Report 05 – Analysis of Competences Measured in TEA

TEST OF ENGLISH FOR AVIATION

Introduction

Fulcher & Davidson state that all testing is indirect since "from test performance we obtain a score, and from the score we draw inferences about the constructs the test is designed to measure" (2010: 64). This is a 'construct-centred approach' as described by Messick (1994: 17):

> "A construct-centred approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviours or behaviours should reveal those constructs, and what tasks or situations should elicit those behaviours? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring and rubrics."

In the development of TEA, the language competences intended to be measured according to the testing context (as described by ICAO) were identified and are defined conceptually in ***Report 01 - Description of Test Purpose, Specifications & Construction***. The following table lists those language competences and they are numbered here for reference purposes. The table shows at-a-glance which competences were intended to be engaged by each of the 6 tasks in TEA (a ● indicates a primary focus; a ○ indicates a secondary focus). (In ***Report 03 - Description of Tasks & Instructions***, the procedures and conditions under which performance is elicited and observed are specified.)

	Competence	Part 1 Interview	Part 2A	Part 2B	Part 2C	Part 3 Pictures	Part 3 Discussion
1	Talk about familiar, common, concrete and work-related topics specific to the candidate's role in aviation	●					○
2	Talk about familiar, common, concrete and work-related topics common to pilot – controller roles in general	●	●			●	●
3	Talk about aviation-related topics in a broader context			●	●	●	●

4	Use a range of basic and complex grammatical structures as appropriate to the function of the task	●	●	●	●	●	●
5	Use a range of work-related vocabulary	●	●	●	●	●	●
6	When lacking vocabulary, use circumlocution strategies	○	○	○	○	●	●
7	Produce connected stretches of language, sometimes at length	●				●	●
8	Use a range of phonological features (sound, stress, rhythm, and intonation) to produce speech intelligible to the international aviation community	●	●	●	●	●	●
9	Understand and recall the specific details of short messages delivered by both pilots and controllers in plain English in non-routine situations at different phases of flight (tower, ground, departure, en-route, approach)		●				
10	Understand a range of native and non-native speakers in terms of accent and rate of speech		●	●	●		
11	Process linguistic difficulties – tense	○	●	●	●		●

12	Process linguistic difficulties – modality		●	●	●	●
13	Process linguistic difficulties – lower frequency work-related vocabulary		●	●	●	●
14	Process linguistic difficulties – negation		●	●	●	
15	Process linguistic difficulties – contraction		●	●	●	
16	Recognise the illocutionary force (the communication purpose) of the speaker	●	●			●
17	Understand and respond to short messages describing linguistic or situational complications or an unexpected turn of events (in an aviation context).			●	●	
18	Where necessary, demonstrate discourse management strategies to resolve misunderstanding	●	●	●	●	●
19	Manage and maintain the speaker-listener relationship	●	○	●	●	●

20	Respond immediately, appropriately and informatively	•	•	•	•		•

In this report, an analysis of three transcribed tests aims to identify how well the competences are engaged and present an argument to support TEA's construct validity. Although many more tests would need to be analysed to present a full picture, by analysing three sample tests the reader can judge the methodology, data and conclusions for themselves (rather than be presented with unsupported data). The tests analysed here can be considered 'normal' samples and indicative of the majority of tests.

The concept of using test response analysis to validate tests is not new. Fulcher and Davidson (2010) describe the theory:

> "A test task is essentially a device that allows the language tester to collect evidence. The evidence is a response from the test-taker, whether this is a tick in a box or an extended contribution to a dialogue. The 'response as evidence' indicates that we are using the responses in order to make inferences about the ability of the test taker to use language in the domains and range of situations defined in the test specifications." (p62)

Methodology

In order to further investigate TEA Version 2010's test quality, the development team conducted an exercise to find out if TEA's test tasks engaged the language competences as intended and, if so, how effective the test tasks were in engaging those competences.

A group of 10 language professionals contributed with minimum requirement of a Diploma in English language teaching and a professional interest in *plain English for Aviation*. Within the group, higher qualifications included Masters qualifications in Language Testing and Linguistics.

The 10 judges were asked to listen to 3 tests (see ***Report 12 – Example Criteria in TEA Assessment: 3 Annotated Test Transcripts*** to view the transcripts) and rate how effectively each part of TEA engages each of the 20 competences listed above. They were also encouraged to make comments as necessary.

The 10 judges were asked to rate their answer to '*How well does the task in TEA engage the 20 language competences?*' on the following Likert scale:

0 = not at all *1 = not very well* *2 = fairly well* *3 = well* *4 = very well*

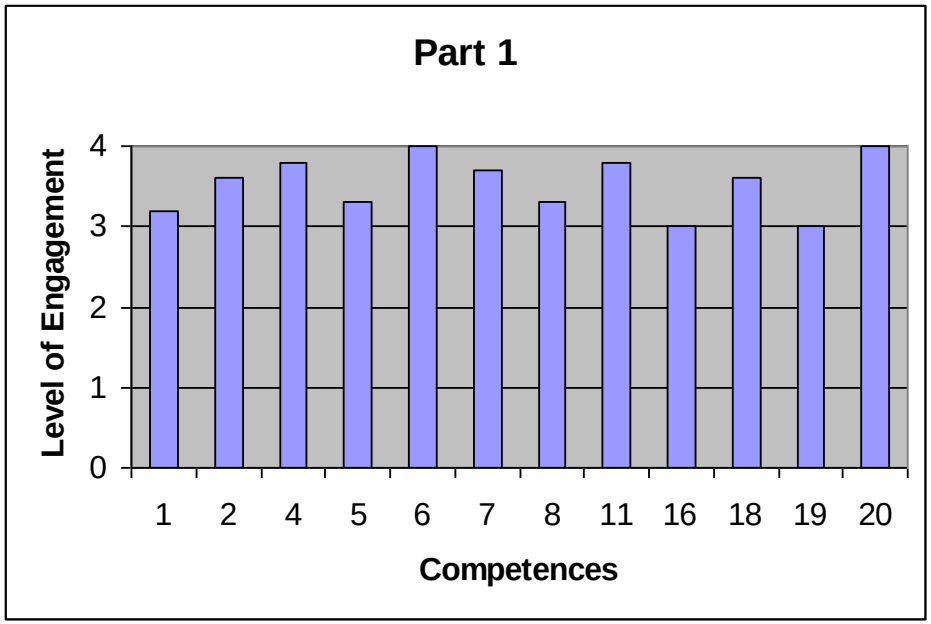
Their ratings and comments were collated to conduct a mixed methods approach to quantitative and qualitative analysis and try to get a fuller picture of test quality.

Data and Discussion

The three tests analysed were recordings of candidates at three distinct proficiency levels in the following order from highest to lowest level: Moroccan ATC (Level 6), Italian Pilot (Level 4), and Russian ATC (Level 3). (For detailed information about assessment of their performances, see ***Report 12 - Example Criteria in TEA Assessment: 3 Annotated Test Transcripts***.) Analysis of each task is presented in turn here along with pertinent comments.

Part 1 Interview

The chart below shows the mean scores of the 10 judges for competence engagement in Part 1:



Sample judge comments:

- *The questions clearly give the candidates the opportunity to talk at length about their duties using different tenses and vocabulary*
- *The lower-level candidates needed to use circumlocution strategies when they couldn't describe things precisely*
- *I think there could be more questions related precisely to the candidates' role in aviation in this part. They talked about pilot-controller roles generally rather than specifically.*
- *Can we say the test engages competences if the candidate doesn't have the competence in the first place? For example, the Russian ATC didn't understand some of the questions – so he wasn't able to 'recognise the illocutionary force of the examiner's questions'. I don't think that's the fault of the test questions, however.*

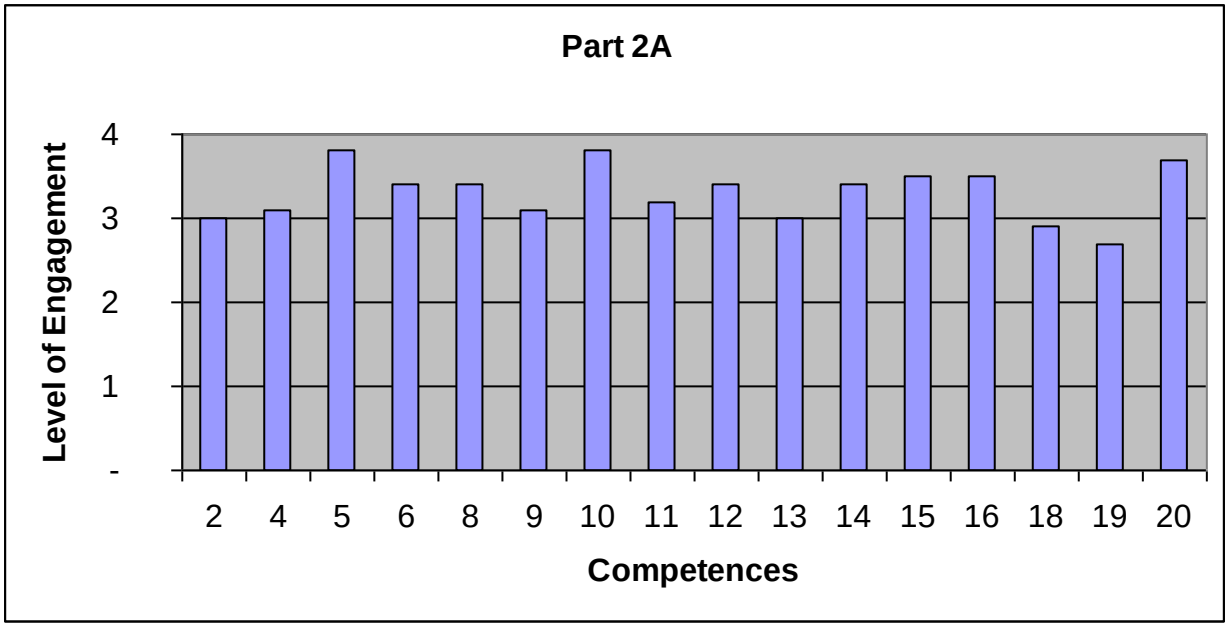
The judges agreed that all of the competences intended to be engaged in Part 1 are engaged 'well' to 'very well'. A few judges gave lower ratings. For competence 1 - *Talk about familiar, common, concrete and work-related topics specific to the candidate's role in aviation* – there were doubts as to whether questions elicited descriptions and explanations of the candidates' specific role, as opposed to their role in aviation generally. The test development team have considered this issue thoroughly during trialling of test items and reached the conclusion that specific questions about candidates' jobs typically elicit responses bound in operational (technical) language – undesirable for this context. For this reason, a broad view of a 'work-related' has been taken (see ***Report 01 – Description of Test Purpose, Specifications & Construction*** for further discussion of this topic).

Interesting comments were made about the lower-level candidate (the Russian ATC at Level 3) and whether the test could be said to be engaging competences that are not present within the candidate himself. Since the judges agreed that Part 1 engages competence 16 - *Recognise the illocutionary force (the communication purpose) of the speaker* – in the higher-level candidates, it seems fair to suggest that deciding if a task engages the necessary competences by observing a candidate who demonstrates a lack of ability is illogical. By contrast, the Italian Pilot (at Level 4) is clearly able to recognise the purpose of the examiner's questions in Part 1.

Regarding competence 18 - *Where necessary, demonstrate discourse management strategies to resolve misunderstanding* – it would seem that the engagement of this competence is more likely to be observable in lower-level candidates since they are more likely to need to seek clarification. A few judges suggested that Part 1 may be simple for higher-level candidates, thus negating the opportunity to engage the ability to resolve misunderstanding. However, it is good testing practice begin with a task that eases candidates into the test before the linguistic challenges increase further into the test. Furthermore, it is still important that the test engages those competences which may be viewed as 'simpler' in order to assess candidates against the ICAO Descriptors.

Part 2A

The chart below shows the mean scores of the 10 judges for competence engagement in Part 2A:



Sample judge comments:

- *The candidates are exposed to a range of vocabulary, structures and speech types. Most of the competences are engaged well.*
- *The recordings certainly test the 3 candidates' ability to recall. The Russian ATC doesn't understand them all – but can we be sure that the higher-level candidates 'understand' the recordings through simple recall?*
- *The candidates didn't have the chance to truly 'resolve misunderstanding' – but they do have to show they can ask for something to be repeated or to express what it is they don't really understand.*
- *Candidates are exposed to a range of vocabulary – through different situations – but are 10 recordings enough to explore a 'range' adequately.*

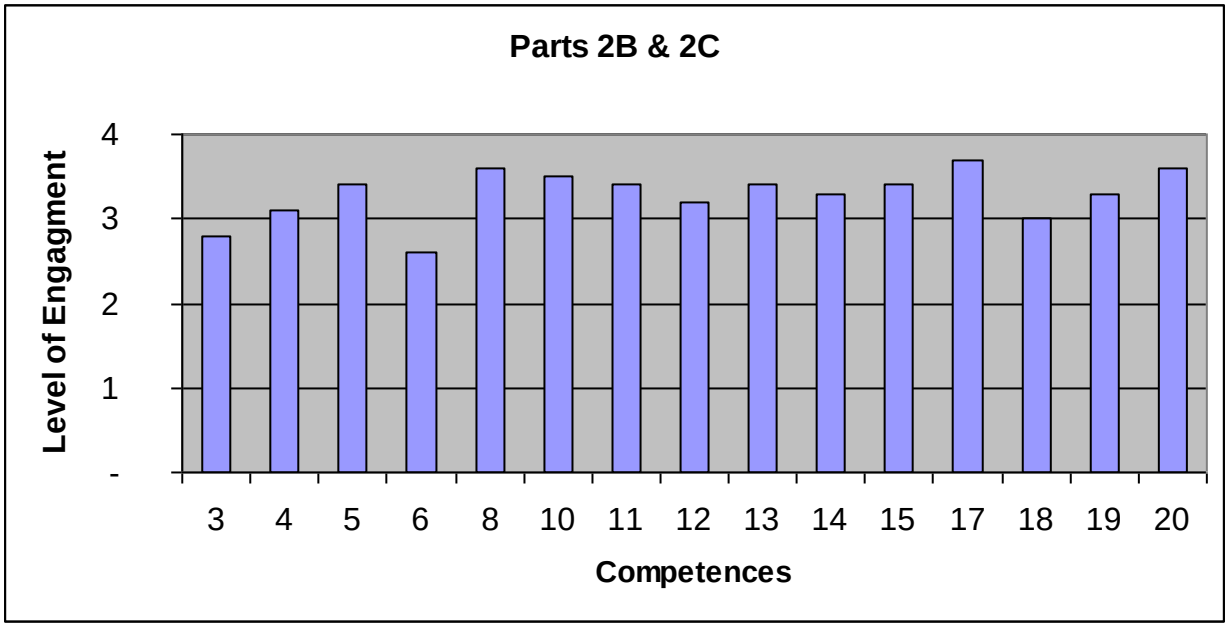
The judges agreed that all of the competences intended to be engaged in Part 2A are engaged 'well' to 'very well', apart from competences 18 and 19 – those which refer to interactive ability – which were rated, on average, as being engaged 'fairly well'. This is predictable since 2A is a semi-direct delivered listening section and, although candidates do show their ability to seek clarification, express and resolve misunderstanding, the nature of assessment in 2A decrees that interaction with the interlocutor is inappropriate. Although the development team have given thought to making 2A more interactive and 'two-way', the obvious difficulties in training interlocutors to behave consistently and not aid candidates in what is essentially a 'listening test' makes the concept unmanageable. The threat to reliability of interaction between candidate and interlocutor during Part 2 is too great to consider.

The matter of 'comprehension' is an interesting one. Since the lower-level candidate clearly has problems demonstrating understanding of the 6 of the 10 recordings, comprehension clearly plays a part in his (in)ability to recall them. Academic research has linked the ability to recall meaningful linguistic chunks with language proficiency, suggesting that the stronger performance in Part 2A of the Moroccan ATC and the Italian pilot is linked to comprehension ability rather than a simple ability to repeat a recording parrot-fashion through working memory. Academic research into short-term working memory and the 'phonological loop' – described by Baddeley (2000) "a temporary phonological store in which auditory memory traces decay over a period of a few seconds, unless revived by articulatory rehearsal" (p419) – has revealed that recall is complicated by lack of comprehension. Baddeley reports that "[i]f asked to recall a sequence of unrelated words, subjects typically begin to make errors once the number of words exceeds five or six. However, if the words comprise a meaningful sentence, then a span of 16 or more is possible" (p419). This suggests that if a test candidate has little or no understanding of a longer recording, he will not be able to accurately repeat what he has heard.

Regarding the potential lack of evidence garnered from 10 items, the development team trialled 20 recordings in Part 2A to expose candidates to a bigger range of situations, structures, and speech types but the resulting data suggested that the longer task did not differentiate between candidates any differently from the 10-recording task and the general candidate feedback was that a shorter form was preferable. See ***Report 03 - Description of Tasks & Instructions*** for further information.

Parts 2B & 2C

The chart below shows the mean scores of the 10 judges for competence engagement in Parts 2B and 2C (analysed jointly since the tasks intend to engage the same competences):



Sample judge comments:

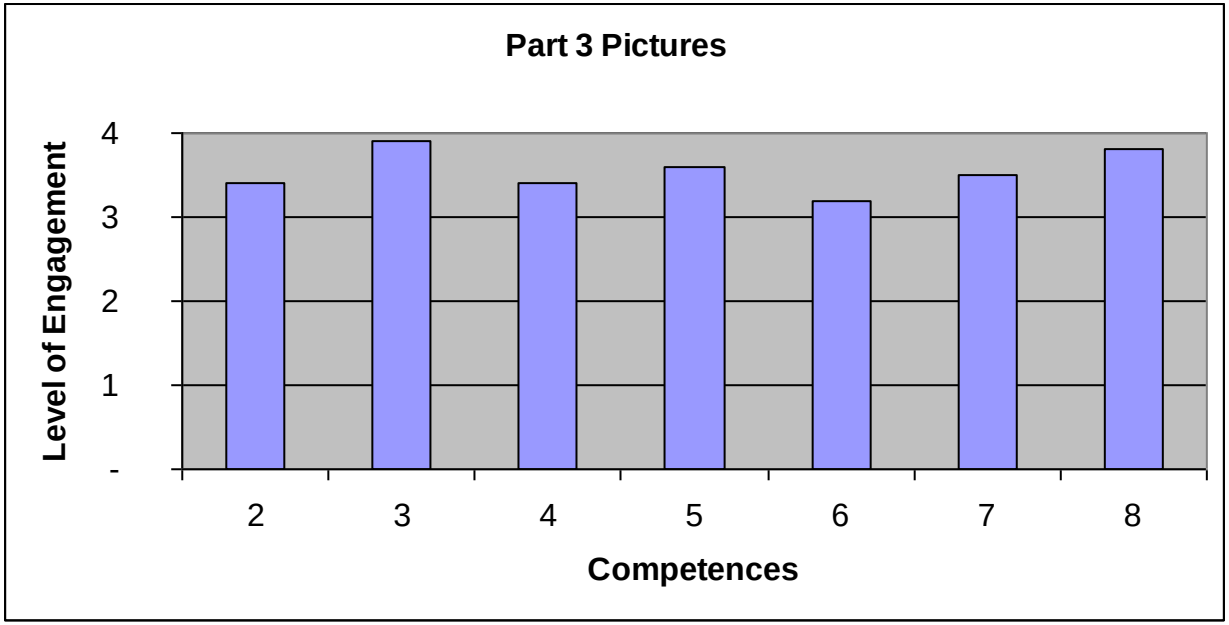
- *Responding to the tasks appropriately and informatively demands comprehension of the short recordings – and there is clearly a range of difficulty in this case.*
- *The lower-level candidates certainly seem to have more difficulty processing understanding of the recordings in these tasks, suggesting the need to utilise a variety of listening sub-skills as described by competences 10 to 15.*
- *The range of structures elicited suggests that competence 4 is not engaged particularly well, but the functional target of the 2 tasks – i.e. asking questions or giving advice - limit that to a strong extent.*

- *More recordings would mean we could say that the listening competences were being better engaged – but what length should a test be?*

The judges agreed that all of the competences intended to be engaged in Parts 2B and 2C are engaged 'well' to 'very well', apart from competence 3 (*Talk about aviation-related topics in a broader context*) and competence 6 (*When lacking vocabulary, use circumlocution strategies*) – which were rated, on average, as being engaged 'fairly well'. Crucially, competence 17 – *Understand and respond to short messages describing linguistic or situational complications or an unexpected turn of events (in an aviation context)* – was rated highly.

Part 3 Pictures

The chart below shows the mean scores of the 10 judges for competence engagement in Part 3, the pictures task:



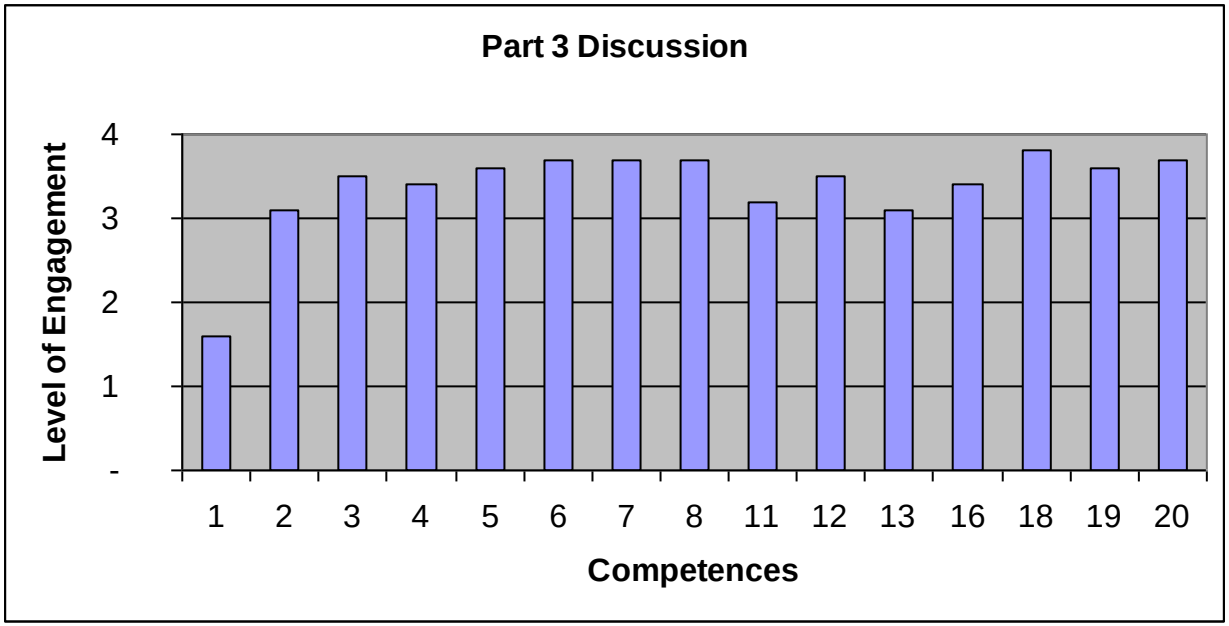
Sample judge comments:

- *From the range of structures elicited from the Moroccan ATC in this task I can say that it engages competence 4 well. The lower-level candidate doesn't respond in such a varied fashion but, clearly, that is not the fault of the task.*
- *Assessing the engagement of a broad range of vocabulary, and competence in paraphrasing, is perhaps as dependent on the pictures as the candidate's level – i.e. that they depict enough of the common and the obscure to elicit 'a range' and the potential need to paraphrase. The Moroccan ATC produced responses of between 120 and 140 words per picture, four times as many as the Russian ATC's first response but still had some trouble finding the vocabulary he wanted to describe the pictures precisely.*
- *Some connections between the pictures and within the pictures are made in all six candidate descriptions in this task.*
- *If more time were allowed, would the potentially bigger elicitation be valuable?*

The judges agreed that all of the competences intended to be engaged in the picture task are engaged 'well' to 'very well'. It was interesting to consider the effect of extending the allowed response time. Trials demonstrated that candidates spoke more slowly and with fewer mistakes but that the range of vocabulary elicited was only marginally extended. This suggested that a time limit of 1 minute was appropriate for assessing candidates under test conditions while still engaging the desired competences.

Part 3 Discussion

The chart below shows the mean scores of the 10 judges for competence engagement in Part 3, the discussion task:



Sample judge comments:

- *The discussion with the Italian pilot is a clear example of a two-way human interaction – the candidate is responding to the examiner and also has the freedom to clarify what is required of him.*
- *The topics are certainly broader than other parts of the test, vocabulary range is engaged fully.*
- *The range of question types forces tests the candidate's comprehension skills. Not many variations in tense here though…*
- *There are a lot of different structures elicited in the discussions with the higher-level candidates.*
- *No role-related discussion at all.*

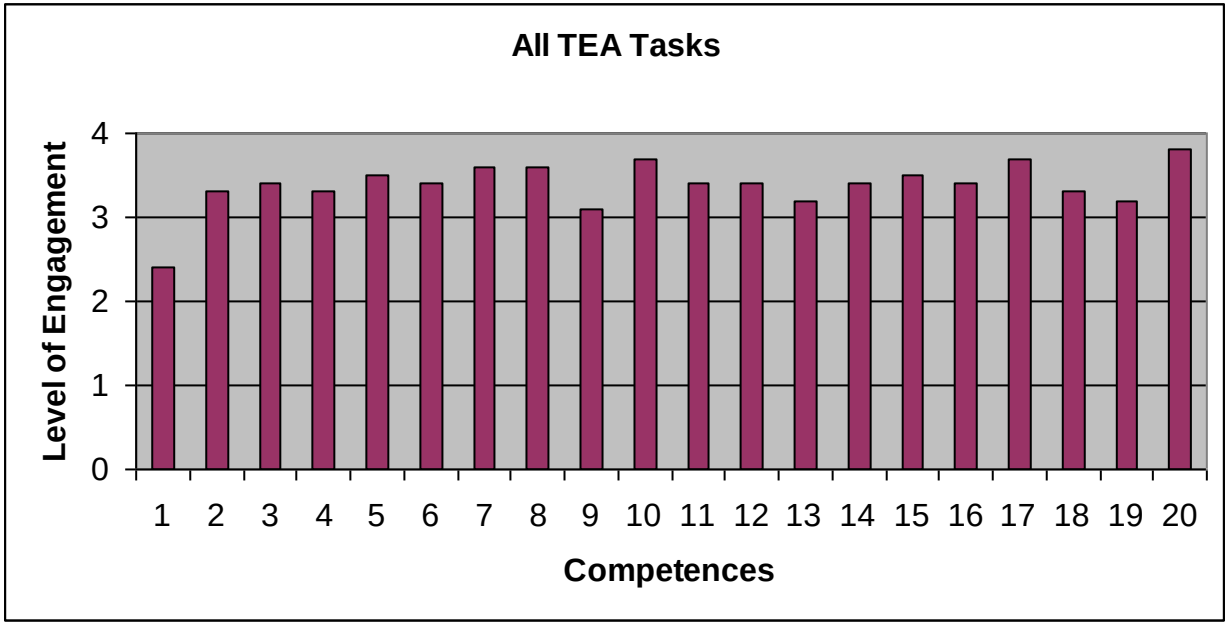
The judges agreed that all of the competences intended to be engaged in the discussion task are engaged 'well' to 'very well', apart from competence 1 – *Talk about familiar, common, concrete and work-related topics specific to the candidate's role in aviation* – which was rated, on average, as being

engaged ‘not very well’. Competence 1 is not a primary focus in this task although candidates talk about their own experiences and/or role sometimes. In different TEA question sets, the questions may focus more on ‘the role of controllers (or pilots)’ in a specific situation, thereby engaging their competence in describing role specific matters.

Conclusions

Although a larger data set would be required to argue for the construct validity of TEA, the three transcribed tests provide enough data to observe the effect of the tasks in TEA and the competences they engage. Clearly, the level of the candidate is a factor in judging competence-engagement but the range of abilities employed by the Moroccan ATC (at Level 6) at least shows that the test tasks allow proficient candidates to demonstrate their abilities. At the lower-levels, and within the bounds of test reliability, the test tasks allow for candidates to demonstrate their ability to resolve miscommunication, paraphrase when lacking vocabulary and manage the interaction.

The chart below displays the judge’s mean scores of competence engagement across the whole test, including all tasks:



The judges agreed that all of the competences intended to be engaged in the test are engaged 'well' to 'very well', apart from competence 1 – *Talk about familiar, common, concrete and work-related topics specific to the candidate's role in aviation* – which was rated, on average, as being engaged 'fairly well'. As discussed above, competence 1 is not a primary focus beyond Part 1 and, furthermore, the test development team have found through extensive trialling that attempts to engage candidates in specific role tasks results in the elicitation of operational or technical language that is un-assessable by the ICAO Descriptors and therefore unsuitable for this testing context. Hence, TEA's 'broad' view of 'work-related' (see ***Report 02 – Overview of Expert Judgements & Action Taken in Test Development*** for further discussion).

Threats to construct validity come in the form of construct under-representation and construct relevance. It can be seen from this study that key linguistic competences – grammatical, lexical and phonological – are engaged in every task. The relevance of the competences, or constructs, engaged in TEA is argued for in ***Report 1 - Description of Test Purpose, Specifications & Construction***.

References

Baddeley, A. (2000) *The episodic buffer: a new component of working memory?* Trends in Cognitive Sciences 4. 11: 417-423.

Fulcher, G. & F. Davidson. (2010). *Language Testing and Assessment*. Routledge: New York

Messick, S. (1994). *The interplay of evidence and consequences in the validation of performance assessments*. Educational Researcher 10, 9-20.