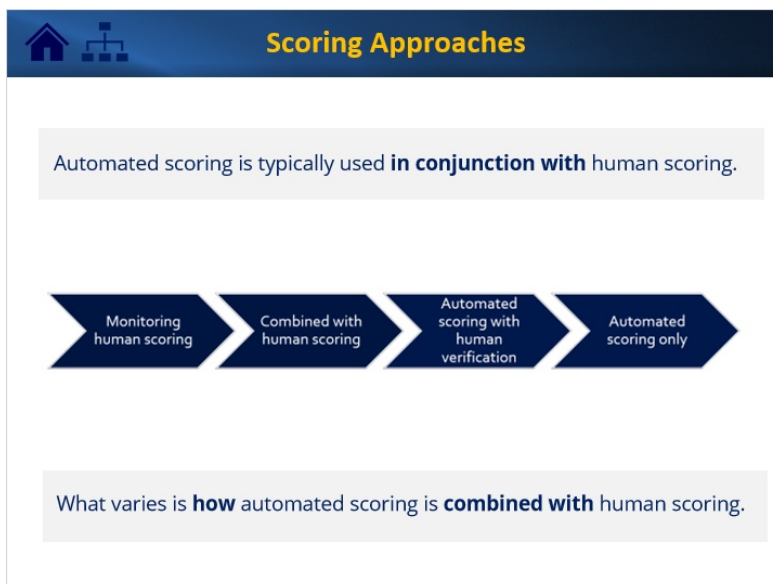
Speaking

Performance Tasks

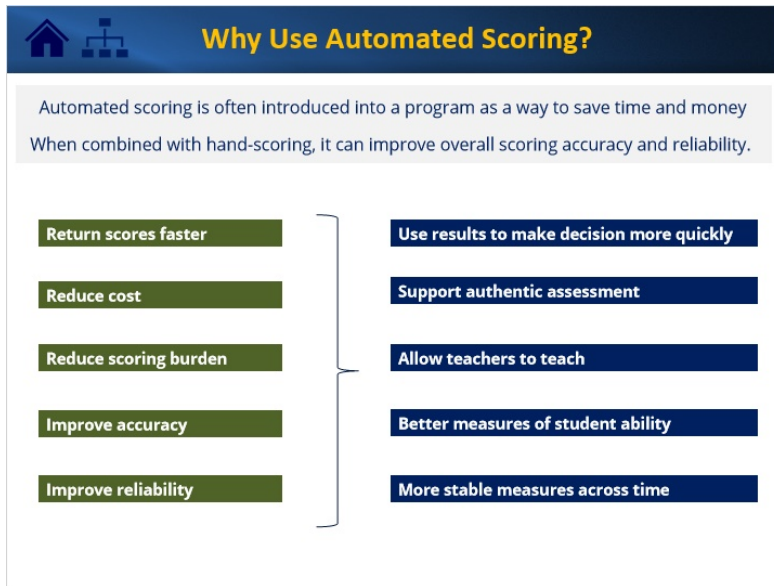
2.4 AS at a High Level



2.5 Scoring Approaches




2.6 Reasons for Use



2.7 Drawbacks



Drawbacks of Automated Scoring

The use of automated scoring does come with costs around complexity.



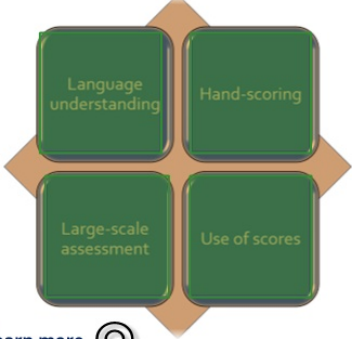
- **More complicated work flows** for approving performance and monitoring scoring
- **Time and resources** to train the engine, including high-quality hand-scored data
- **Changes in scoring** that may have impacts on psychometric scaling
- **Psychometric analyses expand** to include validation
- **Continuous maintenance and improvement** of software to manage emerging new items, responses, and languages


2.8 Topic Selection



Perceptions of Automated Scoring



Automated scoring can be a sensitive subject for people. People's views of language understanding, scoring, and large-scale assessment, and how scores are used influence their perceptions of automated scoring.



Click on each area to learn more 

Section End

2.9 Understanding





Language Understanding

Ellis Page (2003) listed three major objections to automated scoring. Underlying each of these objections is the idea that machines cannot understand language.

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*

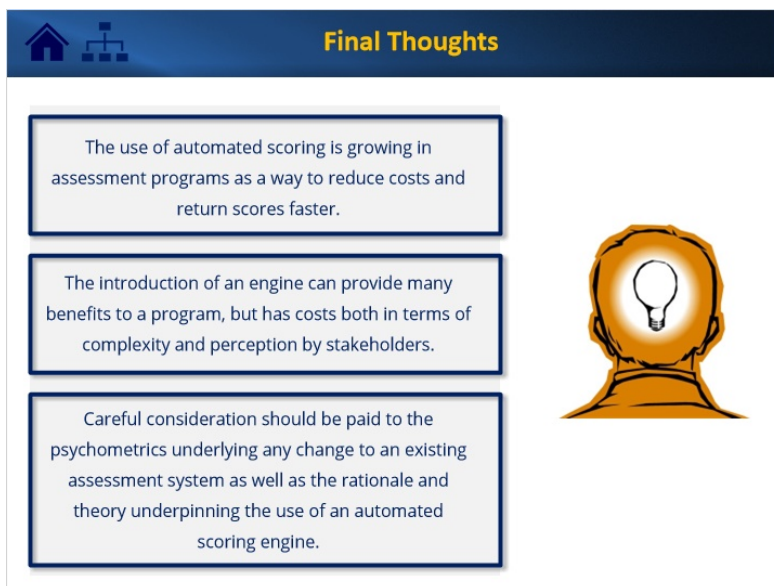
The analysis of automated scoring results is (or will be) part of the psychometric workflow



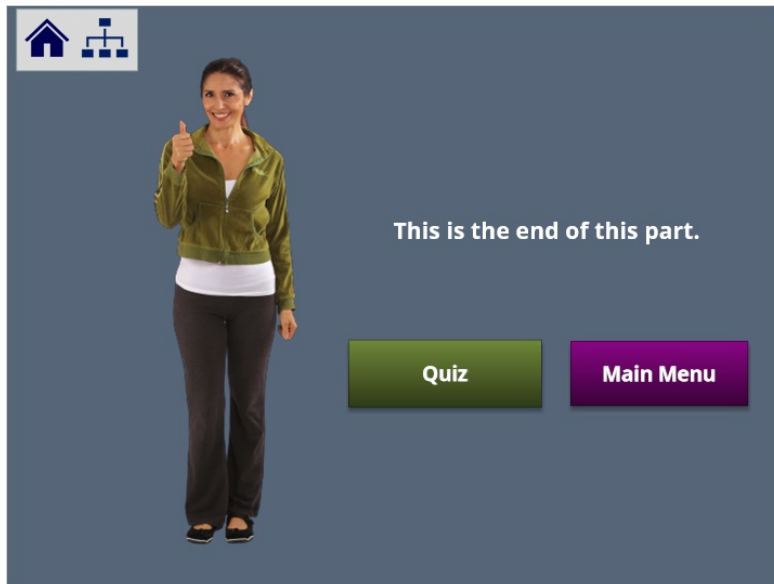
2.14 What Psychometricians Should Know



2.15 Final Thoughts



2.16 Bookend: Section 1





3. Section 2: Scoring Processes


3.1 Cover: Section 2



3.2 Learning Objectives





Learning Objectives



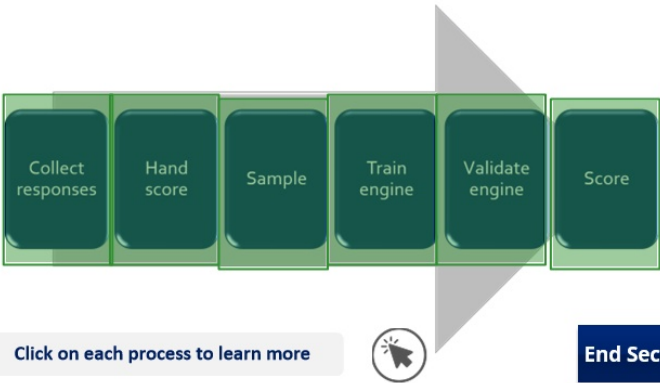
1. Understand what high-level steps are involved in training an automated scoring model
2. Understand how sampling procedures may differ from standard psychometric analyses
3. Understand the steps involved in item training and validation
4. Understand how to monitor automated scoring systems

3.3 Topic Selection



Overview



Like all statistical processes, automated scoring relies on data to estimate parameters for modeling human scoring



Click on each process to learn more

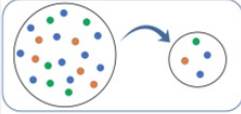
End Section

3.4 Response Collection



Response Collection

Data that the automated scoring engine is trained on should be representative of the intended population



- Responses are typically sampled in conjunction with other psychometric analysis and as part of field-testing, but can be sampled separately as well
- Approximately 2,000 responses for each item are needed for automated scoring models
- The representation of key subgroups should be considered to be able to examine fairness; subgroups often include:
 - ✓ sex
 - ✓ ethnicity
 - ✓ disability status
 - ✓ English language learner status

3.5 Hand-scoring





Hand-scoring

Automated scoring models human scoring, and so human scores should be of the highest quality using the procedures below



- Range-finding, or the process of using committees of experts to determine how to interpret the rubric, is a critical step in ensuring high-quality scoring
- The processes around rater recruitment, hiring, training, and qualification rely on well-defined criteria and training materials should reflect this
- The scoring design should include at least two independent reads, preferably with adjudication of discrepant scores
- Rater scoring should be monitored continuously using back-reads, validity papers, and evaluation of statistical measures

3.6 Sampling





Sampling

Additional sampling is required in automated scoring to protect against over-fitting and to ensure that we can predict how the engine will perform during live testing

Labeled Data	Unlabeled Data
TRAIN	SCORE CLASSIFICATION
VALIDATE	
TEST	

- It is critical to preserve a sample – called a test set – untouched by modeling and selection of models (typically 15-25% of the entire sample)
- Stratified sampling should be used to ensure score point representation because high or low scores are rare for most items
- The type of sub-sampling will be driven by the model-building process
 - ✓ Multiple models are built and evaluated using a validation set
 - ✓ Combining models into an ensemble requires more complex sub-sampling

3.7 Sub-sampling



Sub-Sampling Techniques

There are three main methods used to support the training and evaluation of competing automated scoring models.

Single-sample validation

- Divide training sample into train and validation, typically 85/15 split
- Examine performance on the validation sample
- Simplest method and least computationally intensive
- Results are less stable due to small validation sample size



K-fold Cross validation

- Ensures all records receive a validation score
- Compute metrics on all records
- Supports estimation of variance due to sampling
- Moderately computationally intensive

Bootstrapping

- Allows for estimation of standard error associated with the model statistics
- Validation scores can computed from the non boot-strapped samples
- Very computationally intensive

3.8 Cross-validation



k-fold Cross-validation

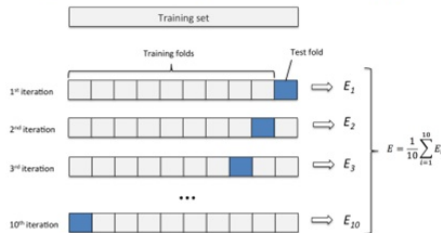
One of the most robust sub-sampling methods is *k*-fold cross validation

Divide sample into *k* even-sized partitions



- Combine *k*-1 partitions, model on partition, and predict *k*th partition
- Do this for each possible combination
- k* can be any number (often *k*=5 or *k*=10)

Every record in sample is part of some validation sample

- You now have predictions on every element in your training sample
- Results should reflect what we expect to see in the population




3.9 Bootstrapping





Bootstrapping

Bootstrapping can also be used if accurate estimation of variance is a concern when determining models



- Bootstrapping is repeated sampling with replacement
- 1/3 of sample is typically not drawn and used for validation
- The average of the bootstrap results is your accuracy estimate
- The standard deviation is the standard error estimate
- Bootstrapping can be computationally intensive

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





Validation

The purpose of validation is to obtain evidence that the chosen model performance will generalize to the intended population



- Validation methods should be used sparingly, preferably only once at the end of the training process
- The performance of the engine is evaluated relative to hand-scoring performance
- Criteria for adequate performance should be defined up-front and automated scoring performance relative to criteria should be provided in a technical report
- Data required for criteria evaluations include a first human score, a second human score, and an engine score and any subgroup data

3.13 Example





Example Evaluation Analysis

Below is a typical output from a report of an automated scoring engine on a writing item with three traits.


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





Fairness

Fairness evaluations examine the measures within subgroup (e.g., females) and may consider all measures or subsets of measures



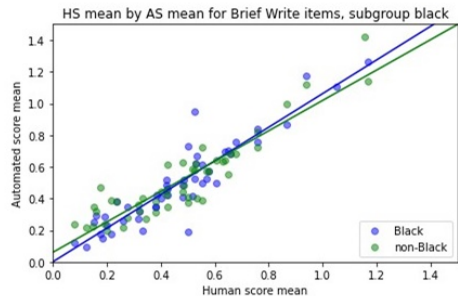
- Fairness is not well-studied in automated scoring; typically, we look for differences in engine performance within subgroup relative to hand-scoring
- Subgroup sizes may be too small to examine fairness for any given item but aggregation of results across items may provide insight into global issues
- Samples should be matched on ability to ensure that the evaluation is examining only the performance of automated scoring (versus difficulty)
- Fairness can be measured by examining the degree of differential item functioning (DIF) across machine-scored and hand-scored items

3.15 Fairness Example





Example Fairness Evaluation

We can examine the scatterplot of automated scoring mean scores and human mean scores by subgroup for samples matched on ability. We should see the lines of best fit overlapping for the two groups AND overlapping with the identity line.




3.16 Mitigation of Issues



Mitigation of Issues



If engines performance is problematic, various mitigation procedures may need to be put in place



Procedures include:


- ✓ re-calibration of statistical models
- ✓ removing the engine from scoring processes
- ✓ routing more responses for human scoring
- ✓ future engine improvement

3.17 Scoring





Scoring

Scoring using automated scoring involves decisions around how to include hand-scoring, how to monitor scores, and mitigation strategies in the event of scoring errors



- For most use cases, hand-scoring should be used during the scoring process to score unusual or hard-to-score responses or to validate engine performance
- Monitoring of scores can include the statistics outlined in the validation section and criteria should be set prior to scoring to be able to detect issues early
 - ✓ Monitoring should be conducted frequently
 - ✓ If issues are identified, the source of error could be the hand-scoring or machine scoring; both require review

3.18 Final Thoughts




Final Thoughts



There are multiple steps in the automated scoring process.


An sound methodological focus at each step will produce benefits for each later step.

An approach that ensures alignment between each step and to the purpose will support a solid validity argument for the use of automated scoring in a program.



3.19 Bookend: Section 2





This is the end of this section.

[Quiz](#)[Main Menu](#)

4. Section 3: Engine Design

4.1 Cover: Section 3

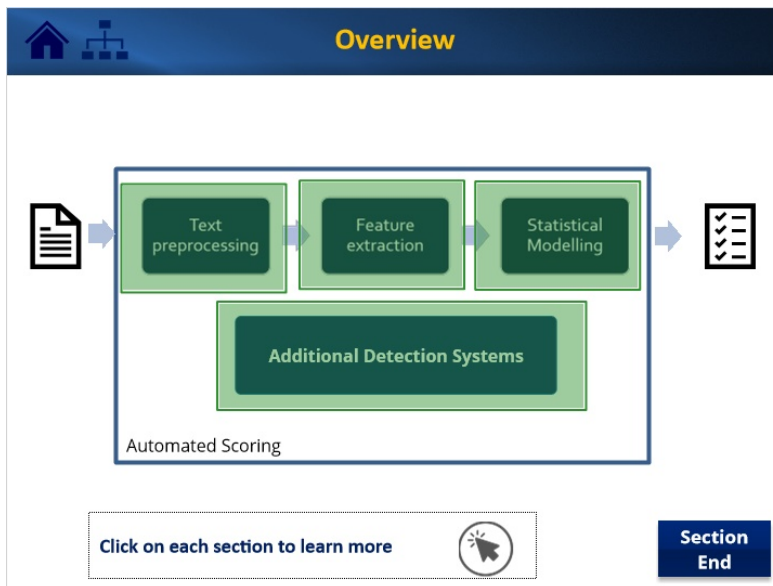


4.2 Learning Objectives

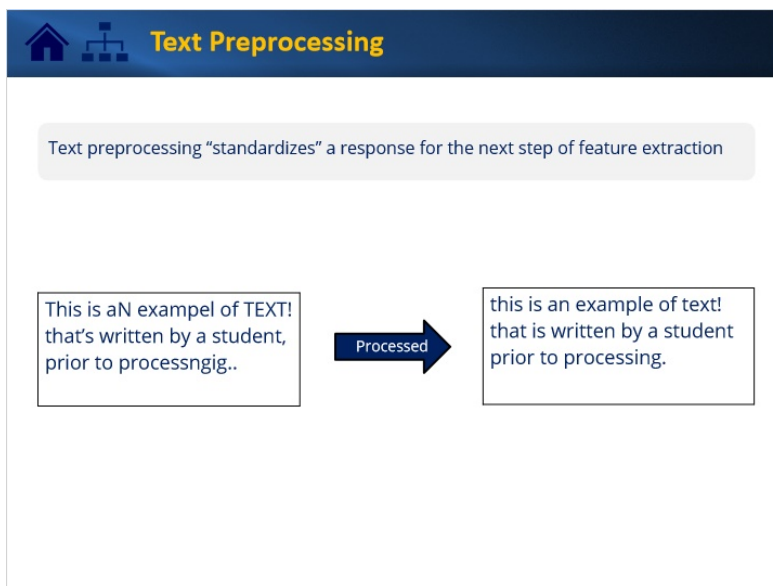
The slide is titled 'Learning Objectives' in yellow text on a dark blue header. Below the header, there is a 3D illustration of a target with a dart hitting the bullseye. At the bottom, there is a list of three learning objectives, each in a separate box with a blue border.

- 1. Understand the high-level processes for how an essay receives a score from an automated scoring system
- 2. Develop intuition around the model building process
- 3. Learn about the different checks that an essay undergoes in addition to receiving a score



4.3 Topic Selection



4.4 Preprocessing





4.5 Feature Extraction (I)



Feature Extraction (I)



Feature extraction takes a text preprocessed response and converts it to a vector or array of numeric values



Feature 1	Feature 2	Feature 3	Feature 4
5	1	1	.8

- Features are predictor variables
- Features can be *explicit* or *implicit*

4.6 Feature Extraction (II)





Feature Extraction (II)

Features are intended to represent the criteria of a rubric at different score levels

Score	Criteria
4	<ul style="list-style-type: none">- Multi-paragraphs- Introduction- Conclusion- Essay is easy to follow
3	<ul style="list-style-type: none">- Multi-paragraphs- Introduction and conclusion- Essay is typically easy to follow, but there are some parts that are unclear
2	<ul style="list-style-type: none">- More than two paragraphs- Missing Introduction or conclusion- Difficult to follow the essay
1	<ul style="list-style-type: none">- Single Paragraph- No introduction or conclusion- Topic of essay is not clear

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




Additional Response Detection Systems


The automated scoring process is trained to predict the score of typical, on-topic responses. Additional classification systems are used to identify responses that fall outside of 'typical.'



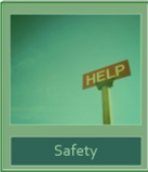
Non-attempts




Aberrant Text



Cheating



Safety

Click on each image to learn more 

[Back](#)

Nonattempts (Slide Layer)

Final Checks: Coding Non-attempts

Codes are assigned to explain why a student received a score of 0 or no score at all

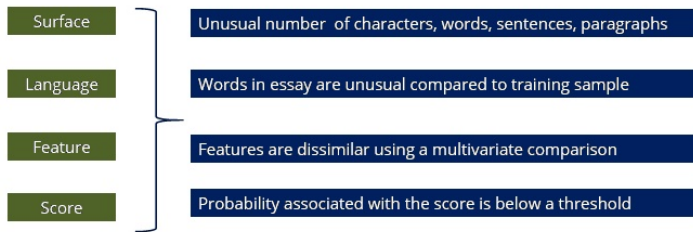
	}	Blank
<i>Djdklddk lkjdik ldds</i>		Gibberish
<i>No, IDK</i>		Refusal
<i>Voy a escribir in espanol.</i>		Non-English

[Back](#)

Aberrant Text (Slide Layer)

Final Checks: Aberrant Text

Essays that are markedly different from the training data are flagged and routed for hand-scoring review

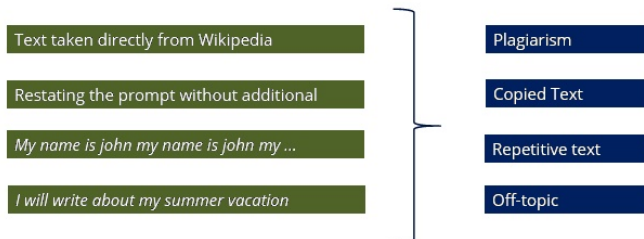


Back

Cheating (Slide Layer)

Final Checks: Cheating

Mitigate criticisms about scoring mis-scoring both unusual essays and essays that are intentionally trying to game the system



Back

Safety (Slide Layer)

Final Checks: Safety



Identify if the student is using the test to disclose a harmful situation

- Students will sometimes report on a test that they are in a harmful situation such as they are being abused or are threatening suicide
- Historically, human graders have been responsible for flagging this type of student writing but the burden is being transitioned to automatic detection systems
- Thousands of pieces of student writing are flagged annually (but are still rare)



Back

4.10 Non Attempts

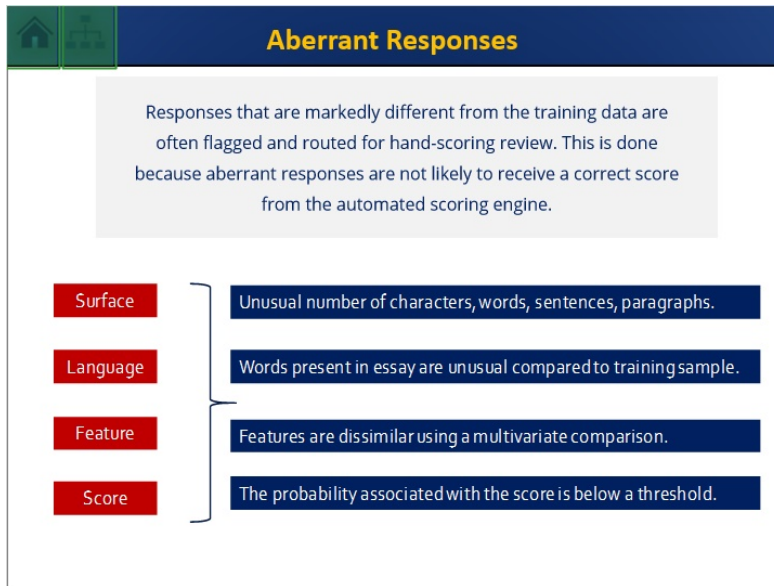


Coding Non-attempts

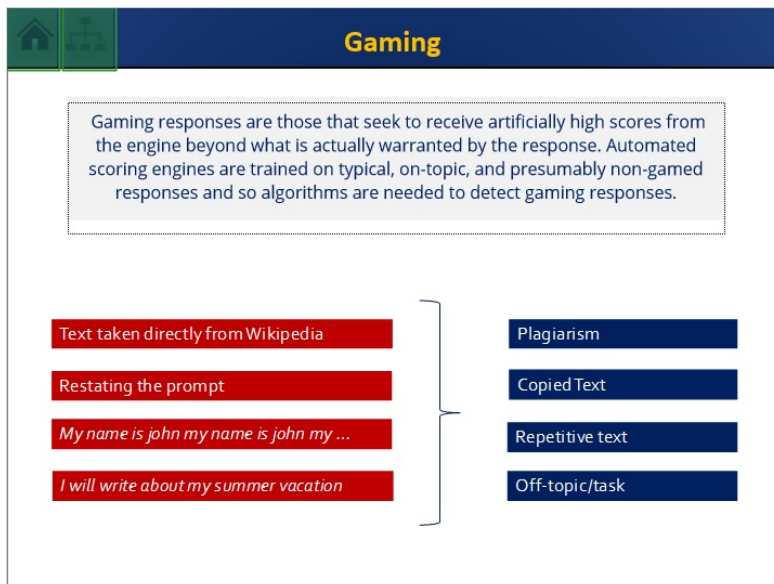
Non-attempts are used to identify responses that do not meet the minimal criteria for a score in the rubric. Typically, codes are used to label the nature of the non-attempt.

	Blank
<i>Djdkdidk l;kjdic llds</i>	Gibberish
<i>No, IDK</i>	Refusal
<i>Voy a escribir en espanol.</i>	Non-English



4.11 Aberrant Responses



4.12 Gaming Responses




4.13 Safety





Safety

Identify whether the student is using the test to disclose a harmful situation.

- Students will sometimes report on a test that they are in a harmful situation, such as they are being abused or are threatening suicide.
- Historically, human raters have been responsible for flagging this type of student writing, but the burden is being transitioned to automatic detection systems.
- Thousands of pieces of student writing are flagged annually (but are still rare).



4.14 Final thoughts




Final Thoughts

We stepped through a high-level conceptualization of how an automated scoring engine works.

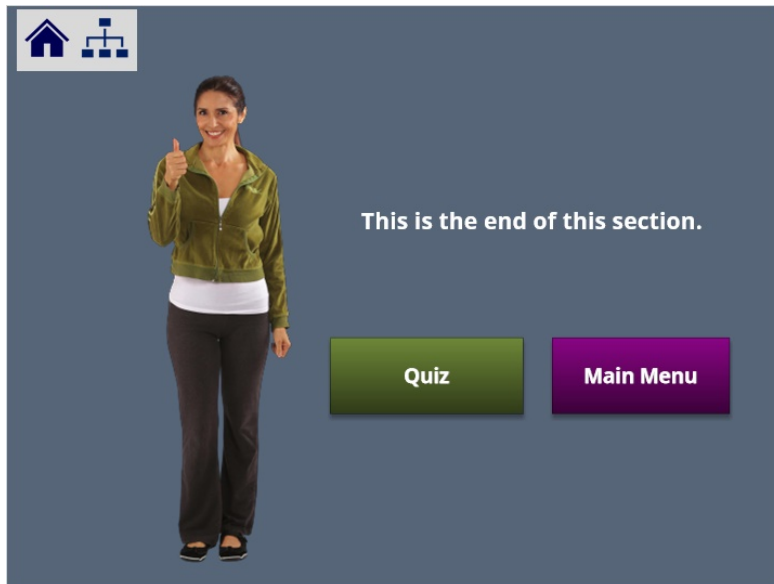
We covered additional text classification systems so that essays that should be receiving scores do, and responses that may be questionable are identified and potentially routed for humans for review.

These systems are in place to ensure valid scores.

The data activity will help to further illustrate the methods of preprocessing, feature extraction, scoring, training and validation.



4.15 Bookend: Section 3





5. Section 4: Validation


5.1 Cover: Section 4



5.2 Learning Objectives





Learning Objectives




1. How to apply validity demands to automated scoring
2. What validity concepts and practices are relevant to automated scoring evaluations
3. What questions and tools might be applied to evaluations
4. How automated scoring fits into the larger context of validity








5.3 Evolution





Evolution of Automated Rater Validation

Increasing Comprehensiveness in Approaches to Validation



-  Human-Machine score agreement used as validity evidence
-  Incorporated test and item design considerations
-  Emphasized the need for quality assurance of the human scores used to train engines
-  Established statistical criteria for automated score quality
-  Provided frameworks for validation to evaluate relationships between constructs, rubrics, responses, and rater behavior
-  Explored inferential methods to evaluate multiple raters (whether human and machine) over item collections
-  Developed systems level approaches that guide decision-making and allow for synthesizing evidence from multiple sources

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



Validity Demands in Automated Scoring

We can break down validity demands into four areas:
scorability, construct validity in engine design, and item/test comparability.
Fairness evaluations should be embedded in each area.

Scorability

- Does the item lend itself to high-quality scoring?
- Do the rubrics and scoring materials support high-quality scoring?
- Do the human rater training materials and methods support high-quality scoring?

Construct validity


- Does the engine design support the assessment of the construct, for all examinees?
- Do empirical analyses support that the construct is not altered?
- Are non-attempt and cheating or otherwise aberrant responses accurately identified?

Item-level comparability

- Does the engine demonstrate comparability to human raters in the aggregate?
- Does the engine demonstrate comparability to human raters for all examinee groups?
- Does the engine exhibit similar relationships to other measures as human rater scores?

Test-level comparability

- Does the inclusion of automated scoring change the internal structure of the test?
- Does the inclusion of automated scoring change the estimated item and person parameters?
- Does the inclusion of automated scoring change how results are reported?

[Click on each block above to learn more](#) 

[End Section](#)

5.7 Scorability



Scorability

The design and hand-scoring of items should support the creation of high-quality scores on which to model



Item Design

Structure

Clarity

Constraints

Response Expectations

Number and relationship of key concepts

Level of inference required

Expected language variation

Examinee

Alignment with instruction

Facility with tools

Experience with item or item type

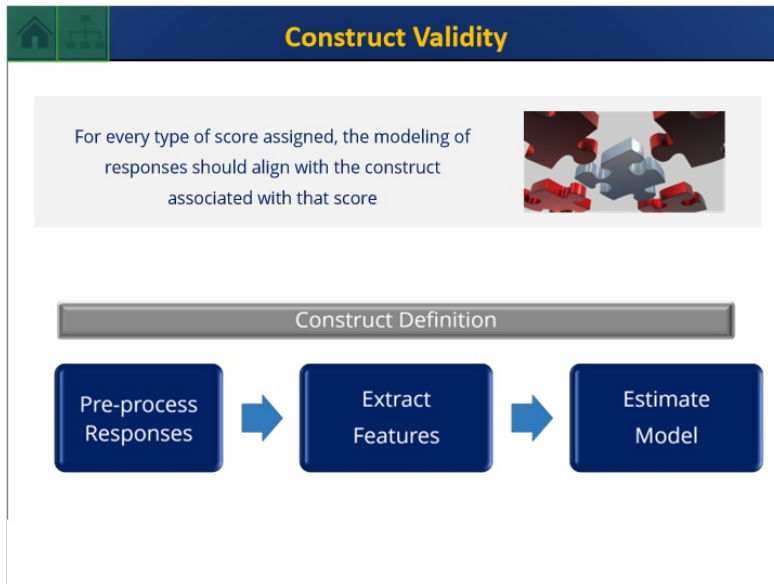
Scoring

Alignment of rubric with item

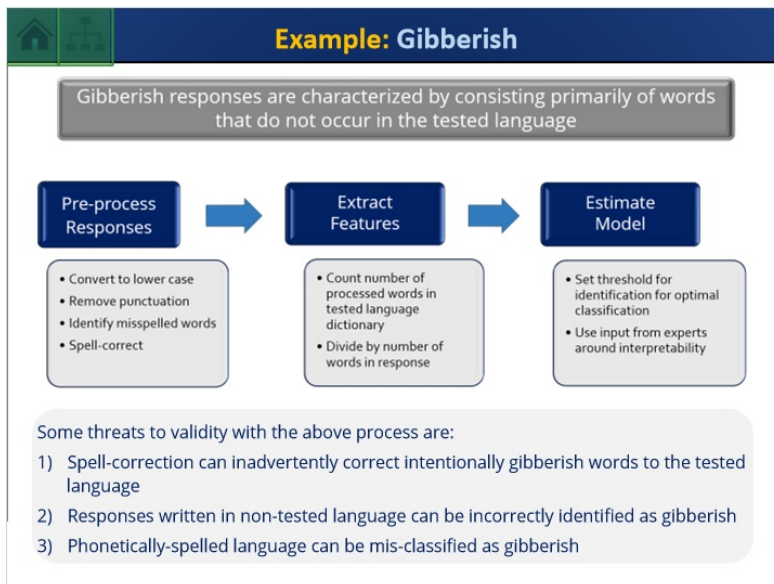
Clarity of rubric

Training and qualification

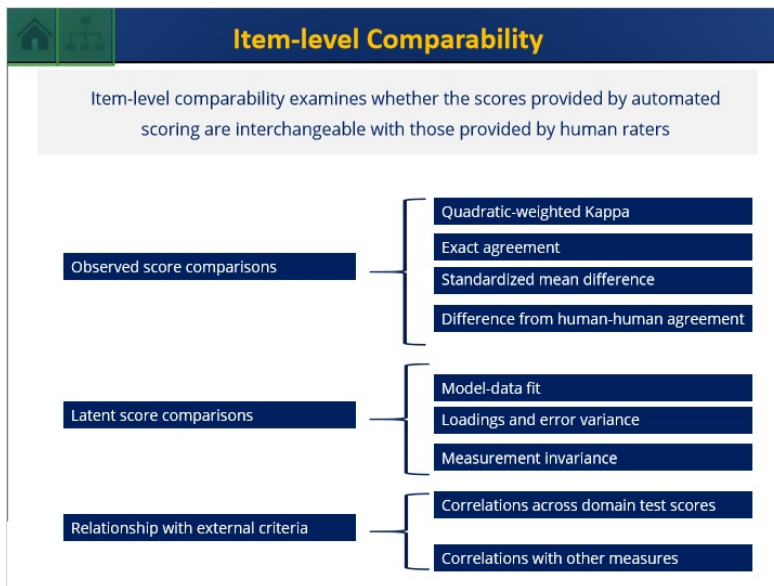
5.8 Construct Validity



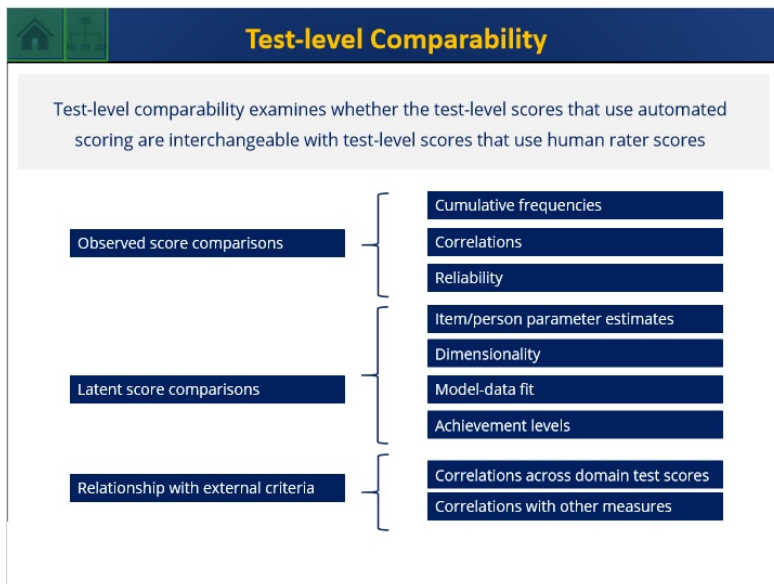
5.9 Example



5.10 Item-level Comparability






5.11 Test-level Comparability



5.12 Standards

 	Standards for High Quality Validation
1.	The rationale for using automated scoring should be clearly articulated and appropriate for the program in which it is used.
2.	The architectural design of the automated scoring system should be grounded in a theoretical approach that aligns with the constructs assessed via the items, rubrics, and other scoring materials.
3.	The automated scoring system should be trained on a representative sample of responses that were hand-scored with a level of quality aligned with program needs.
4.	The validity, reliability, and fairness of automated scoring should be evaluated using a sound methodological and statistical approach and clear evaluation criteria .
5.	The approach for using automated scoring and/or human scoring during test administrations should be based upon scoring performance and aligned to the needs of the program .
6.	A well-defined process for reviewing scoring performance during and after test administrations should exist and there should be a process in place for handling errors or disruptions.
7.	Examinee data should be treated securely and in accordance with the laws and principles that regulate the assessment program.

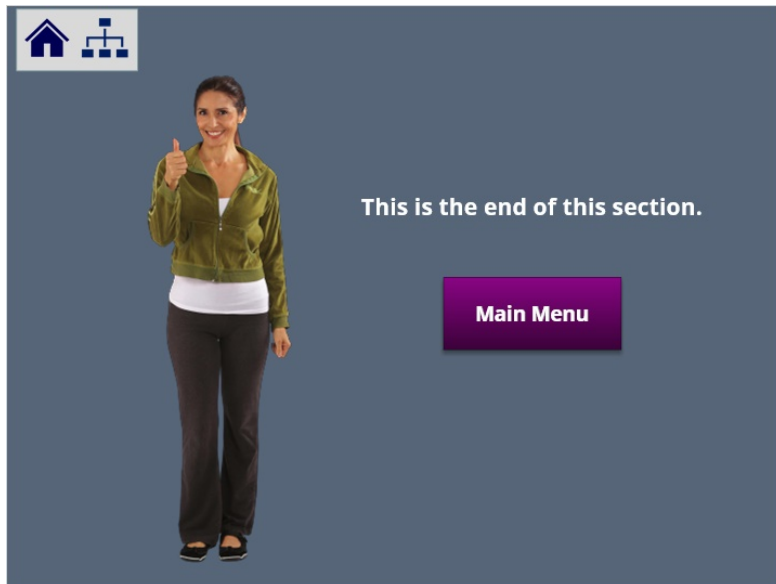
5.13 Closing Thoughts

 	A Few Closing Thoughts
<p>Although validation approaches in automated scoring have advanced in important and substantive ways since 1966, that work is not done.</p>	
<p>Processes and approaches to validation for automated scoring are continually advancing toward the goals of improving the accuracy of automated scores, their evaluation criteria, and of course, to address continued public skepticism of a machine's abilities relative to trusted human raters.</p>	
<p>There are many opportunities for your participation in the continued research and development of automated scoring, and how evidence of the validity of the scores that they produce is established.</p>	

5.14 References

References
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). <i>Standards for educational and psychological testing</i> . Washington, DC: American Educational Research Association.
Page, E. B. (1966). Grading essays by computer: Progress report. In <i>Proceedings of the 1966 invitational conference on testing</i> . Princeton, NJ (pp. 87-100). Princeton, NJ: Educational Testing Service.
Yang, Y., Buckendahl, C. W., J., Piotr J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. <i>Applied Measurement in Education</i> , 15(4), 391-412.
Bennett, R. E. & Behar, I. I. (1998). Validity and automated scoring: It's not only in the scoring. <i>Educational Measurement: Issues and Practice</i> , 4, 9-17.
Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for evaluation and use of automated scoring. <i>Educational Measurement: Issues and Practice</i> , 31, 2-13.
Bennett, R.E., & Zhang, M. (2016). Validity and automate scoring. In F. Drasgow (Ed.), <i>Technology and testing: Improving educational and psychological measurement</i> . New York: NY.
Kieftenbeld, V. & Boyer, M. (2017). Statistically Comparing the Performance of Multiple Automated Raters across Multiple Items. <i>Applied Measurement in Education</i> , 30(2) pp. 117-128.
Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions, <i>Applied Measurement in Education</i> , 31:3, 191-214.
Yan, D. & Bridgeman, B. (2020). Validation of Automated Scoring Systems. In D. Yan, A. Rupp, & P. Foltz (Eds), <i>Handbook of automated scoring: Theory into practice</i> . Boca Raton, FL: Taylor & Francis Group.

5.15 Bookend: Section 4





6. Section 5: Worked Example


6.1 Worked Example



6.2 Learning Objectives





Learning Objectives



1. Understand the potential impact on total test score comparability over a range of automated rater quality
2. Understand how different quality human scores may impact total test score accuracy

6.3 Objective 1



An Evaluation of Test Score Comparability

Objective 1:

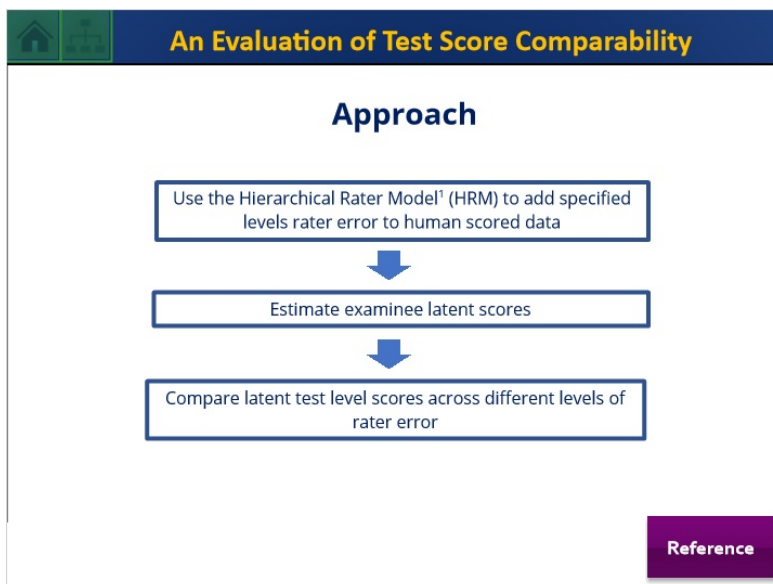
Understand how different quality human scores impact test level accuracy

This can inform decisions about the adequacy of human scores for training and validating an automated rater.

Scenario:

- ~3300 examinees take a 20 item test where 50% of their raw score points are based on human scores
- We want to know the consequence of using scores from human raters that perform at different thresholds for acceptable agreement (0.70, 0.80, 0.90 QWK).

6.4 Approach



Reference (Slide Layer)



Reference

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 Brian W. Junker
CTB/McGraw-Hill 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 unobserved discrete variables, 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 Marascuilo (1995), and the rater handle 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, 1980; Engelhard, 1994, 1996), and we explore ways that information from HRM analyses may improve 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

Back

6.5 HRM



The HRM Model

HRM is a two-stage, three-level hierarchy with observed ratings (X_{ipr}), ideal ratings (ξ_{pi}), and examinee true scores (θ):

$$\begin{cases} \theta_p \sim i.i.d. N(\mu, \sigma^2), & p = 1, \dots, P \\ \xi_{pi} \sim \text{an IRT model}, & i = 1, \dots, I, \text{ for each } p \\ X_{ipr} \sim \text{signal detection model}, & r = 1, \dots, R, \text{ for each } p, i \end{cases}$$

The 2 stages are:

- Stage 1: **signal detection model** that produces an “ideal” rating
- Stage 2: **measurement model** use for estimating latent trait test scores (thetas) using ideal ratings for each examinee response

The 3 levels are:

- Level 1: models the **distribution of ratings** given the quality of the response
- Level 2: models the **distribution of an examinee’s response** given their ability
- Level 3: models the **distribution of the latent trait** theta

6.6 Signal Detection

Using HRM to Simulate Error

The rating process is modeled as a discrete signal detection-like problem relating observed and ideal ratings.

In other words, how likely is rater r to give rating k , given ideal rating ζ ?

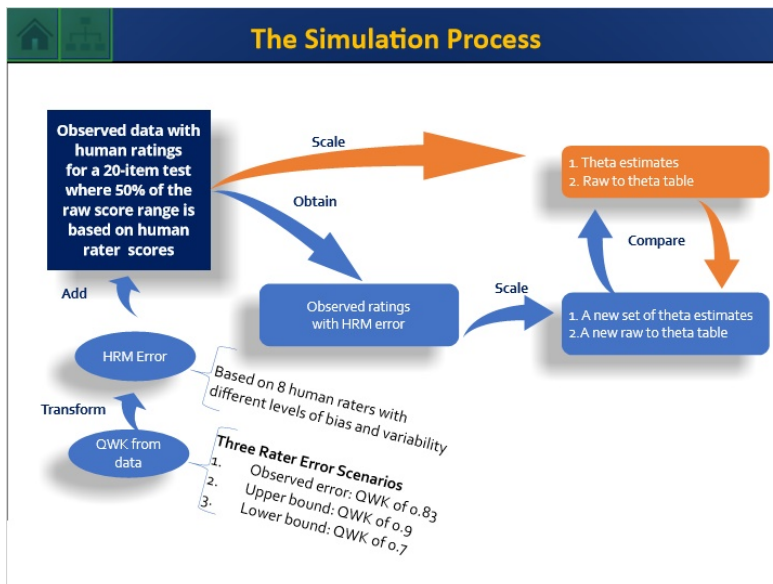
Ideal Rating (ζ)	Observed Rating (k)				
	0	1	2	3	4
0	P_{00r}	P_{01r}	P_{02r}	P_{03r}	P_{04r}
1	P_{10r}	P_{11r}	P_{12r}	P_{13r}	P_{14r}
2	P_{20r}	P_{21r}	P_{22r}	P_{23r}	P_{24r}
3	P_{30r}	P_{31r}	P_{32r}	P_{33r}	P_{34r}
4	P_{40r}	P_{41r}	P_{42r}	P_{43r}	P_{44r}

*Reproduced from Pat: et al. (2002)
Note: $P_{\zeta k} = P(\text{Rater } r \text{ rates } k | \text{ideal rating } \zeta \text{ in each row of this matrix})$

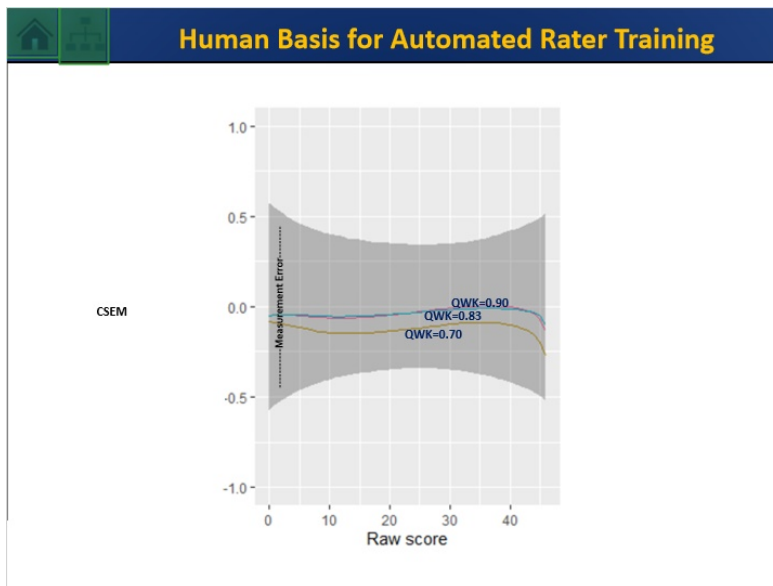
Each row in the table is a unimodal (unfolding) discrete distribution where the mode is the rater bias and the variability represents rater unreliability.

We use this distribution to simulate rater error in vectors of item scores based on parameter values that correspond to the QWK values.

6.7 Simulation Process



6.8 Objective1 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



Setting Quality Targets for Engine Training

Objective 2:

Understand potential impact on test level score comparability over a range of QWK



Approach:

- Using the observed data:
 - ✓ Introduce HRM error corresponding to QWK of 0.70 for a single rater
 - ✓ Separately, introduce HRM error corresponding to QWK of 0.90 for a single rater
- Estimate latent trait scores from these scenarios
- Compare examinee scores to understand the potential impact over the simulated range of engine performance

6.11 Objective 2 Results





6.12 Objective 2 Summary





What Do These Results Tell Us?

- Depending on the how the scores for this test are to be used, we may find that an automated rater performing at this lower bound threshold is not acceptable.



- Alternatively, we might decide that the quality is acceptable and proceed with setting a target as intended.

6.13 Overall summary



Utility of Evaluating Quality Thresholds

- The utility of this type of analysis lies in how it allows us to understand the **potential downstream effects of rater error on an examinee's total test score**, under the conditions of an observed data set.
- If developers are able anticipate the **potential effects of rater bias and variability** in the context of a specific test design, and using observed data, they can make possibly **increasingly informed decisions about the quality criteria** that are used for determining the acceptability of an automated rater.

6.14 Bookend: Section 2





7. Section 6: Data Activity

7.1 Cover: Section 6




7.2 Learning Objectives





Learn how to use the app

Learning Objectives



1. Get background information on the essays that are loaded into the app
2. Use a Shiny app to develop intuition around building an automated scoring model
3. Explore basic examples of features and statistical methods used in automated scoring


7.3 ASAP items



Background: Data Source

These essays are from a Hewlett Foundation data science competition



[Automated Student Assessment Prize \(ASAP\)](#)



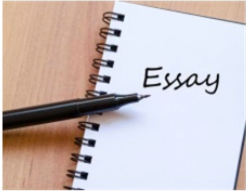
Two sets of essays are loaded into the app:

- **Prompt 1:** Write a letter to your local newspaper in which you state your opinion on the effects that computers have on people. Persuade the readers to agree with you.
- **Prompt 2:** Write a persuasive essay to a newspaper reflecting your views on censorship in libraries

7.4 ASAP essays




Background: Data Source





Atypical characteristics of these essays:

- Anonymized (replaced proper nouns with generic tags; Denver <- @LOCATION)
- Single paragraphs
- Rubric combines multiple traits into a single score
- No resolved score
- Transcribed from handwritten responses



How might these characteristics impact the scoring process (for either a hand-scorer or a model)?

7.5 Prompt 1 Rubric



Rubric: Computers (Prompt 1)

Score Point 1: An undeveloped response that may take a position but offers no more than very minimal support. Typical elements:

- Contains few or vague details.
- Is awkward and fragmented.
- May be difficult to read and understand.
- May show no awareness of audience.

Score Point 2: An under-developed response that may or may not take a position. Typical elements:

- Contains only general reasons with unelaborated and/or list-like details.
- Shows little or no evidence of organization.
- May be awkward and confused or simplistic.
- May show little awareness of audience.

Score Point 3: A minimally-developed response that may take a position, but with inadequate support and details. Typical elements:

- Has reasons with minimal elaboration and more general than specific details.
- Shows some organization.
- May be awkward in parts with few transitions.
- Shows some awareness of audience.

Score Point 4: A somewhat-developed response that takes a position and provides adequate support. Typical elements:

- Has adequately elaborated reasons with a mix of general and specific details.
- Shows satisfactory organization.
- May be somewhat fluent with some transitional language.
- Shows adequate awareness of audience.

Score Point 5: A developed response that takes a clear position and provides reasonably persuasive support. Typical elements:

- Has moderately well elaborated reasons with mostly specific details.
- Exhibits generally strong organization.
- May be moderately fluent with transitional language throughout.
- May show a consistent awareness of audience.

Score Point 6: A well-developed response that takes a clear and thoughtful position and provides persuasive support. Typical elements:


- Has fully elaborated reasons with specific details.
- Exhibits strong organization.
- Is fluent and uses sophisticated transitional language.
- May show a heightened awareness of audience.

7.6 Prompt 2 Rubric

Rubric: Censorship (Prompt 2)	
<p>Rubric Guidelines—Domain 1: Writing Applications</p> <p>Score Point 6: A Score Point 6 paper is rare. It fully accomplishes the task in a thorough and insightful manner and has a distinctive quality that sets it apart as an outstanding performance.</p> <p>Idea and Content Does the writing sample fully accomplish the task (e.g., support an opinion, summarize, tell a story, or write an article)? Does it</p> <ul style="list-style-type: none"> present a unifying theme or main idea without going off on tangents? stay completely focused on topic and task? <p>Does the writing sample include thorough, relevant, and complete ideas? Does it</p> <ul style="list-style-type: none"> include in-depth information and exceptional supporting details that are fully developed? fully explore many facets of the topic? <p>Organization Are the ideas in the writing sample organized logically? Does the writing</p> <ul style="list-style-type: none"> present a meaningful, cohesive whole with a beginning, a middle, and an end (i.e., include an inviting introduction and a strong conclusion)? progress in an order that enhances meaning? include smooth transitions between ideas, sentences, and paragraphs to enhance meaning of text (i.e., have a clear connection of ideas and use topic sentences)? <p>Style Does the writing sample exhibit exceptional word usage? Does it</p> <ul style="list-style-type: none"> include vocabulary to make explanations detailed and precise, descriptions rich, and actions clear and vivid (e.g., varied word choices, action words, appropriate modifiers, sensory details)? demonstrate control of a challenging vocabulary? <p>Does the writing sample demonstrate exceptional writing technique?</p> <ul style="list-style-type: none"> is the writing exceptionally fluent? does it include varied sentence patterns, including complex sentences? does it demonstrate use of writer's techniques (e.g., literary conventions such as imagery and dialogue and/or literary genres such as humor and suspense)? <p>Tone Does the writing sample demonstrate effective adjustment of language and tone to task and reader? Does it</p> <ul style="list-style-type: none"> exhibit appropriate register (e.g., formal, personal, or dialect) to suit task? demonstrate a strong sense of audience? exhibit an original perspective (e.g., authoritative, lively, and/or exciting)? 	<p>Score Point 3: A Score Point 3 paper represents a performance that minimally accomplishes the task. Some elements of development, organization, and writing style are weak.</p> <p>Idea and Content Does the writing sample minimally accomplish the task (e.g., support an opinion, summarize, tell a story, or write an article)? Does it</p> <ul style="list-style-type: none"> attempt a unifying theme or main idea? stay somewhat focused on topic and task? <p>Does the writing sample include some relevant ideas? Does it</p> <ul style="list-style-type: none"> include some information with only a few details, or list ideas without supporting details? explore some facets of the topic? <p>Organization Is there an attempt to logically organize ideas in the writing sample? Does the writing</p> <ul style="list-style-type: none"> have a beginning, a middle, or an end that may be weak or absent? demonstrate an attempt to progress in an order that enhances meaning? (Progression of text may sometimes be unclear or out of order.) demonstrate an attempt to include transitions? (Are some topic sentences used? Are transitions between sentences and paragraphs weak or absent?) <p>Style Does the writing sample exhibit ordinary word usage? Does it</p> <ul style="list-style-type: none"> contain basic vocabulary, with words that are predictable and common? demonstrate some control of vocabulary? <p>Does the writing sample demonstrate average writing technique?</p> <ul style="list-style-type: none"> is the writing generally fluent? does it contain mostly simple sentences (although there may be an attempt at more varied sentence patterns)? is it generally ordinary and predictable? <p>Tone Does the writing sample demonstrate an attempt to adjust language and tone to task and reader? Does it</p> <ul style="list-style-type: none"> demonstrate a difficulty in establishing a register (e.g., formal, personal, or dialect)? demonstrate little sense of audience? generally lack an original perspective?

This is an example of rubric – criteria for score of 6 and score of 3

7.7 Features

Features (Predictor Variables)	
<p>Simply counts intended to capture information pertaining to different elements of student writing:</p> <ul style="list-style-type: none"> Commas <ul style="list-style-type: none"> Intends to capture an element of the writing style Discourse Elements <ul style="list-style-type: none"> Count of words commonly used in argumentation Gets at the organization of the response Verbs (past tense) <ul style="list-style-type: none"> Another signal of writing style Introduction to the idea of using parts-of-speech as features in a model Grammatical Errors <ul style="list-style-type: none"> Focuses on readability of the text Unique Words <ul style="list-style-type: none"> Addresses lexical diversity 	
	<ul style="list-style-type: none"> Why might be counts be problematic features to score student essays? Semantic features would be preferable but they are more complex to develop and interpret and are not used in this example



7.8 Reminder



Recap: Important Reminder

1. We are using these essays because they are **publicly available**
2. We are using these features to provide a **simple demonstration** that can be easily explained
3. Ideally, we would have:
 - ✓ A rubric for which each trait captures a **unique aspect** of student writing
 - ✓ Features (that are not simply counts) that have been **specifically engineered** to assess each trait

7.9 Choose adventure




Choose Your Own Adventure

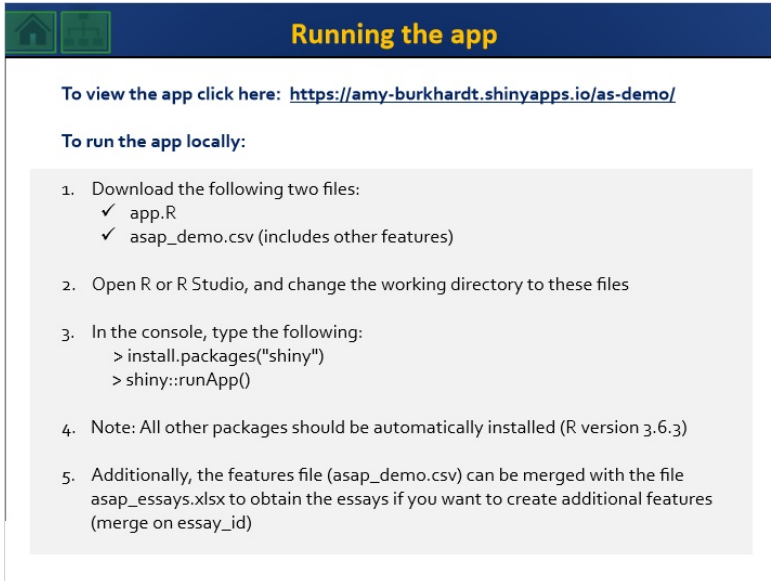
1. First become familiar with the app and then return to the slides

OR

2. First step through the remainder of the slides to get a better sense of the activity's purpose and functionality



7.10 Running the App



Running the app

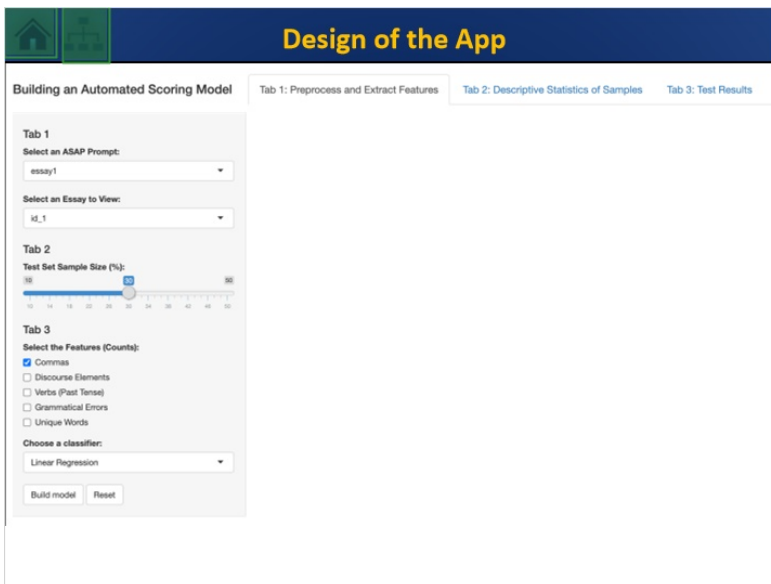
To view the app click here: <https://amy-burkhardt.shinyapps.io/as-demo/>

To run the app locally:

1. Download the following two files:
 - ✓ app.R
 - ✓ asap_demo.csv (includes other features)
2. Open R or R Studio, and change the working directory to these files
3. In the console, type the following:

```
> install.packages("shiny")
> shiny::runApp()
```
4. Note: All other packages should be automatically installed (R version 3.6.3)
5. Additionally, the features file (asap_demo.csv) can be merged with the file asap_essays.xlsx to obtain the essays if you want to create additional features (merge on essay_id)

7.11 App Design



Design of the App

Building an Automated Scoring Model

Tab 1: Preprocess and Extract Features Tab 2: Descriptive Statistics of Samples Tab 3: Test Results

Tab 1

Select an ASAP Prompt:
essay1

Select an Essay to View:
id_1

Tab 2

Test Set Sample Size (%):
15

Tab 3

Select the Features (Counts):
☒ Commas
☐ Discourse Elements
☐ Verbs (Past Tense)
☐ Grammatical Errors
☐ Unique Words

Choose a classifier:
Linear Regression

Build model Reset

7.12 Tab 1

Tab 1: Preprocess and Extract Features

Tab 1

Select an ASAP Prompt:

essay1

Select an Essay to View:

id_3

Tab 1: Preprocess and Extract Features

Tab 2: Descriptive Statistics of Samples

Tab 3: Test Results

Extracted Features

Rater 1 Score	essay_id	Commas	Discourse Elements	Verbs (Past Tense)	Grammatical Errors	Unique Words
4	126	1	16	1	17	150

Unprocessed Essay

Dear, BORGANIZATION! concerned with an issue that people are using computers and not exsisting. This is a very bad thing and we need to let the word out of what can happen if we don't exercise. The first reason why I disagree of the fact that computers help people is because computers really affect your health you will become fat you start having problems with your heart and it can lead to something fatal. Another reason why I think computers are healthy is because computers affect the economy alot in so many ways for example computers use up so much electricity it affects the power in all states. Computers also run jobs and this is very bad for the economy people will be addicted to computers for so long they wont bother looking for jobs they will just stay on the computer and gamble or buy stuff. The last reason why I think computers are bad is because of some of the illegal stuff people do on computers there are alot of cyber predators that prey on kids and they end up finding the kid and kidnapp them. There are also other illegal stuff like bad websites or popups that can get you arrested so if kids go on computers they will not no what something means and they will click on it and it can traumatize a kid their whole life. So NUM1 I hope you can take this letter into recognition and do something about it cause this computers can affect my health your health and even your childrens health. That is why I have written this letter to you. If you agree with me I thank you.

Processed Essay

Dear, 'ORGANIZATION!' concerned with an issue that people are using computers and not exsisting. This is a very bad thing and we need to let the word out of what can happen if we do n't exercise. The first reason why I disagree of the fact that computers help people is because computers really affect your health you will become fat you start having problems with your heart and it can lead to something fatal. Another reason why I think computers are healthy is because computers affect the economy alot in so many ways for example computers use up so much electricity it affects the power in all states. Computers also run jobs and this is very bad for the economy people will be addicted to computers for so long they wont bother looking for jobs they will just stay on the computer and gamble or buy stuff. The last reason why I think computers are bad is because of some of the illegal stuff people do on computers there are alot of cyber predators that prey on kids and they end up finding the kid and kidnapp them. There are also other illegal stuff like bad websites or popups that can get you arrested so if kids go on computers they will not no what something means and they will click on it and it can traumatize a kid their whole life. So NUM1 I hope you can take this letter into recognition and do something about it cause this computers can affect my health your health and even your childrens health. That is why I have written this letter to you. If you agree with me I thank you.

Sentence Tokenization

[[Dear, 'ORGANIZATION!' concerned with an issue that people are using computers and not exsisting. This is a very bad thing and we need to let the word out of what can happen if we do n't exercise. The first reason why I disagree of the fact that computers help people is because computers really affect your health you will become fat you start having problems with your heart and it can lead to something fatal. Another reason why I think computers are healthy is because computers affect the economy alot in so many ways for example computers use up so much electricity it affects the power in all states. Computers also run jobs and this is very bad for the economy people will be addicted to computers for so long they wont bother looking for jobs they will just stay on the computer and gamble or buy stuff. The last reason why I think computers are bad is because of some of the illegal stuff people do on computers there are alot of cyber predators that prey on kids and they end up finding the kid and kidnapp them. There are also other illegal stuff like bad websites or popups that can get you arrested so if kids go on computers they will not no what something means and they will click on it and it can traumatize a kid their whole life. So NUM1 I hope you can take this letter into recognition and do something about it cause this computers can affect my health your health and even your childrens health. That is why I have written this letter to you. If you agree with me I thank you.]]

7.13 Tab 1 Questions

Tab 1: Preprocess and Extract Features

Below are some guiding questions to help you think about this first tab.

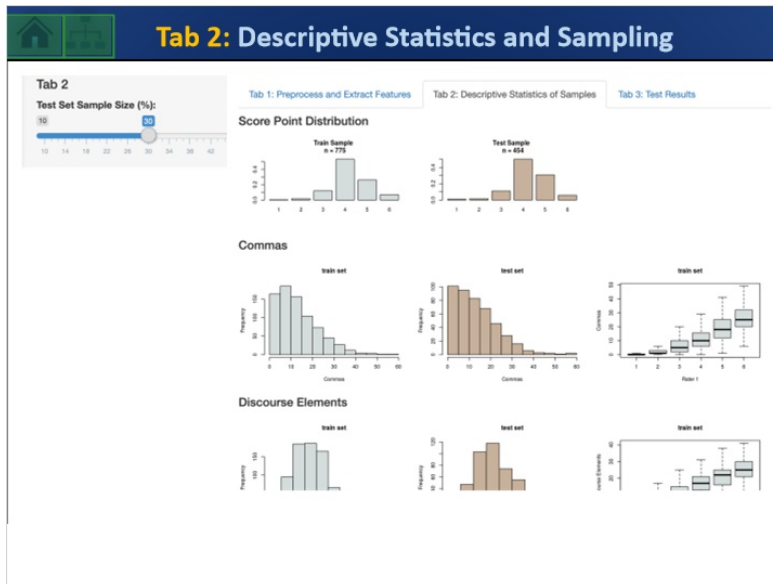
Essays

- Processed Essay:** In this example, we only engaged in a small amount of preprocessing. Why might we not want to do a lot of preprocessing? What other preprocessing could we have done?
- Sentence Tokenization:** Why might we want to segment the essay into sentences? Can you think of features that we could extract from the sentences?
- Parts-of-speech Tags:** Note how we count the number of words per each tag. Why might we want to represent the essay this way? Can you think of features that we can extract from these tags?

Feature Extraction

- Note the extracted features for the individual essay at the top of the screen. Can you think of other features that might be useful in predicting the Rater 1 score?
- Select the other prompt from the dropdown and review its essays. Does anything stand out about the content, features, or hand-scores of these essays?

7.14 Tab 2



7.15 Tab 2 Questions

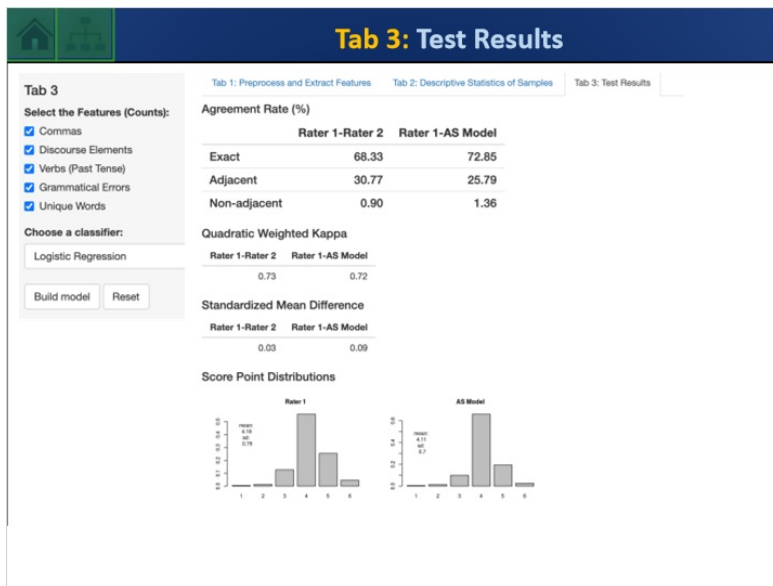
Tab 2: Descriptive Statistics and Sampling

Below are some guiding questions to help you think about the second tab.

- Score Point Distributions:** Compare the distribution of score points for the training and test set. Does the shape of distribution remain similar for both, regardless of the sample size? Given the low frequency of the 1, 2 and 6 score points, does taking a random sample seem appropriate, or should we use a stratified sample?
- Features:** Note the distribution of features for the training and test set. Does anything stand out about these distributions within each sample, and across the train and test sample?
- Box plots:** Note the relationships between the rater 1 score and the feature values. Do any of the relationships surprise you?

Test Sample Size. Select a test sample of 10% of the data and review the results in the tab. Then, select a test sample of 50%, and review the results. What stands out to you when you do this?

7.16 Tab 3



7.17 Tab 3 Questions



Tab 3: Test Results

Below are some guiding questions to help you think about the third tab.

- Select the features:** Which feature seems to be the most useful in predicting human scores? Why might this feature be so predictive, and why might this be problematic in operational use of an automated scoring system?
- Choose a Classifier:** Which classifier seems to produce the best model? How are you making this determination?
- Features and Classifier:** Which combination of features and classifier results in the best model?

If you think you found the optimal classifier and features for operational use, you should test the model one more time on a held-out test sample (which we don't have in the demo). Why is it important to do this?

7.18 Tab 3 Questions (2)




Tab 3: Test Results

Finally, we can examine chance variability in our results due to sampling.

Chance Variability (and Validating the Model)

Follow these steps:

1. Specify the test sample size (e.g., 40%), select at least one feature, and click "build model". Take note of the model performance using one of the metrics (e.g., quadratic weighted kappa). Click the reset button.
2. Holding the test sample size and features constant, repeat step #1 several times to observe the impact of chance variability on model performance.



How might re-sampling (with replacement) be used as a different approach to validate your model?

7.19 Bookend: Section 2





This is the end of this section.

[Main Menu](#)

7.20 Module Cover (END)

