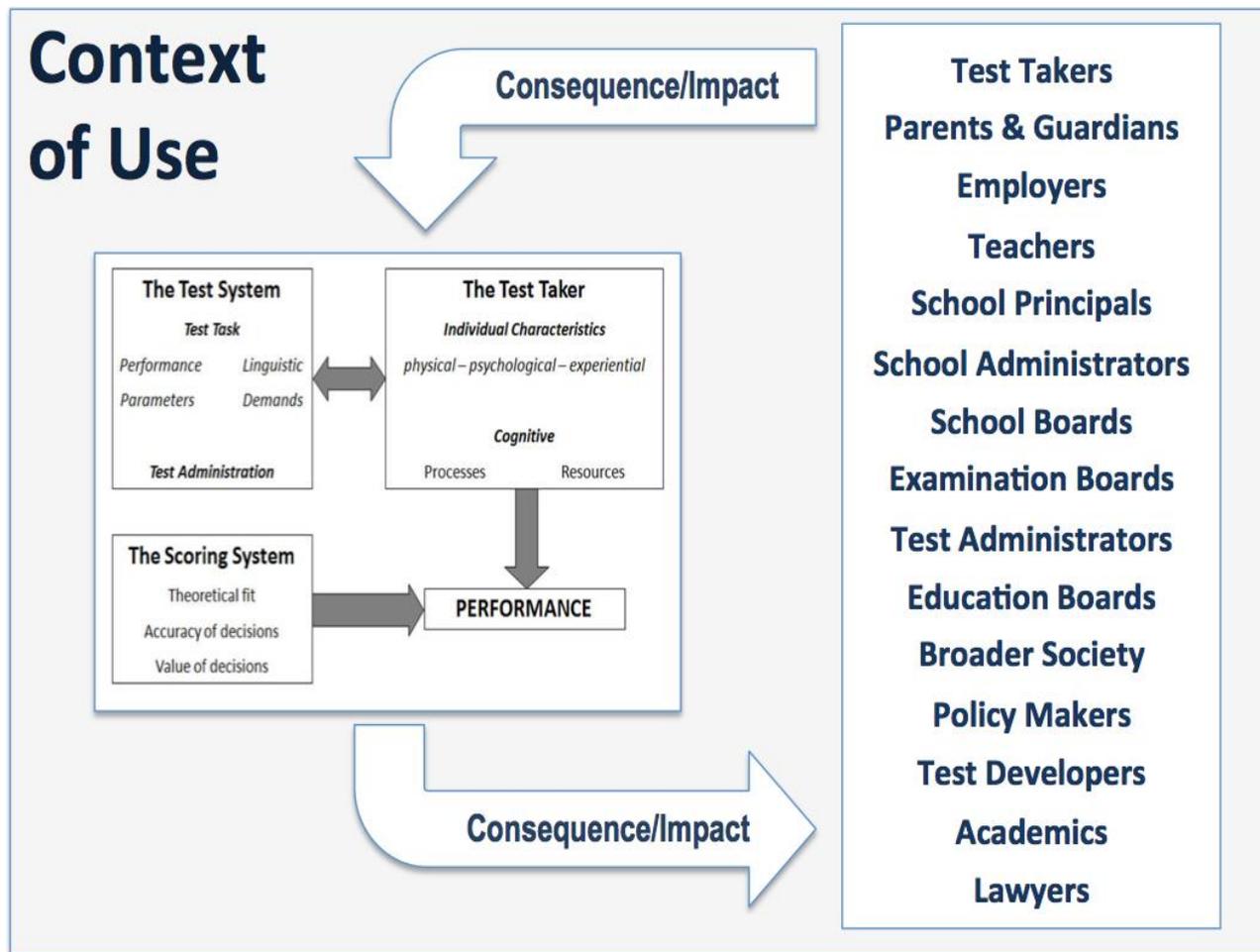**SHERYL COOKE, JUDITH FAIRBAIRN, JAMIE DUNLEA, KEVIN RUTHERFORD, RICHARD SPIBY**
LTF 24 NOVEMBER 2018

A TEST DEVELOPMENT MODEL FOR DEVELOPING A LOCALISED TEST

TEST DEVELOPMENT AND VALIDATION APPROACH

SOCIO-COGNITIVE MODEL (O'SULLIVAN 2014; IN PRESS)

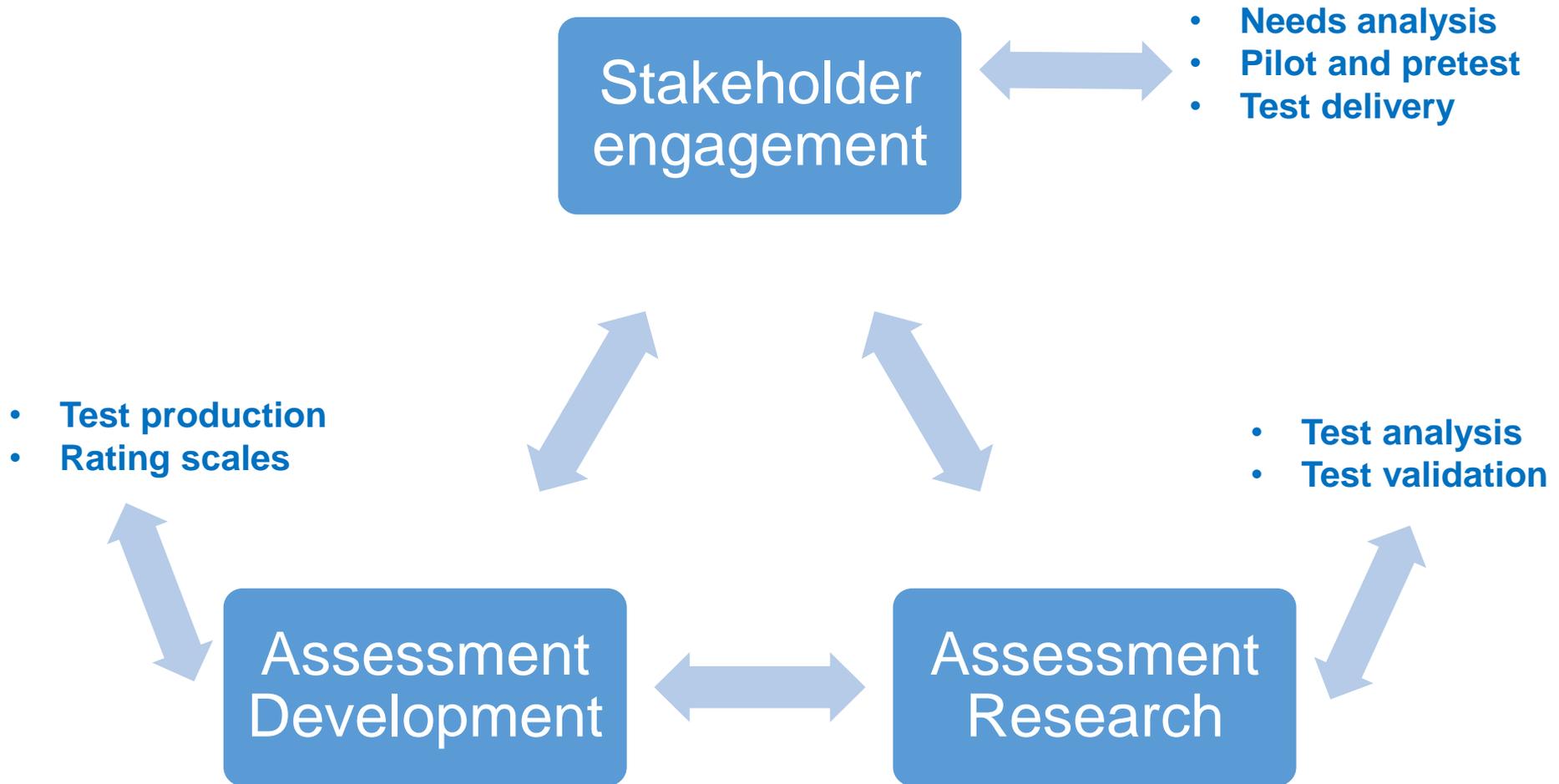# LOCALISATION

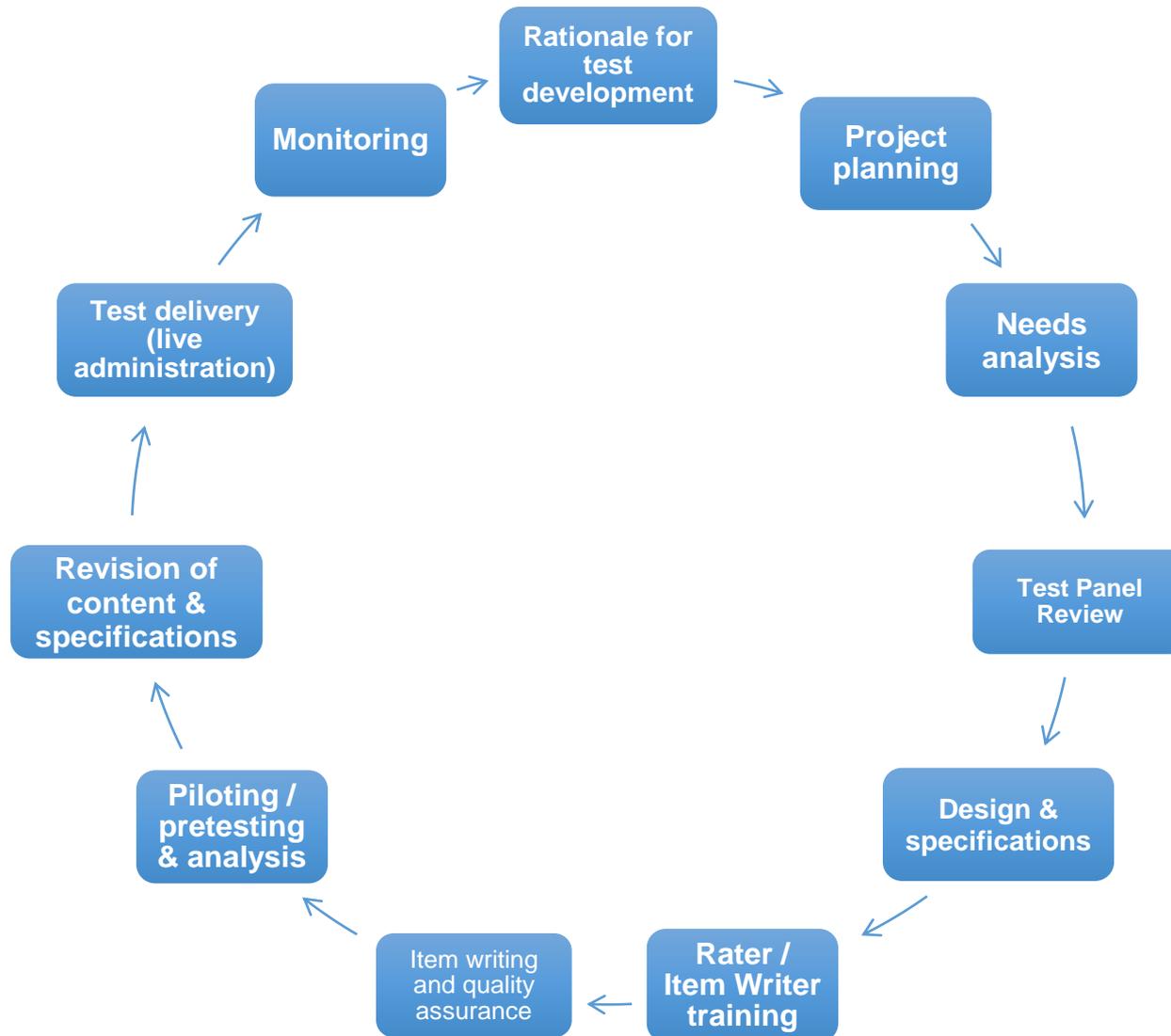Customising tests
for local needs

Locally

appropriate

test

Globally

Neutral

LOCALISATION

Level	Description	Examples
Level 0	Aptis General (or other existing variant) in a **full, four-skills** package	User selects a four-skills package of any Aptis (General or variant) available for use.
Level 1	Options for localisation are limited to selection from a **fixed range of pre-existing features**, such as delivery mode and/or components	User is able to select the skills to be tested and/or the mode of delivery that are appropriate. For example, the Reading package (Core component + Reading component) of Aptis General, taken as a pen-and-paper administration.
Level 2	Contextual localisation: **lexical, topical modification**	Development of specifications for generating items using existing task formats but with topics, vocabulary, etc. relevant for specific domains (e.g. Aptis for Teachers, projects in India, Rwanda, Saudia Arabia).
Level 3	**Structural reassembly**: changing the number of items, proficiency levels targeted, etc., **using item-bank content**.	Developing a test of reading targeted at a specific level, e.g. B1, using existing task types and items of known difficulty calibrated to the Aptis reading scale.
Level 4	Partial **re-definition of target construct** from existing variants. Will involve **developing different task types** to elicit different aspects of performance,	Developing new task types that are more relevant for a specific population of test-takers, while remaining within the overall framework of the Aptis test system (*e.g. Aptis for Teens*).
Level 5	The construct and/or other aspects of the test system are changed to such an extent that **the test will no longer be a variant within the system**.	For example, developing a matriculation test for uses within a formal secondary educational context; developing a certification test available to individuals rather than organisations, etc.

LOCALISATION

Stakeholder engagement

- **Needs analysis**
- **Pilot and pretest**
- **Test delivery**

- **Test production**
- **Rating scales**

Assessment Development

Assessment Research

- **Test analysis**
- **Test validation**

TEST DEVELOPMENT CYCLE

STAKEHOLDER ENGAGEMENT

STAKEHOLDER ENGAGEMENT & VALIDITY

Messick's 1989 unified approach to validity
The socio-cognitive model: test-taker at the core

Important to understand

- Target Language Use (TLU) domain the test-takers are being tested for

- test-taker profile

- context of test delivery

Stakeholders:

"the test designers, teachers, students, score users, governments or any other individual or group that has an interest in how the scores are used and whether they are useful for a given context" (2007, p. 14).

Fulcher & Davidson (2007)

THE CONTEXT

English Medium of Instruction (EMI) high school in China
International curriculum
Prestigious:

- *likely precursors to top international academic opportunities*
- *places to develop bilingual or near-bilingual proficiency in academic English*

3,000 applicants per year
Extremely high-performing students apply
10% acceptance rate
Battery of placement tests: English proficiency one of the key discriminating factors

Existing Aptis versions:
- *Aptis for Teens*
- *Aptis General*
- *Aptis Advanced*

Not fit-for-purpose:
- *Too easy*
- *Irrelevant topics*

STEPS IN STAKEHOLDER ENGAGEMENT

- Feedback and reflection

- In-depth discussion with management - Needs Analysis tool

  - decision-making criteria

  - decision-making process

  - test-administration requirements and constraints

  - teaching faculty

- Survey of teachers

- Analysis of curriculum, textbooks and examination papers

- Limited classroom observation

NEEDS ANALYSIS TOOL

Understanding the English Language Needs of Your Students - August 2017

* 3. Which of the following do your students need to listen to for their studies? (Both live and broadcast through media)

	Never (not needed)	Rarely	Sometimes
Classroom lectures and presentations	○	○	○
Descriptions and explanations of processes or systems	○	○	○
Debates and discussions	○	○	○	○
Announcements and instructions	○	○	○	○
Interviews and opinions	○	○	○	○
Extended narratives and performances	○	○	○	○
Fact-rich information in a format			

Which of the following do your students need to read in English for their studies? (Both paper-based and on-screen)

Which of the following do your students need to listen to in English for their studies? (Both live and broadcast through media)

Test Development Needs Analysis

Preliminary NA
Wednesday, 21 June 2017
Place
Conducted by: Sheryl Cooke

Content and background

Who marked the test?
What rating scale was used for writing assessment?
What was your experience with Aptis for Teens?
What was your experience with Aptis General?
What problem are you trying to solve?
What's the most important thing about the test for you?

B   Test-takers
How old are they?
Where do they come from
What's their educational b
Gender?
Is it likely that they've prep
Are they generally computer literate?
What curriculum do they follow before they arrive?
What about after they arrive at the school?
What English use does this age group typically engage in?
Do you have any test-takers with special needs?

Test-takers

C   Test use
What's the purpose of the te
What decisions will be made
How much weight does the E
decision?
Who sees the test results?
What consequences are ther
don't?
Are the test results used for any other purpose? e.g. placement in classes, diagnostic, etc.
Would it be useful if they could be?
Are practice materials required?

Test use

D   Construct
What ability are you trying to te
Which skills?
Why not speaking?
What are the children expected
If you did want information about their English ability, what information would be most useful?
Construct irrelevance - what not to test?

Construct

E   Test administration
How many test-takers?
How man
How long
How man
Why pap
What turn
How shou
What TRF information do you want?
How do you ensure test security?
Are you open to computer delivery of the test?
Do they take copies of the test home with them?

Test administration

RESULTS OF STAKEHOLDER ENGAGEMENT

Key requirements and challenges

- *Discriminate at the higher levels (B1-C2)*

- *Reflective of actual communicative demands in the classroom*

- *Reflective of textbook topics*

- *Timing during test administration*

Survey of teachers and analysis of textbooks & exams facilitated a more granular understanding of the TLU domain:

- *Communicative tasks students are required to engage in*

- *Topics and text genres they encounter*

- *Variety of Englishes they are exposed to*

Writing Needs

- Reflective journals or learning diary entries 4%
- Academic essays and arguments 11%
- Notes and minutes e.g. in response to aural, textual or visual (graphic) input 9%
- Reviews, e.g. article, book or literature reviews 9%
- Handouts/Powerpoints (for presentations reports, etc) 11%
- Summaries and reports, e.g. in response to aural, textual or visual (graphic) input 11%
- Produce posters (for presentation or display) 6%
- Creative writing 6%
- Formal emails (or letters), e.g. to organisations 5%
- Social media, e.g. blogs, wikis and journals 4%
- Informal emails (or letters), e.g. to friends 4%
- Complete forms 7%
- Articles for newspapers or magazines (e.g. school newspaper) 5%
- Proposals for school projects 5%
- Study or job applications with Resumes / CVs 3%

Task: Write an article for an online magazine using a set of graphically-presented research notes

Task: Write an email to a university professor about a topic related to possible future study.

13

ASSESSMENT DEVELOPMENT

LOCALISATION

Level	Description	Examples
Level 0	Aptis General (or other existing variant) in a **full, four-skills** package	User selects a four-skills package of any Aptis (General or variant) available for use.
Level 1	Options for localisation are limited to selection from a **fixed range of pre-existing features**, such as delivery mode and/or components	User is able to select the skills to be tested and/or the mode of delivery that are appropriate. For example, the Reading package (Core component + Reading component) of Aptis General, taken as a pen-and-paper administration.
Level 2	Contextual localisation: **lexical, topical modification**	Development of specifications for generating items using existing task formats but with topics, vocabulary, etc. relevant for specific domains (e.g. Aptis for Teachers, projects in India, Rwanda, Saudia Arabia).
Level 3	**Structural reassembly**: changing the number of items, proficiency levels targeted, etc., **using item-bank content**	Developing a test of reading targeted at a specific level, e.g. B1, using existing task types and items of known difficulty calibrated to the Aptis reading scale.
Level 4	Partial **re-definition of target construct** from existing variants. Will involve **developing different task types** to elicit different aspects of performance,	Developing new task types that are more relevant for a specific population of test-takers, while remaining within the overall framework of the Aptis test system (*e.g. Aptis for Teens*).
Level 5	The construct and/or other aspects of the test system are changed to such an extent that **the test will no longer be a variant within the system**.	For example, developing a matriculation test for uses within a formal secondary educational context; developing a certification test available to individuals rather than organisations, etc.

Test	Aptis Teens Advanced	Com...	...eaking	Task	Task 4

				C1	C2
Skill focus	Responding to qu...				
Task Level (CEFR)	A1				
Task description	The candidate ...they have just presented (either the poster task or the argum... the task is expected to simulate questions from ... or argumentative speech. The poster or ar...				
Instructions to candidates	...ut your [poster presentation or ...in speaking when you hear this				
Presentation of ...	...al non-verbal (e.g. photo)				
Response ...					
Planning ...					

Descri...					
Length of questions					
Lexical Level					
Content knowledge	Ge...				
Cultural specificity	Neut...				
Nature of information	Only			...stract	
Relevant Domain	Pu...			...sonal	
Topic	Referencing th...				

Description	Candidate must prov... ...question to achieve the task.				
Length of response	45 seconds for each resp... candidate to speak for the full ...			...e will generally require the	
Lexis /grammar	Demonstration of grammatical co... sufficient to respond to all questions.			...g of 5 (out of 6) for the task. C1 lexis	
Rating scale for task	A task-specific holistic rating scale is used... performance is required to achieve score ba...			...e is a 7-point scale from 0-6. A C1 level ...mances at a C2 level.	

Rating scale (overlay):

6 C2
- Responses to all three questions are clear and well-structured and show the following features
- Uses a range of complex grammar constructions accurately. Some very minor errors occur but do not impede understanding.
- Uses a range of vocabulary to discuss the topic required by the task. No awkward usage or inappropriate lexical choices.
- Pronunciation is clearly intelligible.
- Backtracking and reformulations do not interrupt the flow of speech.
- A range of cohesive devices are used to clearly indicate the links between ideas.

5 C1
- Responses to all three questions are on topic and show the following features
- Uses a range of complex grammar constructions accurately. Some minor errors occur but do not impede understanding.
- Uses a range of vocabulary to discuss the topics required by the task. Some awkward usage or slightly inappropriate lexical choices.
- Pronunciation is clearly intelligible.
- Backtracking and reformulations do not interrupt the flow of speech.
- A range of cohesive devices are used to clearly indicate the links between ideas.

4 B2.2
- Responses to all three questions are on topic and show the following features
- Some complex grammar constructions used accurately. Errors do not lead to misunderstanding.
- Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.
- Pronunciation is intelligible. Mispronunciations do not put a strain on the listener.
- Some pausing while searching for vocabulary but this does not put a strain on the listener.
- A limited number of cohesive devices are used to indicate the links between ideas.

3 B2.1
- Responses to two questions are on topic and show the following features
- Some complex grammar constructions used accurately. Errors do not lead to misunderstanding.
- Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not lead to misunderstanding.
- Pronunciation is intelligible. Mispronunciations do not put a strain on the listener.
- Some pausing while searching for vocabulary but this does not put a strain on the listener.
- A limited number of cohesive devices are used to indicate the links between ideas.

2 B1.2
- Responses to at least two questions are on topic and show the following features
- Control of simple grammatical structures. Errors occur when attempting complex structures.
- Limitations in vocabulary make it difficult to deal fully with the task.
- Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener.
- Noticeable pausing, false starts, reformulations and repetition. Links between ideas are not always clearly indicated.

1 B1.1
- Response to one question is on topic and shows the following features
- Control of simple grammatical structures. Errors occur when attempting complex structures.
- Limitations in vocabulary make it difficult to deal fully with the task.
- Pronunciation is intelligible but occasional mispronunciations put an occasional strain on the listener.
- Noticeable pausing, false starts, reformulations and repetition. Links between ideas are not always clearly indicated.

0 A1/A2
- Uses only simple cohesive devices.
- Performance not sufficient for B1, or the responses are completely off-topic (memorised or guessing).

PRINCIPLED COMPROMISES

- Time to develop tests – how long does it take to develop at test?

- Length of time for test sections – where can we reduce test taker time?

- Variety of Englishes

- Item writer training

- CBT / pen & paper format – logistics

SOCIOCOGNITIVE MODEL
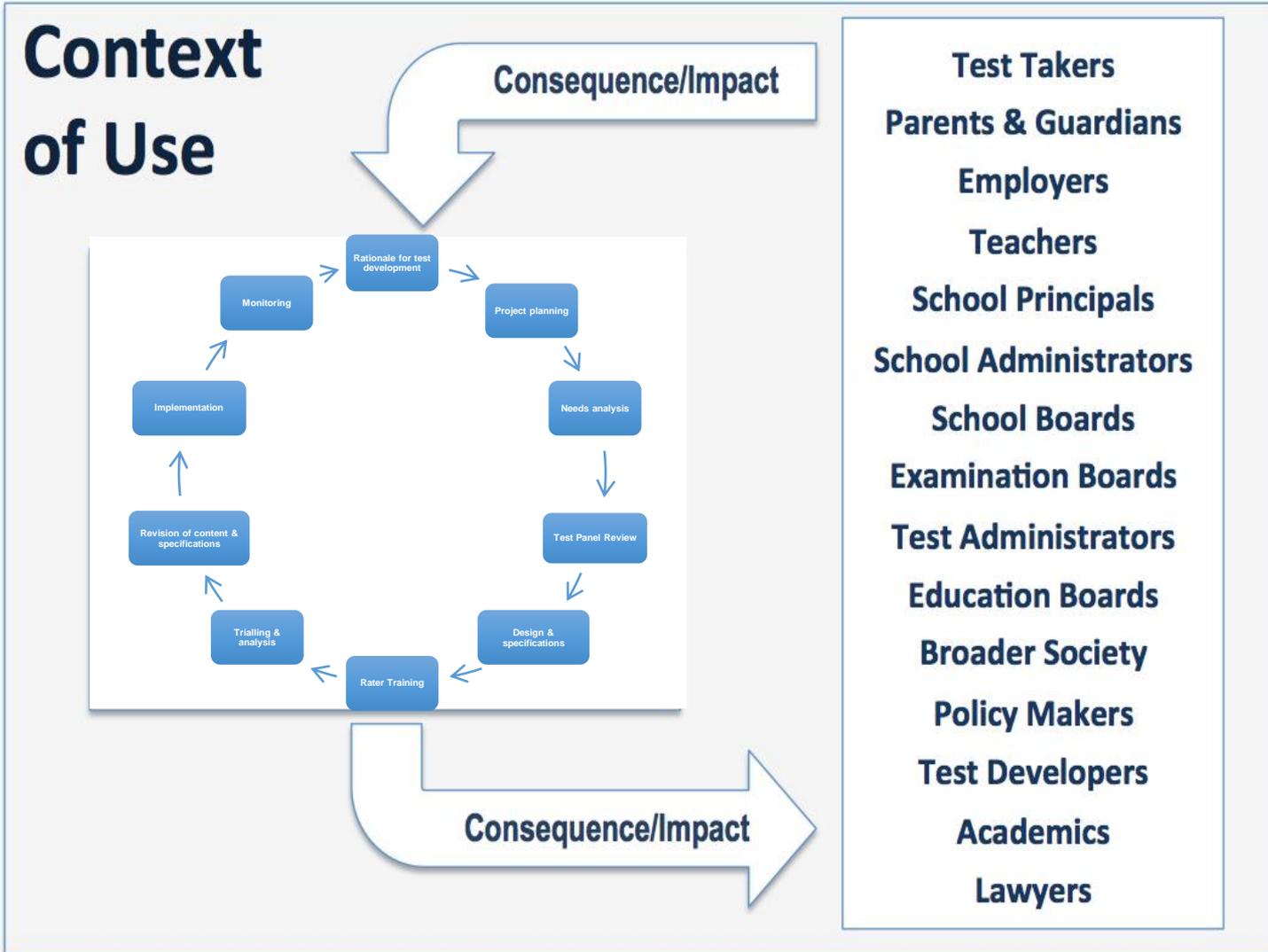
CONCLUSIONS

- Collaboration between teams

- Localisation and engagement with stakeholders

- Principled compromises

THANK YOU

Sheryl.Cooke@britishcouncil.org

Judith.Fairbairn@britishcouncil.org

Jamie.Dunlea@britishcouncil.org

Kevin.Rutherford@britishcouncil.org

Richard.Spiby@britishcouncil.org