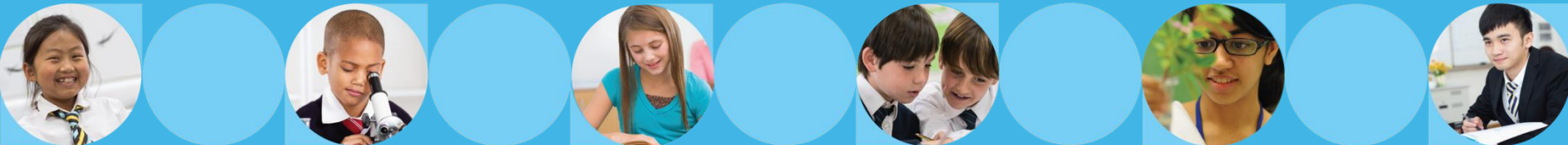


A framework for describing comparability between alternative assessments

Stuart Shaw, Vicki Crisp and Sarah Hughes

IAEA, Baku

September, 2019



Project aims – Phase 1

Research Question: *What guidance could test developers follow to ensure that outcomes from on-screen and on-paper versions of tests are comparable?*



- ▶ describe comparability in terms of the standards of an assessment
- ▶ consider comparability intentions and claims
- ▶ monitor and evaluate if intentions and claims are met
- ▶ reassure customers of comparability
- ▶ offer a means for evaluating other assessments (including our competitors')

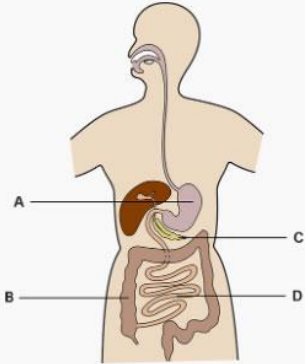
On-screen and paper-based examples

Stage 8 Science Paper 1

https://cambridge.inspera.no/player/?assessmentRunId=28581503&context=...

Candidate ID
Connected 1 week, 43 hours, 23 minutes remaining

The diagram shows the digestive system.



For each letter select the correct organ and its function in the digestive system.
One has been done for you.

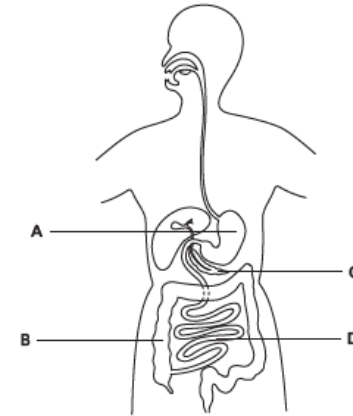
letter	organ	function
A	stomach	digests protein
B	<input type="text"/>	<input type="text"/>
C	<input type="text"/>	<input type="text"/>
D	<input type="text"/> <ul style="list-style-type: none"> large intestine pancreas small intestine stomach 	<input type="text"/>

Reset [3]

© UCLES 2017

2

1 The diagram shows the digestive system.



Draw a line from each letter to the correct organ and its function in the digestive system.

One has been done for you.

organ	letter	function
large intestine	A	absorbs water
stomach	B	digests protein
small intestine	C	produces enzymes
pancreas	D	absorbs digested food

For
Teacher's
Use

[3]

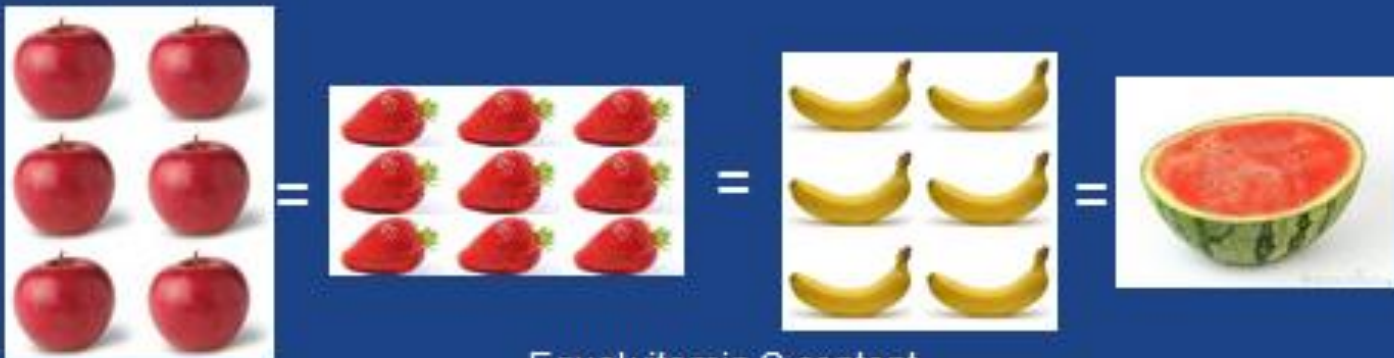
Comparability of what?



Equal weight



Equal sugar content



Equal vitamin C content

Comparability Standards



Content

Value of relevance of the content



Demand

KSU required for success



Marking

How marks are assigned



Awarding

Performance worthy of the grade

Comparability of

Content standards	Demand standards	Marking standards	Awarding standards
<p>If it is the intention that content standards are comparable across tests, the following need to be fulfilled:</p> <ul style="list-style-type: none">• subject domains are the same across tests• subject topics are the same across tests• whole test content coverage is the same across tests	<p>If it is the intention that demand standards are comparable across tests, the following need to be fulfilled:</p> <ul style="list-style-type: none">• knowledge, understanding and skills (e.g. Assessment Objectives) assessed are the same across tests• the range of kinds of questions or tasks are the same across tests (e.g. similar balance of MCQ, short answer, essay)• the test environment does not affect the nature of the teaching and learning• the test environment is easy to use and students have been given sufficient opportunity for familiarisation with the test environment• the cognitive processes (as supported by tools) are the same across tests as far as we can tell• the possible effects of any differences in response format are carefully considered (e.g. for on-screen tests, the effects of typing rather than writing on paper, or of using a drop down list rather than circling a response on paper)	<p>If it is the intention that marking standards are comparable across tests, the following need to be fulfilled:</p> <ul style="list-style-type: none">• the mark schemes reward the same knowledge, skills and understanding• the application of the mark scheme is the same across tests with markers complying with marking guidance and requirements equally across tests• the way that student responses are presented to markers needs to give equal opportunity for accurate marking across tests• marker competence/accuracy is the same across tests (ideally, the same specific markers are used for both tests)• markers are standardised appropriately for both tests and appropriate quality assurance processes are used for both tests• auto-marking (if used) and human marking are both sufficiently accurate and reward intended constructs (only relevant if comparing an on-screen test to a paper-based test)	<p>If it is the intention that awarding standards are comparable across tests, the following need to be fulfilled:</p> <ul style="list-style-type: none">• awarding is conducted separately for different tests with potentially different grade thresholds (thus ensuring comparability of awarding standards between tests even if there are differences in content, demand or marking standard)• the awarding process is the same across tests (e.g. use of judgemental and statistical evidence, methods of recording awarding decisions)• sufficient data is available to compare across tests (e.g. entry sizes, benchmark centres, syllabus pairs, knowledge of the characteristics of the candidates entering for each test)• awarding standards are maintained over time across tests

Table 2: Comparability recording form: a structure for describing test comparability across modes

Completed by (name)..... (Job Role)..... Date.....

Assessment name and code.....

1. Standard	2. Is it intended that there should be comparability between modes in terms of each standard?	3. Comparability features – these should be the same across modes if comparability between modes is intended for that standard	4. What are the differences between modes, if any, in terms of these features?	5. How have the differences been addressed (if they have been)?	6. For the standards where comparability is intended, are you satisfied that there is sufficient comparability?
Content standards		Subject domains			
		Subject topics			
		Sub-topics			
		Whole test coverage			
Demand standards		Knowledge, understanding and skills			
		Range of kinds of questions			
		Teaching and learning			
		Test environment ease of use and opportunity for familiarisation			
		Cognitive processes			
		Response format			
Marking standards		Mark schemes			
		Application of the mark scheme			
		The way that student responses are presented to markers			
		Marker competence/accuracy			
		Standardisation methods			
		Quality assurance processes			
		Any auto-marking is sufficiently accurate and rewards intended constructs			
Awarding standards		Awarding conducted separately for different modes			
		Awarding process			
		Sufficient data is available			
		Awarding standards are			

Project aims – Phase 2

- ▶ A pilot using two assessment contexts



to pilot the comparability framework and revise it if needed



to identify the most suitable personnel to use the framework and recording form



to provide guidance on the use of the framework and completion of the recording form

Phase 2: Method

- ▶ Scoping project and assessment contexts with colleagues
- ▶ Assessment contexts selected:
 - ▶ On-screen and paper-based tests: Stage 8 Progression tests in science for 2018, Papers 1 and 2
 - ▶ An Alternative to Practical paper and a Practical test: IGCSE Chemistry (0620, Time zone X, Papers 51 and 61 for June 2017)
- ▶ Initial exercise: We attempted to apply the framework and form to these assessments – appeared viable
- ▶ Main piloting: An assessor familiar with the assessments was asked to:
 - ▶ Read the phase 1 report
 - ▶ Re-familiarise themselves with the assessment materials
 - ▶ Attempt to complete the comparability recording form
 - ▶ Complete a questionnaire about the framework and form

Outcomes

Understandable/useable

- ▶ The framework was generally considered understandable
- ▶ Demand standards were considered most difficult to comprehend, but also most thought-provoking in terms of possible differences between papers or modes (e.g. cognitive processes)
- ▶ A few cells left blank - due to individual not being involved in a stage of assessment process?
- ▶ Columns 4 and 5 – if analysing retrospectively, differences already addressed as far as possible
- ▶ Some differences identified but answered ‘yes’ to column 6. Bias?
Genuinely trivial differences?

Outcomes

Usefulness

- ▶ Considered useful in terms of providing criteria for evaluating comparability (e.g. rather than just focusing on content)
- ▶ Provides evidence to support stated claims for comparability
- ▶ Could be used for other contexts (e.g. time-zoned papers)

Frequency of use

- ▶ Not thought to be necessary every time parallel assessments are written
- ▶ Possible use: when syllabuses reviewed; first time a alternative assessment is created (to parallel an existing assessment)

Comparability recording form: a structure for describing test comparability across tests

Completed by (name)..... (Job Role)..... Date.....

Assessment name and code.....

1. Standard	2. Is it intended that there should be comparability between tests in terms of each standard?	3. Comparability features – these should be the same across tests if comparability between tests is intended for that standard	4. What are the differences between tests, if any, in terms of these features? (Notes can be included on actions taken to minimise differences)	5. How have the differences been addressed (if they have been)?	5. For the standards where comparability is intended, are you satisfied that there is sufficient comparability?
Content standards		Subject domains			
		Subject topics			
		Sub-topics			
		Whole test coverage			
Demand standards		Knowledge, understanding and skills (e.g. Assessment Objectives)			
		Range of kinds of questions/tasks			
		Teaching and learning			
		Test environment ease of use and opportunity for familiarisation			
		Cognitive processes			
		Response format			
Marking standards		Mark schemes			
		Application of the mark scheme			
		The way that student responses are presented to markers			
		Marker competence/accuracy			
		Standardisation methods and any other quality assurance processes			
		Any auto-marking is sufficiently accurate and rewards intended constructs (if relevant)			
Awarding standards		Awarding conducted separately for different tests			
		Awarding process			
		Sufficient data is available			
		Awarding standards are maintained over time			

Who should complete the form?

- ▶ Range of personnel involved in different stages of assessment process could complete parts of the form
- ▶ Product Managers could complete the intended comparability claims and then manage completion of form by relevant personnel
- ▶ Possible general pattern (adjust if appropriate):

1 Standard	2. Is it intended that there should be comparability between tests in terms of each standard?	4. What are the differences between tests, if any, in terms of these features? (Notes can be included on actions taken to minimise differences)	5. For the standards where comparability is intended, are you satisfied that there is sufficient comparability?
Content standards	Product Manager	Setter and Reviser	Product Manager
Demand standards	Product Manager	Setter and Reviser	Product Manager
Marking standards	Product Manager	Principal Examiner	Product Manager
Awarding standards	Product Manager	Principal Examiner and awarding team	Product Manager

Next steps and application of comparability framework

'Should'

- ▶ Use for development and redevelopments
- ▶ Apply to any optional assessments
- ▶ Use during development and/or after

- ▶ Attend to all cells
- ▶ Not use it as a checklist

- ▶ Help develop examiners
- ▶ Inform grading
- ▶ Include relevant personnel

- ▶ Consider how much information to share with the public

- ▶ Be stored somewhere central



... it is essential that the use of the framework be formally documented by retention of completed forms and other records.



Cambridge Assessment
International Education

Learn more!

Getting in touch with Cambridge is easy

Email info@cambridgeinternational.org
or telephone +44 1223 553554

