Automated scoring of Critical Thinking in Accounting

Peter W. Foltz & Mark Rosenstein
Pearson

Matthew Schultz, Joshua Stopek
AICPA

The Writing Problem

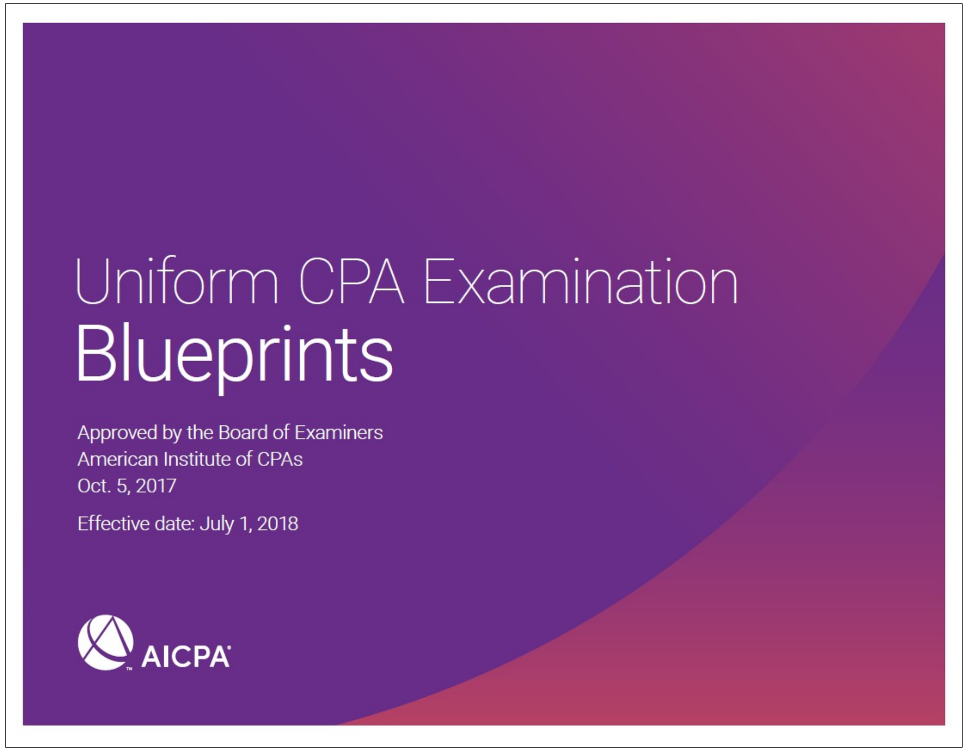
We hear about the importance of writing and critical thinking from instructors. Writing, however, is difficult to integrate into weekly assignments, core curriculum, and high stakes tests.

People want **authentic tasks**, which often can only be assessed by humans

It takes **too much time** to grade essays and provide feedback.

For AICPA, they wanted to assess critical thinking in a high stakes accounting exam while keeping assessment costs low

CPA Exam Blueprints



Uniform CPA Examination
Blueprints

Approved by the Board of Examiners
American Institute of CPAs
Oct. 5, 2017

Effective date: July 1, 2018

AICPA

aicpa.org/examblueprints

Content Organization & Weighting

Skill Allocation & Weighting

Representative Tasks

References

Exam basics

4 x 4

Exam sections

hours of testing per section

18

months to pass all sections

75

or higher on each section

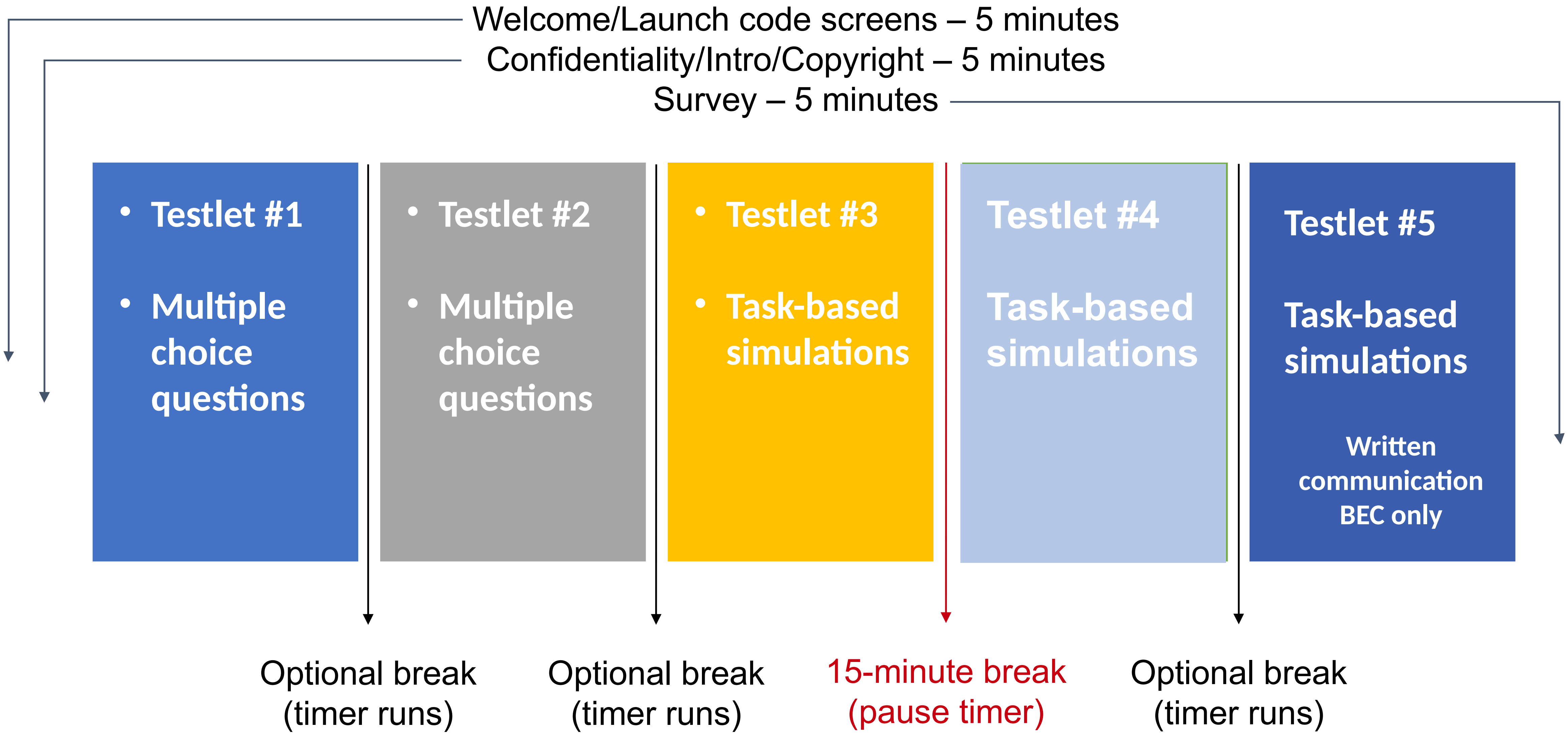
5

testlets per section

15

minute break option to pause the Exam timer

Exam structure



Welcome/Launch code screens – 5 minutes

Confidentiality/Intro/Copyright – 5 minutes

Survey – 5 minutes

- Testlet #1
- Multiple choice questions

- Testlet #2
- Multiple choice questions

- Testlet #3
- Task-based simulations

Testlet #4

Task-based simulations

Testlet #5

Task-based simulations

Written communication BEC only

Optional break (timer runs)

Optional break (timer runs)

15-minute break (pause timer)

Optional break (timer runs)

Higher order skills

Critical thinking

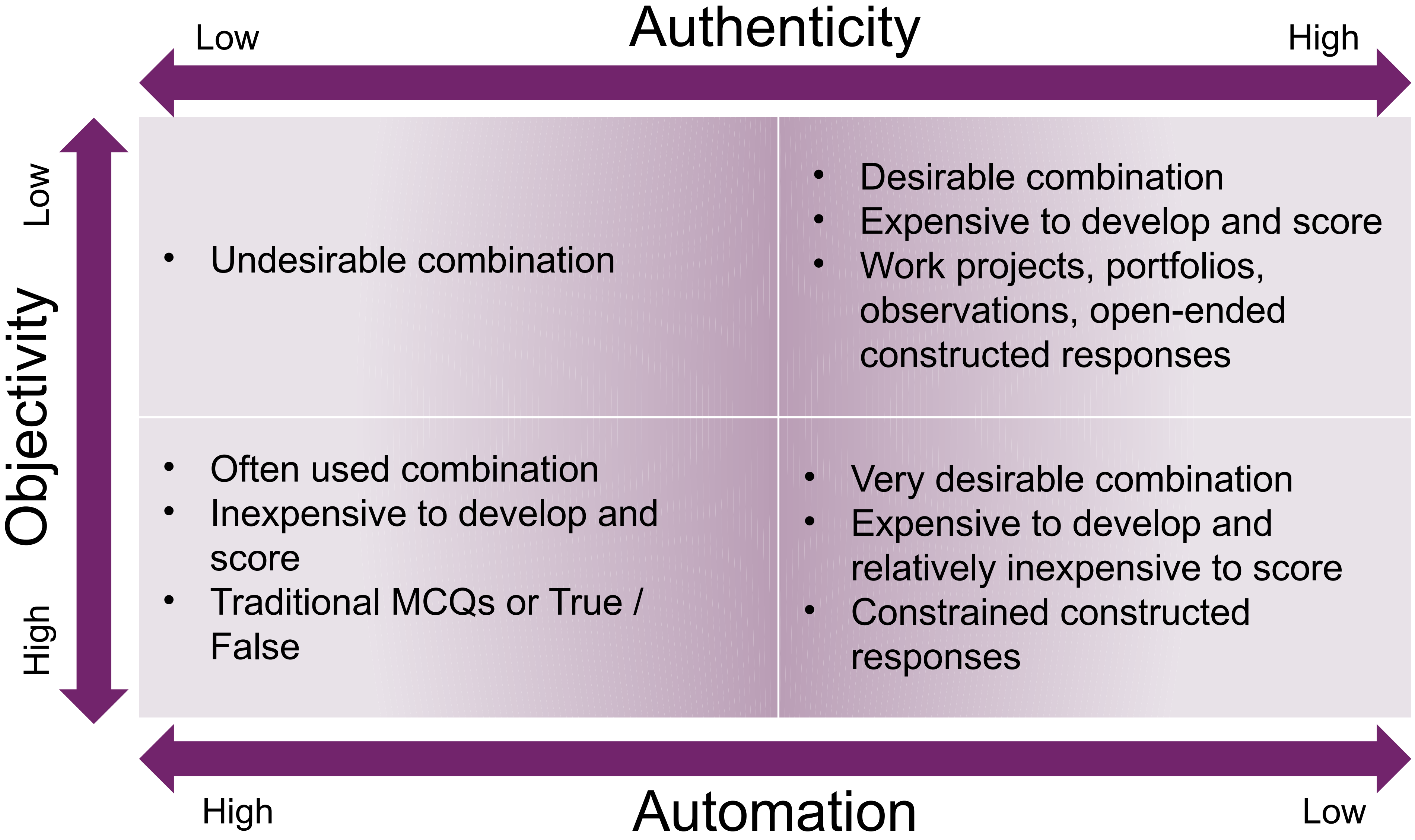
Analytical ability

Problem solving

Professional skepticism

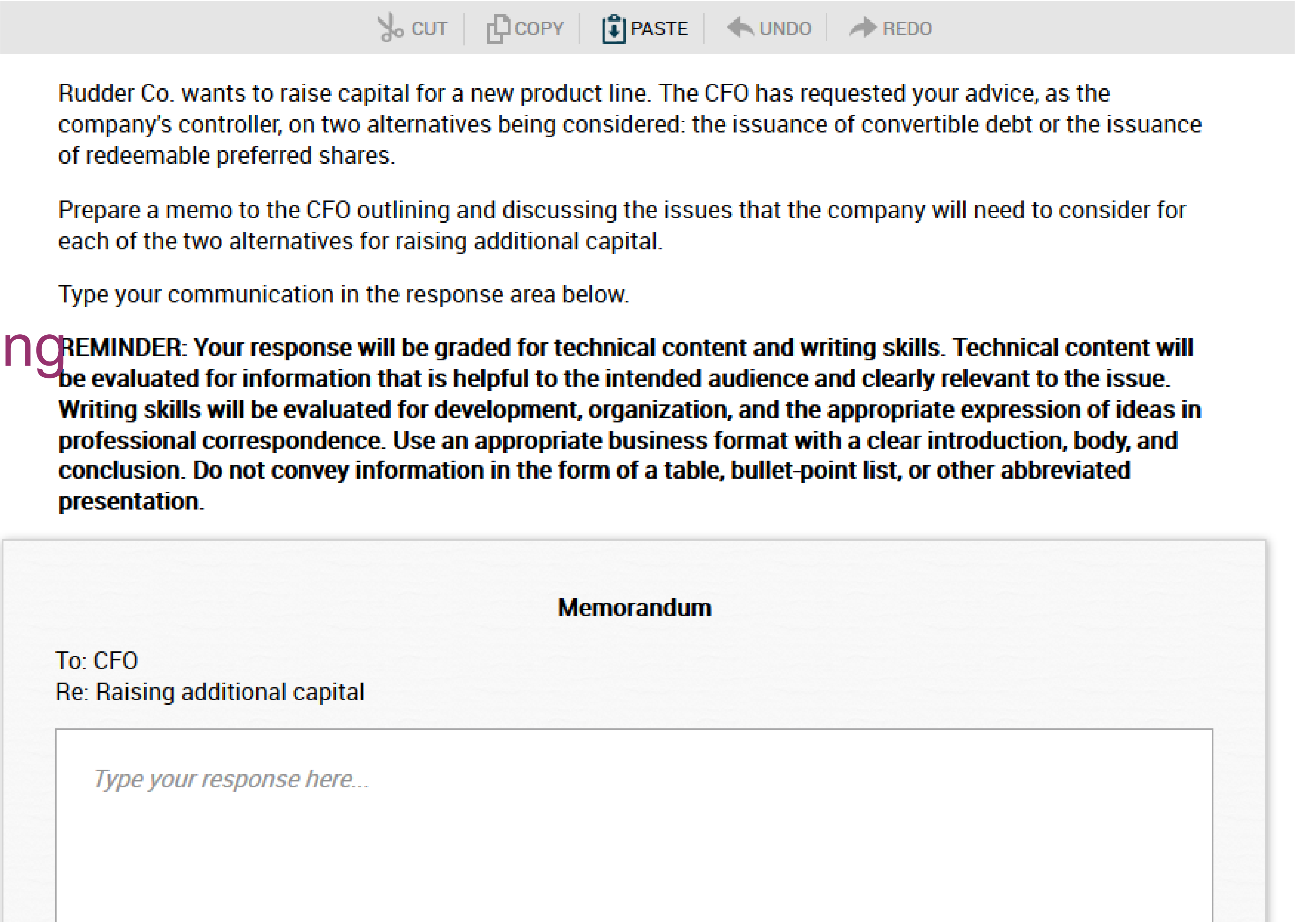
Effective communication

Objectivity vs. Automation vs. Authenticity



	Authenticity: Low Automation: High	Authenticity: High Automation: Low
Objectivity: Low	• Undesirable combination	• Desirable combination • Expensive to develop and score • Work projects, portfolios, observations, open-ended constructed responses
Objectivity: High	• Often used combination • Inexpensive to develop and score • Traditional MCQs or True / False	• Very desirable combination • Expensive to develop and relatively inexpensive to score • Constrained constructed responses

AICPA’s starting point



Rudder Co. wants to raise capital for a new product line. The CFO has requested your advice, as the company's controller, on two alternatives being considered: the issuance of convertible debt or the issuance of redeemable preferred shares.

Prepare a memo to the CFO outlining and discussing the issues that the company will need to consider for each of the two alternatives for raising additional capital.

Type your communication in the response area below.

REMINDER: Your response will be graded for technical content and writing skills. Technical content will be evaluated for information that is helpful to the intended audience and clearly relevant to the issue. Writing skills will be evaluated for development, organization, and the appropriate expression of ideas in professional correspondence. Use an appropriate business format with a clear introduction, body, and conclusion. Do not convey information in the form of a table, bullet-point list, or other abbreviated presentation.

Memorandum

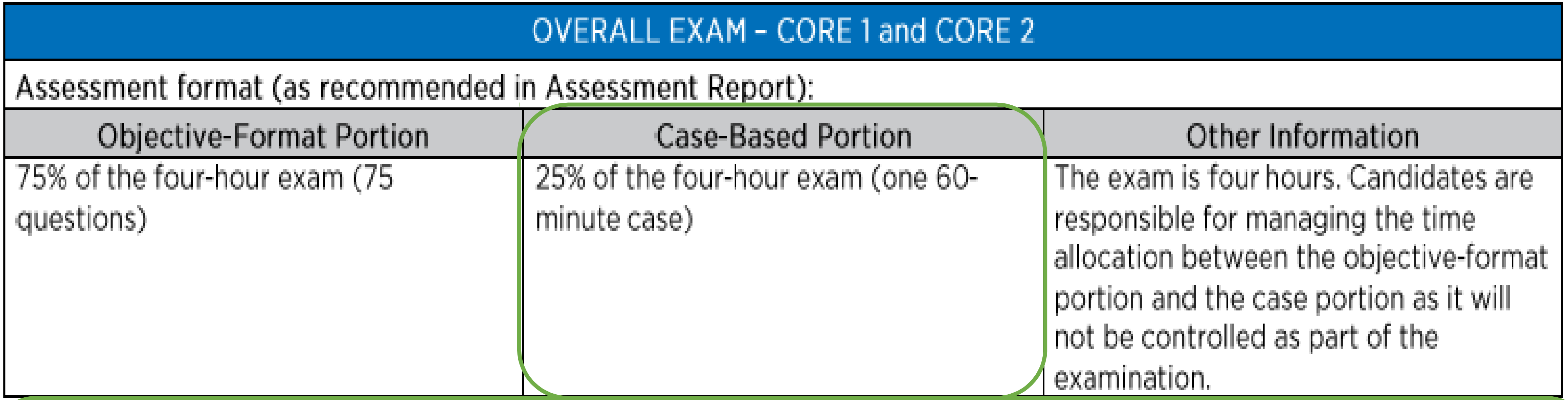
To: CFO
Re: Raising additional capital

Type your response here...

Synergy

CPA CHARTERED PROFESSIONAL ACCOUNTANTS CANADA

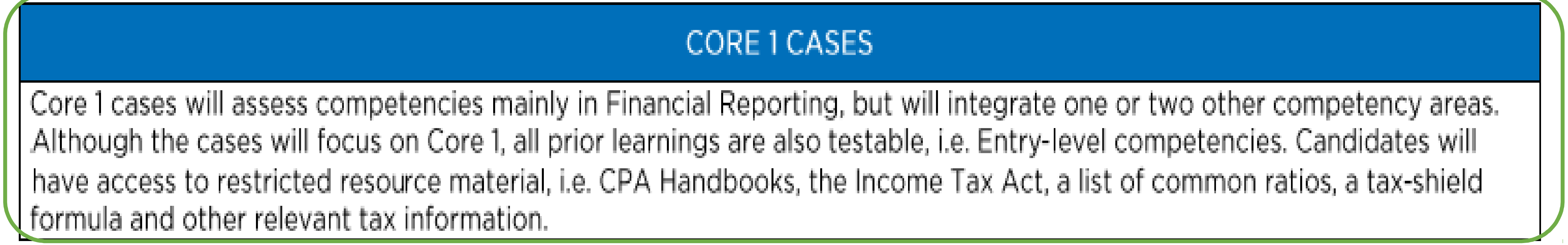
AICPA®



OVERALL EXAM – CORE 1 and CORE 2

Assessment format (as recommended in Assessment Report):

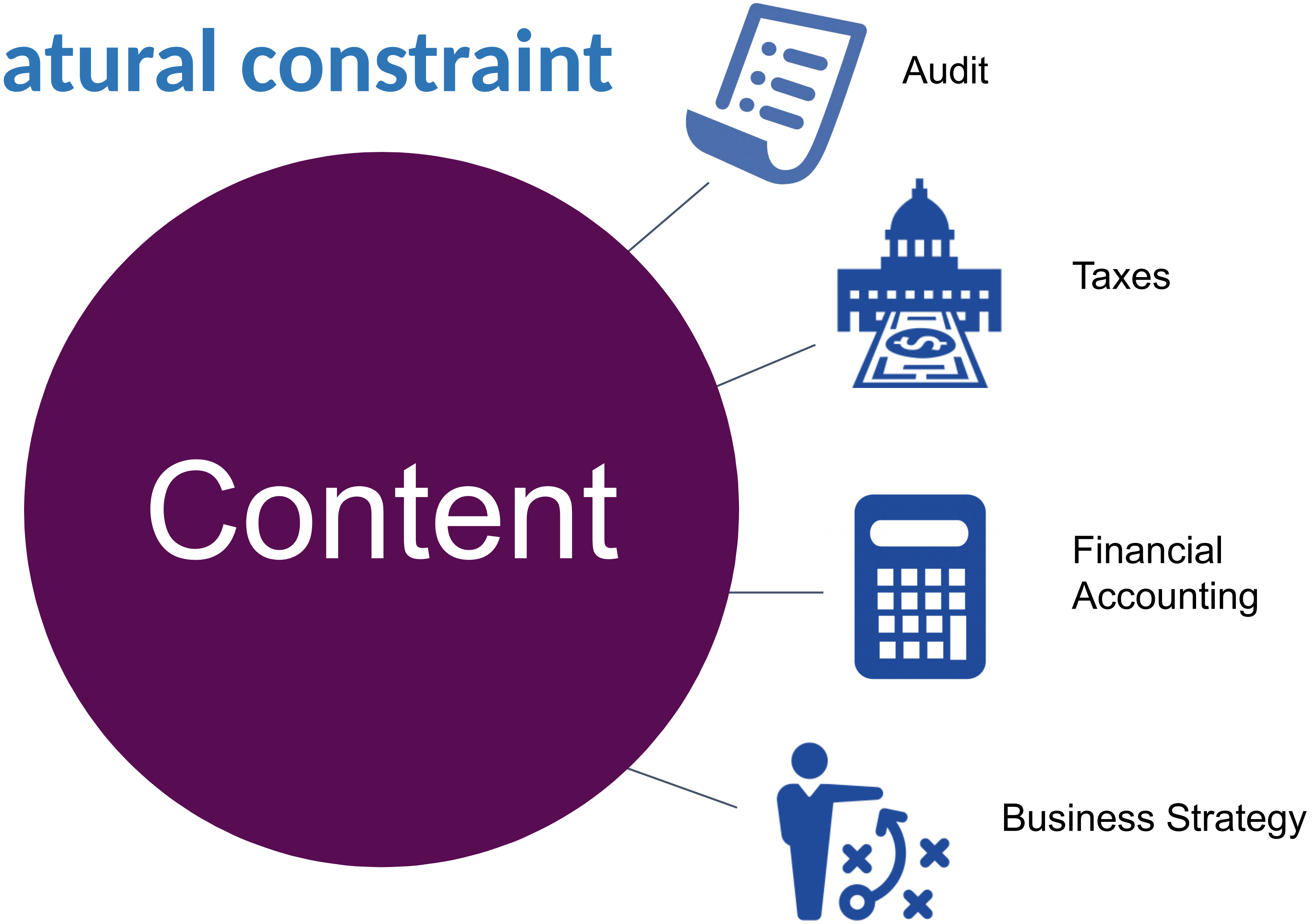
Objective-Format Portion	Case-Based Portion	Other Information
75% of the four-hour exam (75 questions)	25% of the four-hour exam (one 60-minute case)	The exam is four hours. Candidates are responsible for managing the time allocation between the objective-format portion and the case portion as it will not be controlled as part of the examination.



CORE 1 CASES

Core 1 cases will assess competencies mainly in Financial Reporting, but will integrate one or two other competency areas. Although the cases will focus on Core 1, all prior learnings are also testable, i.e. Entry-level competencies. Candidates will have access to restricted resource material, i.e. CPA Handbooks, the Income Tax Act, a list of common ratios, a tax-shield formula and other relevant tax information.

Size natural constraint



Content

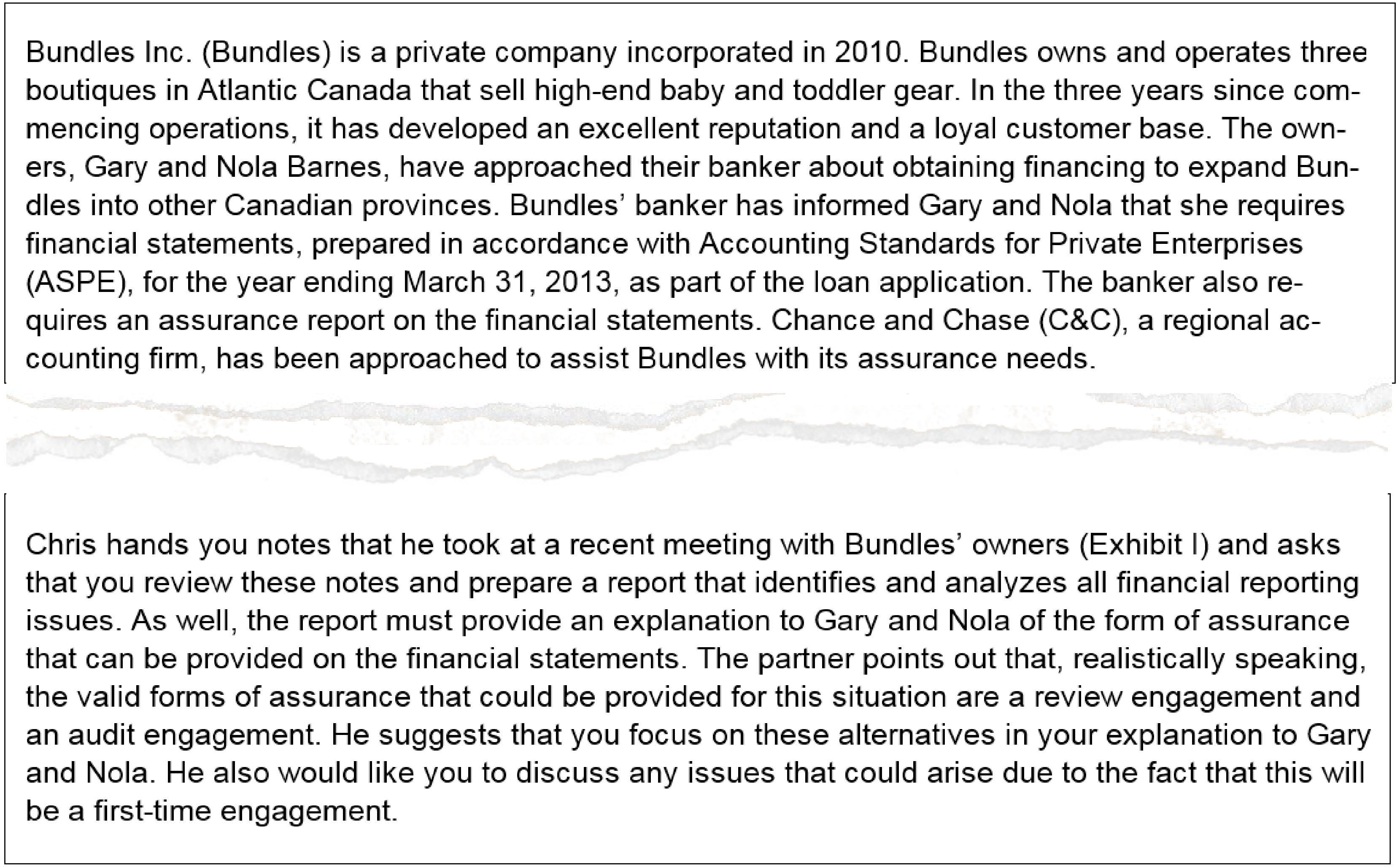
Audit

Taxes

Financial Accounting

Business Strategy

Stimuli



Bundles Inc. (Bundles) is a private company incorporated in 2010. Bundles owns and operates three boutiques in Atlantic Canada that sell high-end baby and toddler gear. In the three years since commencing operations, it has developed an excellent reputation and a loyal customer base. The owners, Gary and Nola Barnes, have approached their banker about obtaining financing to expand Bundles into other Canadian provinces. Bundles' banker has informed Gary and Nola that she requires financial statements, prepared in accordance with Accounting Standards for Private Enterprises (ASPE), for the year ending March 31, 2013, as part of the loan application. The banker also requires an assurance report on the financial statements. Chance and Chase (C&C), a regional accounting firm, has been approached to assist Bundles with its assurance needs.

Chris hands you notes that he took at a recent meeting with Bundles' owners (Exhibit I) and asks that you review these notes and prepare a report that identifies and analyzes all financial reporting issues. As well, the report must provide an explanation to Gary and Nola of the form of assurance that can be provided on the financial statements. The partner points out that, realistically speaking, the valid forms of assurance that could be provided for this situation are a review engagement and an audit engagement. He suggests that you focus on these alternatives in your explanation to Gary and Nola. He also would like you to discuss any issues that could arise due to the fact that this will be a first-time engagement.

4-6 pages of information

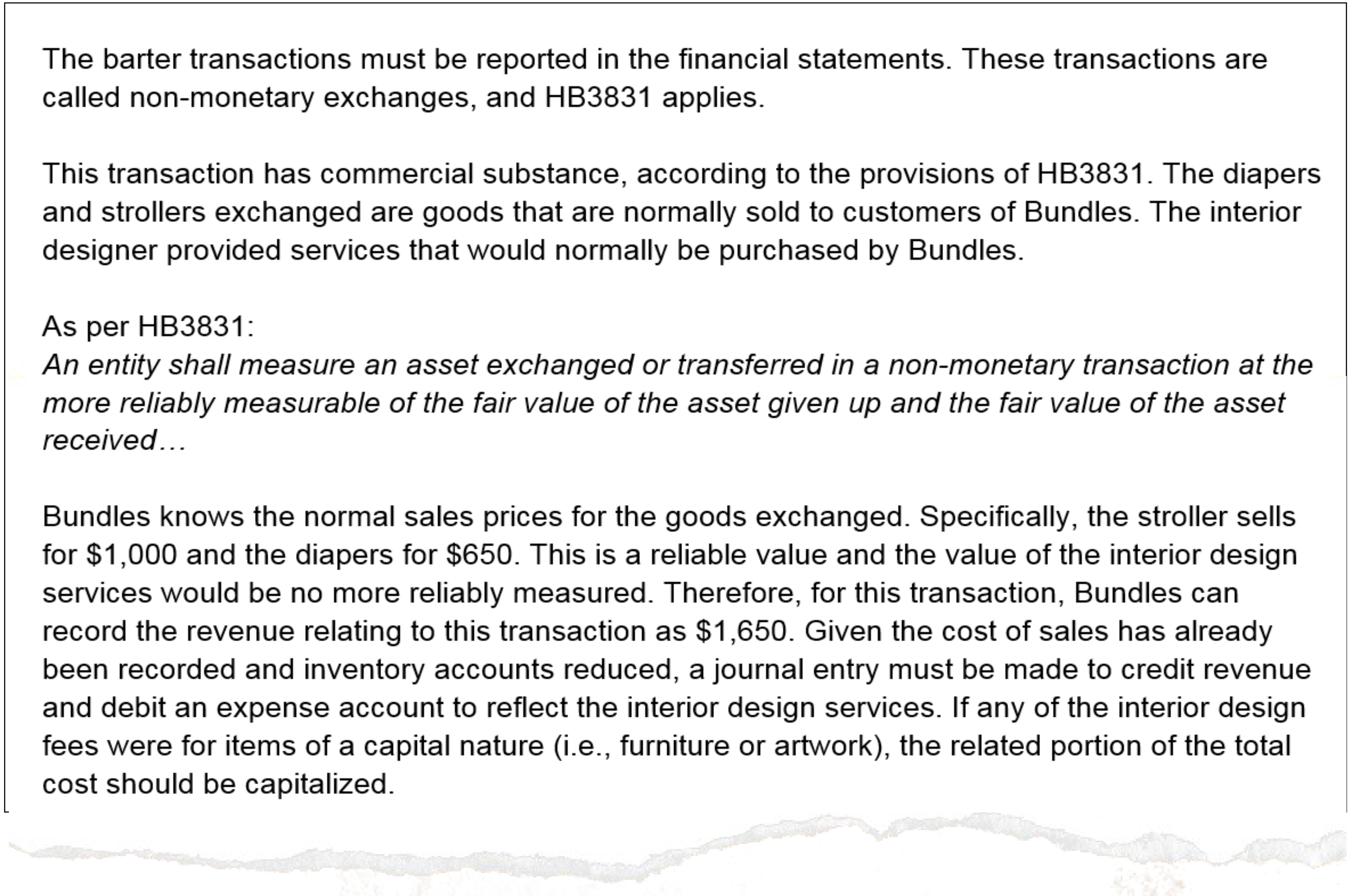
Inventory Adjustments

Bartering

Consignment

Intangibles

Model answer - bartering



The barter transactions must be reported in the financial statements. These transactions are called non-monetary exchanges, and HB3831 applies.

This transaction has commercial substance, according to the provisions of HB3831. The diapers and strollers exchanged are goods that are normally sold to customers of Bundles. The interior designer provided services that would normally be purchased by Bundles.

As per HB3831:
An entity shall measure an asset exchanged or transferred in a non-monetary transaction at the more reliably measurable of the fair value of the asset given up and the fair value of the asset received…

Bundles knows the normal sales prices for the goods exchanged. Specifically, the stroller sells for $1,000 and the diapers for $650. This is a reliable value and the value of the interior design services would be no more reliably measured. Therefore, for this transaction, Bundles can record the revenue relating to this transaction as $1,650. Given the cost of sales has already been recorded and inventory accounts reduced, a journal entry must be made to credit revenue and debit an expense account to reflect the interior design services. If any of the interior design fees were for items of a capital nature (i.e., furniture or artwork), the related portion of the total cost should be capitalized.

Complete response: 4-6 pages

Rubrics

Assessment Opportunity #1 (FR) - Issue X

Issue X	Identify	Analyze	Conclude	Amortization
Assessment Opportunity 1 Overall Assessment	**NC**	**RC**	**C**	**CD**

Assessment Opportunity #2 (FR) - Issue Y ***(Note, some issues have multiple subcontent that they're looking for)***

Content A	Identify (Weakness)	Analyze (Implication)	Recommendation	
Content B	Identify (Weakness)	Analyze (Implication)	Recommendation	
Assessment Opportunity 1 Overall Assessment	**NC**	**RC**	**C**	**CD**

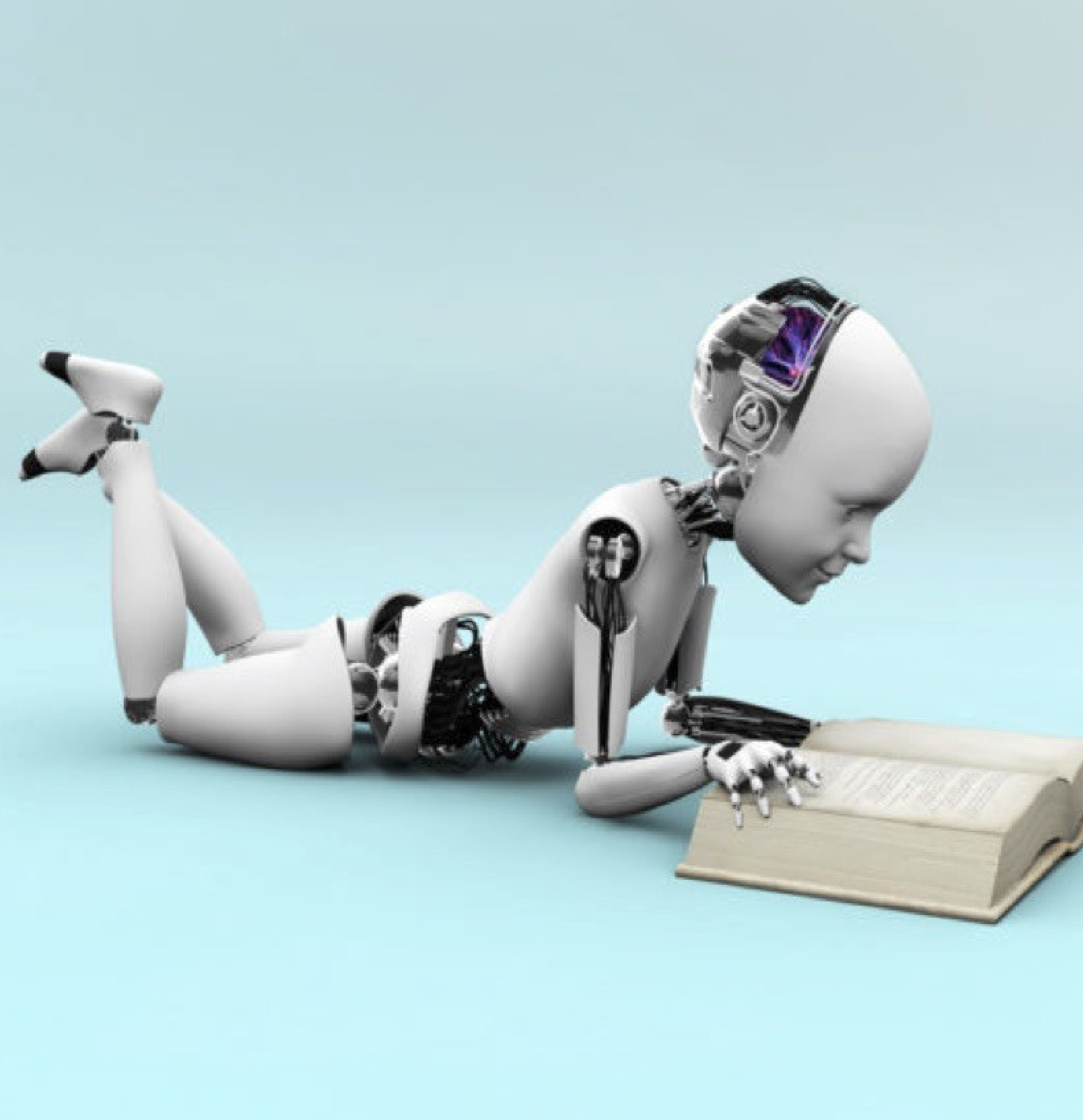
NC: Not Competent

RC: Reaching Competence

C: Competence

CD: Competent w/ Distinction

Building an automated scoring model



Creating a scoring model

- **Inferring teacher scoring behavior**
- Computer learns background knowledge of the domain by “reading” a large amount of text (corpus)
- Computer is trained on a large sample (200-1000+) of human-scored/annotated essays
- Analyze construct relevant features
- Machine learning to infer the combination and weighting (or types of feedback) for particular writing traits

Stages in developing automated scoring

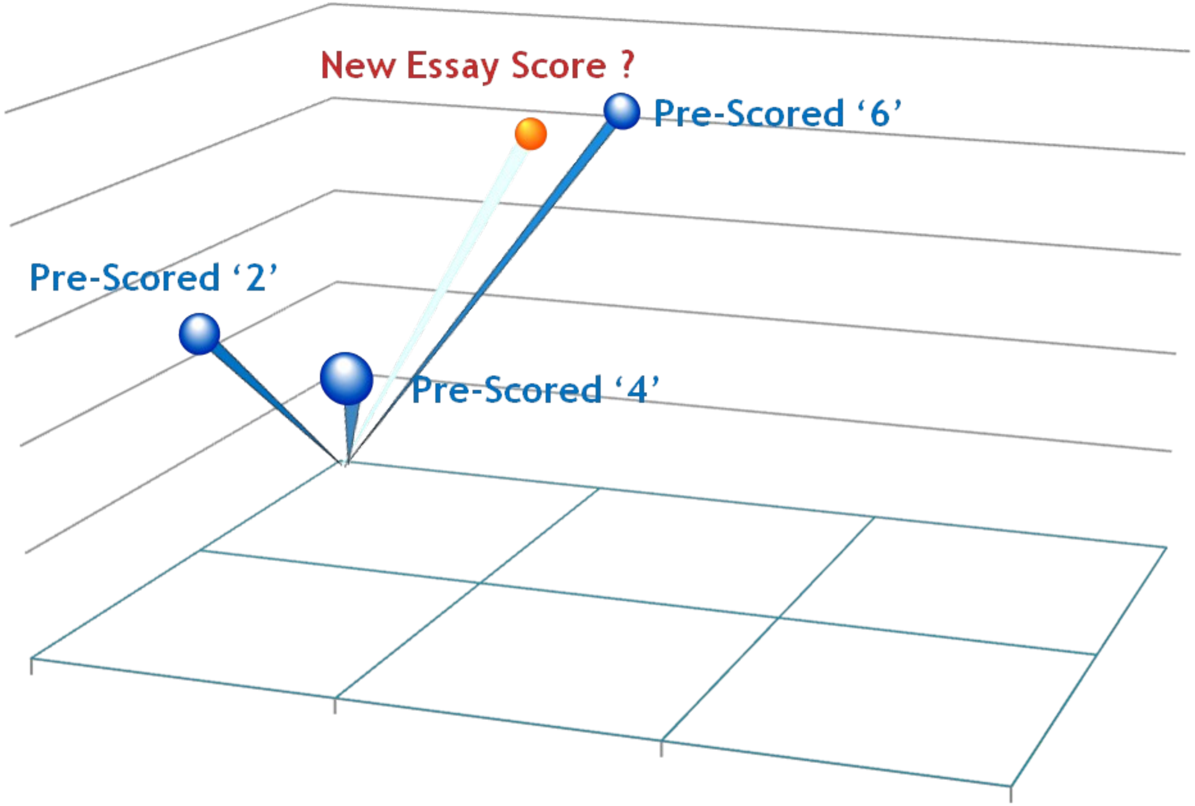
- **Collect training essays**
  - 100s to 1000 per prompt topic
- **Analyze Language Features**
  - Content/domain features
  - Writing features
- **Build Scoring model**
  - Machine Learning to weigh and combine features
- **Validation**
- **Deployment**

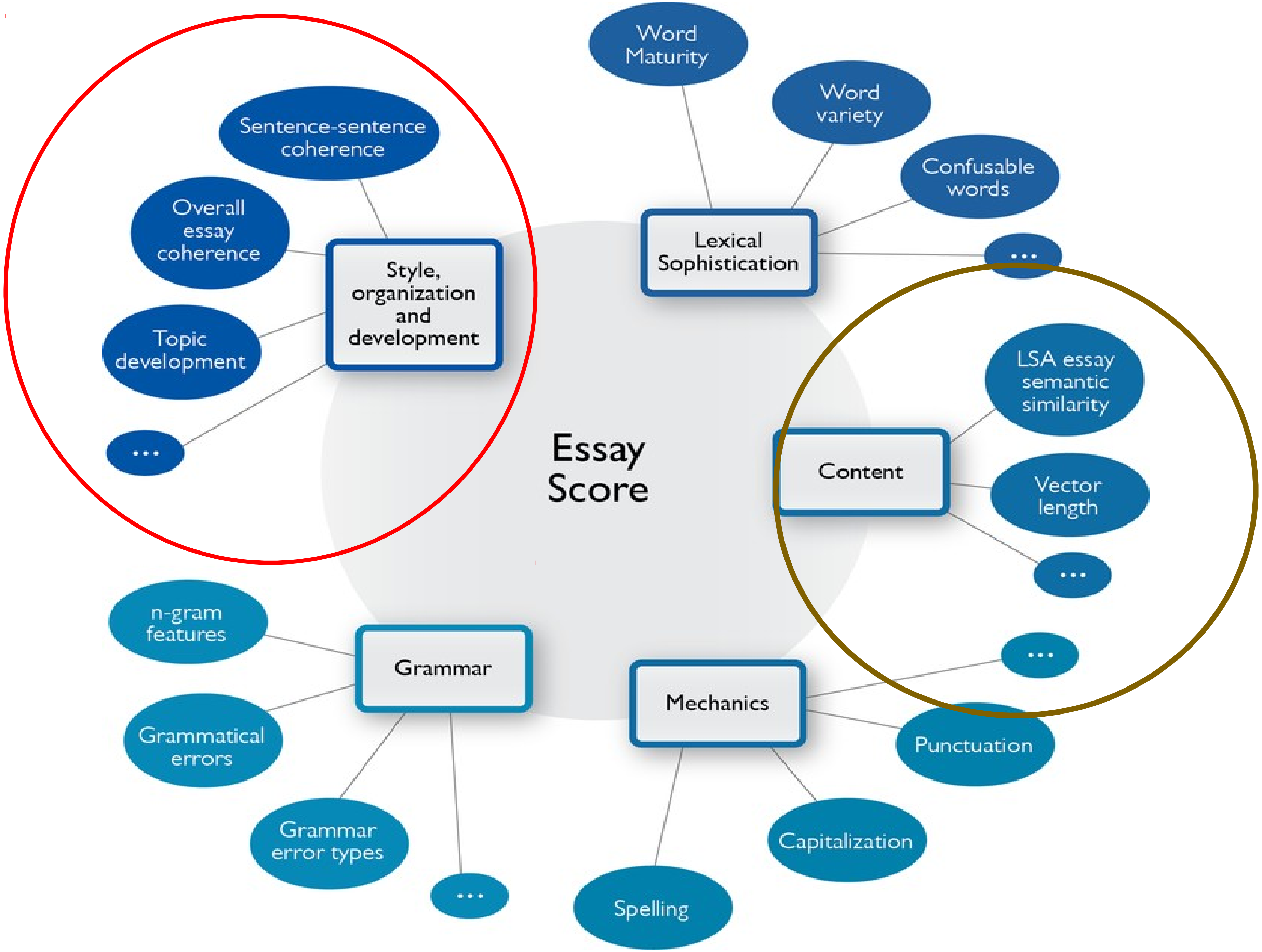
How to score content

- AI-based models learn semantic content (meaning) of domain
- Model trained on corpus of domain-relevant content
  - (e.g., accounting texts or general language corpus)
- Computational semantic model
  - Compares essays against other essays with known scores
  - Computes coverage of content in AOs

Content Scoring

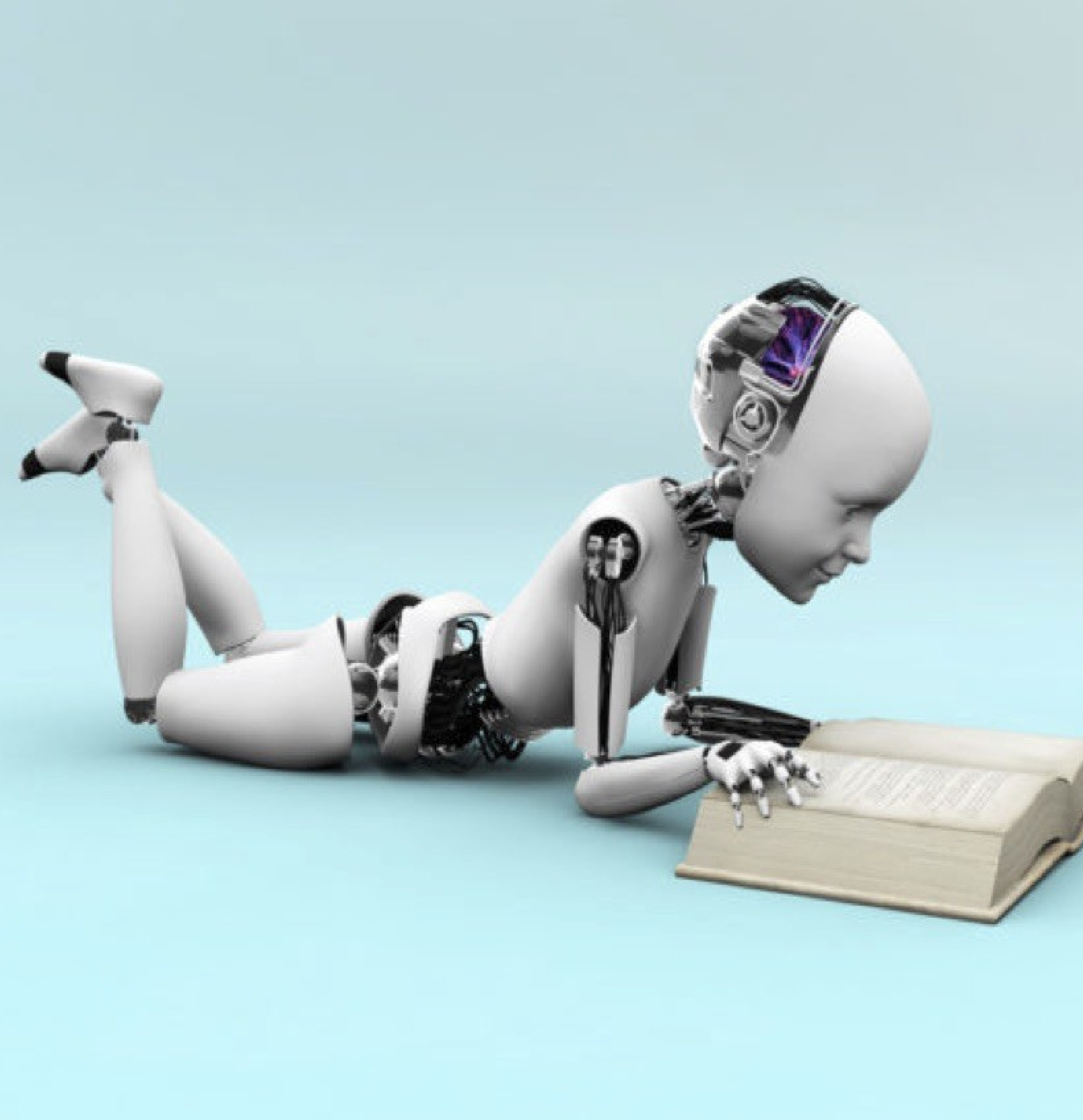
- Essays (or AOs) represented as a vector based on semantic features
- New essays vectors compared against pre-scored essays




Essay Score
Style, organization and development
Sentence-sentence coherence
Overall essay coherence
Topic development
...
Lexical Sophistication
Word Maturity
Word variety
Confusable words
...
Content
LSA essay semantic similarity
Vector length
...
Grammar
n-gram features
Grammatical errors
Grammar error types
...
Mechanics
...
Punctuation
Capitalization
Spelling

Pearson

Results of Modeling



Modeling goals

- **How well can automated scoring using the Intelligent Essay Assessor assess higher order skills?**
- Predict examinee overall scores on written assessment
- Predict performance for individual assessment opportunities (AOs)
- Predict overall pass/fail of exam*
- How well does it work compared to human performance?
- Explore potential use models and considerations for operational implementation

Data

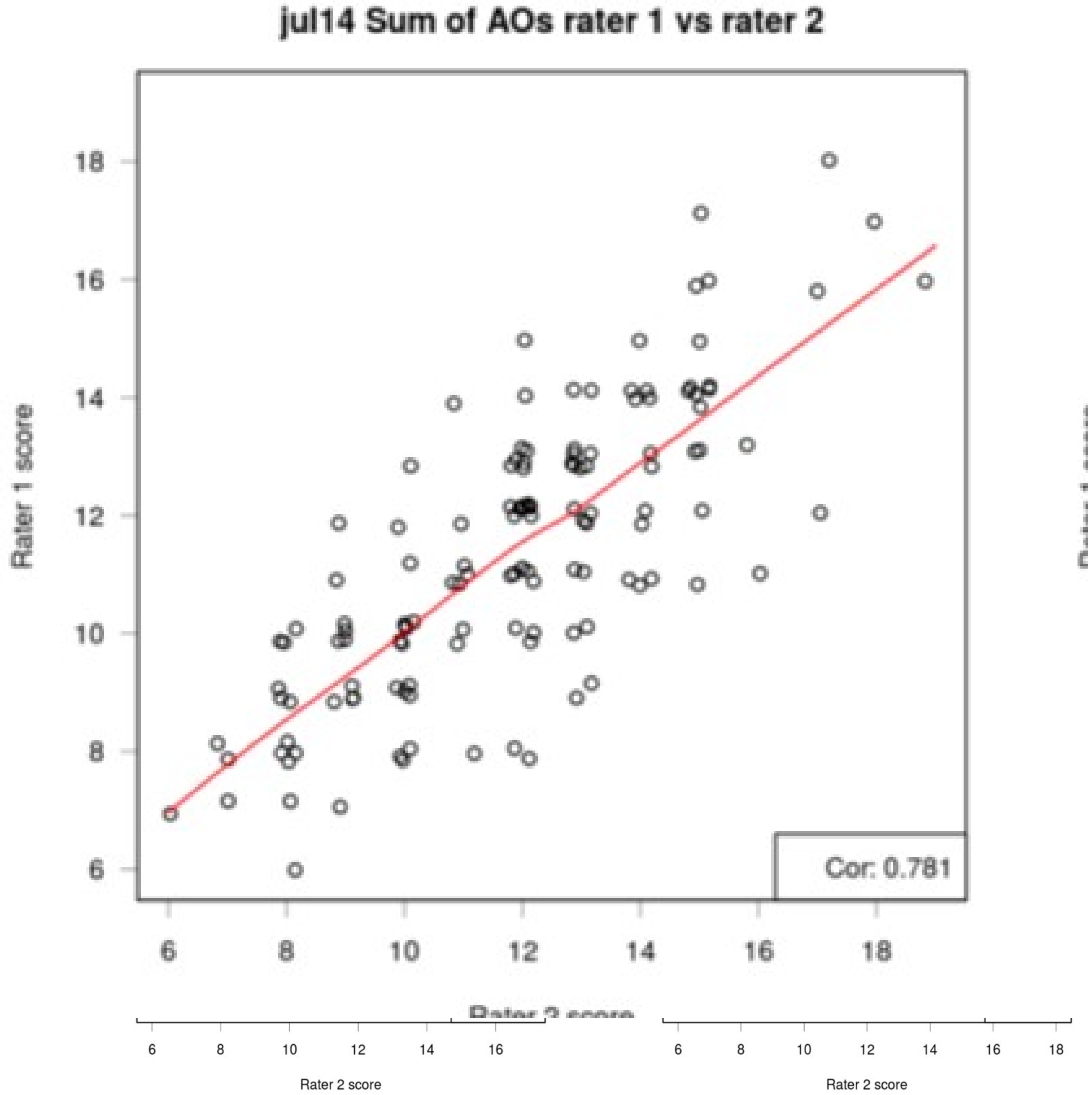
- Four long format constructed response item administrations from CPA Canada
  - Three separate items, one repeated
- Test-taker responses
  - Mean length: ~1200 words
- Overall human scores (0-30 scale)
- Human scores on six Assessment Opportunities (AOs) (0-4 scale)
- Overall pass/fail of the entire examination

Human rating

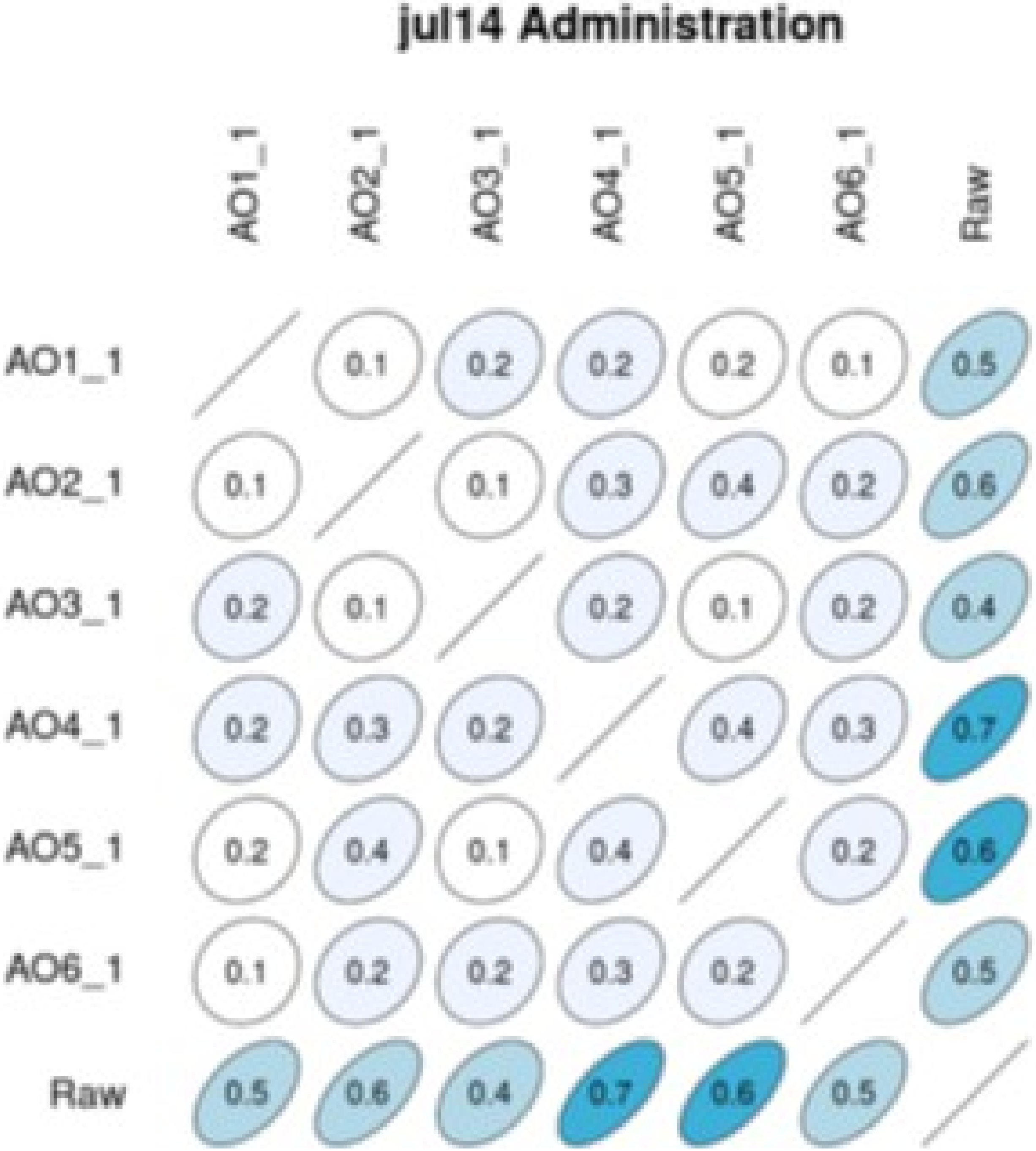
- 21-37% AOs received “second” score and resolution/final score.
- Inter-rater agreement of summed AO scores

Administration	Correlation	Exact agreement	Adjacent agreement
14-Jul	0.78	31.1	69.6
14-Nov	0.67	26.5	64.8
15-Mar	0.74	23.1	62.0
15-Aug	0.73	30.1	65.4

Inter-rater agreement



Independence of AOs



jul14 Administration

	AO1_1	AO2_1	AO3_1	AO4_1	AO5_1	AO6_1	Raw
AO1_1		0.1	0.2	0.2	0.2	0.1	0.5
AO2_1	0.1		0.1	0.3	0.4	0.2	0.6
AO3_1	0.2	0.1		0.2	0.1	0.2	0.4
AO4_1	0.2	0.3	0.2		0.4	0.3	0.7
AO5_1	0.2	0.4	0.1	0.4		0.2	0.6
AO6_1	0.1	0.2	0.2	0.3	0.2		0.5
Raw	0.5	0.6	0.4	0.7	0.6	0.5	

Developing an automated scoring model

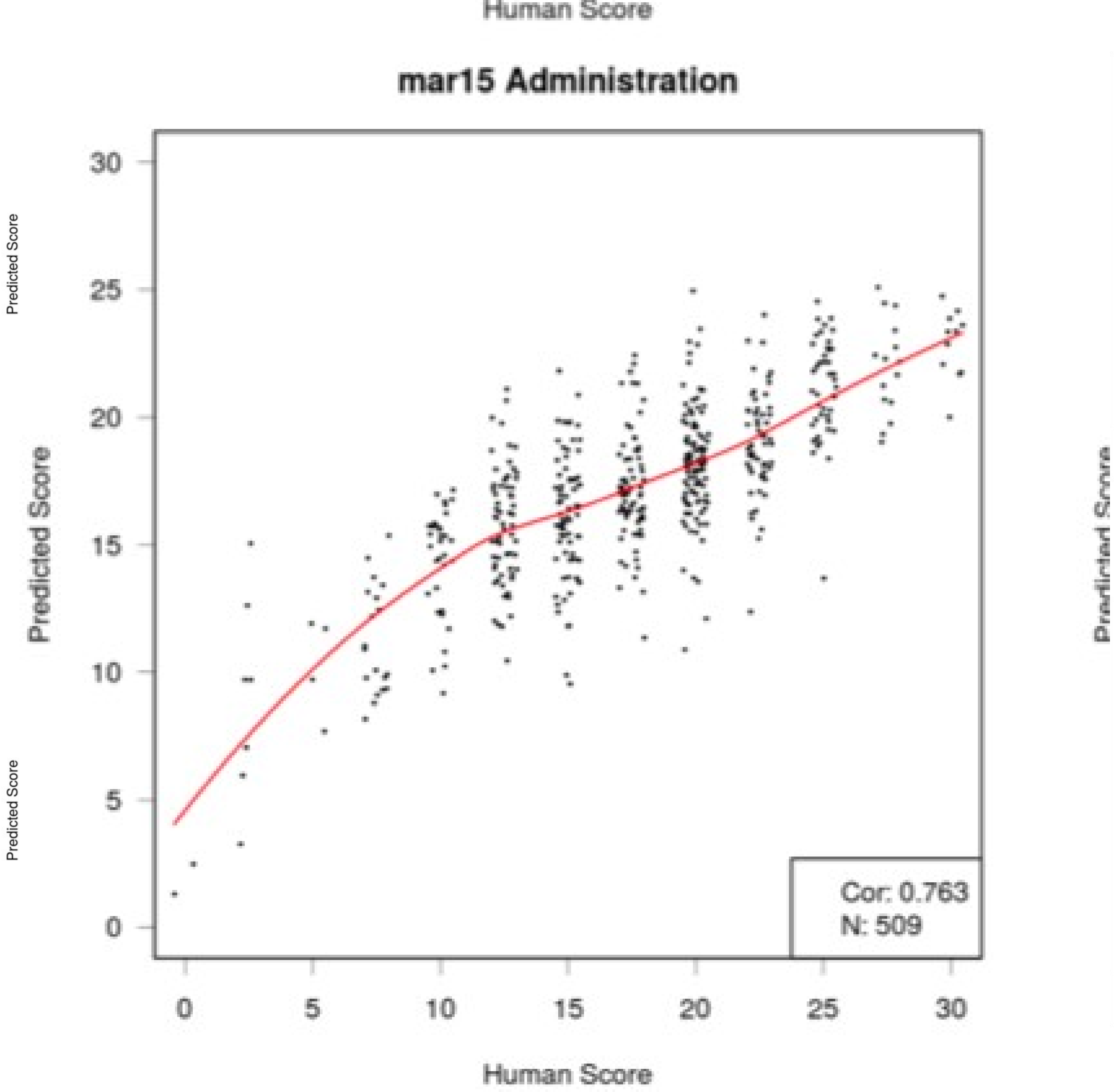
Administration	Training Set size	Held out test set size
14-Jul	400	236
14-Nov	400	1749
25-Mar	400	509
15-Aug	300	70

Model 1:
Predicting Combined scores of AOs

Agreement for Overall score model

admin	Computer to human correlation
14-Jul	0.79
14-Nov	0.75
15-Mar	0.76
15-Aug	0.79

Agreement


Human Score
mar15 Administration
Predicted Score
Predicted Score
Predicted Score
30
25
20
15
10
5
0
0
5
10
15
20
25
30
Human Score
Cor: 0.763
N: 509
Predicted Score

Model 2:
Predicting individual
AO scores

Exact Agreement on AO scores (1-4 range)

- 24 Models (6 AOs by 4 Administrations)

	Computer to human exact agreement	Human to human exact agreement
Mean	62%	67%
Min	39%	48%
Max	88%	88%

Model 3: Predicting Overall scores using combined models

Combining individual AO models with the holistic model

admin	Human – Human correlation
14-Jul	0.78
14-Nov	0.67
15-Mar	0.74
15-Aug	0.73

Implications/Conclusions

Conclusions

- Individual AO models perform slightly below the level of human performance
- Holistic models perform at or above the level of human performance
- Aggregated scores based on multiple AO models can perform better than humans
- Combining AO model with Holistic models further improves performance.

- *Automated scoring models built on training data can perform at a level of accuracy equivalent, or slightly better* than that of human scorers.*
  - *Note difficult to impute performance as better than human without external validation measures.

Implications/Use cases

- Automated scoring can assess higher order skills in complex scenarios

- Formative
  - Oslo/GLP

- Summative
  - Automated scoring used as a check for human scorers.
  - One human and one automated rater.
  - Automated scorer as the primary scorer with human backreads and additional checks.
  - Detection of responses near critical boundaries such pass/fail thresholds.

Help Wanted
Test Scorers
(apply within)

Questions?