Testing
Robert A. Buckmaster

In this article I will look at testing. I will critique the existing testing paradigm and suggest that we should focus on Integrated Scenario-based Competency Tests for our testing needs.

I will consider:

1. Current Testing Options
2. Quality of Information
3. Trade-offs: Test Costs and Benefits
4. Playing to our Strengths
5. Integrated Scenario-based Competency Tests
6. Conclusions

1 Current Testing Options

At the moment we use testing at different stages of our courses:

1. Placement Testing: a quick test to ensure our learners are placed at an appropriate level. These are most commonly a multiple-choice test with an optional spoken and written component.
2. Progress Testing: tests taken after a period of work to see if the learners have mastered the material. These tests commonly reflect similar task types to coursebook and practice/workbook task types.
3. Diagnostic Testing: Tests done to find out if the learners have problems with certain items. Placement tests and Progress tests are also kinds of diagnostic tests.
4. End of Course Testing: A progress test writ large.
5. Proficiency Testing: A test at the level the learners are supposed to be at but not based on the content of the course, but on representative content the learners should be able to cope with.

In-House vs External Tests

In-house tests are mostly of the placement and progress kind. An institution might assemble a multiple-choice placement test, or rely on coursebook progress tests, or assemble a final proficiency test from published practice tests materials from external test providers like Cambridge or IELTS. These have the advantage of being ready-made and subject to some quality control.

External tests can be bought in from organisations like Cambridge, Pearson and IELTS. . Proficiency tests like Cambridge Advanced and IELTS, and Pearson tests suffer from weaknesses which are systemic in testing and make them less than ideal from our perspective. These are:

- An over-reliance on discrete item testing e.g. multiple choice items. In Reading, Listening and Use of English tests multiple choice tasks are by far the most common tasks. These have the advantage that the rating is 'objective' (if the question and options are properly written), and they are easy to score. However, they have weaknesses: the candidate has to read and understand the question and options; only selected parts of the texts are tested; candidates can guess answers and have 33% or 25% chance of being correct, or even 50% by eliminating distractors, or they have a straight 50% chance on True or False questions.

- Other tasks chosen include gap-fills. While these have the advantage of being a '*natural*' task (one where it is clear what the test taker has to do; and little or no training is required), gap-fills only test discrete items throughout a text (whether reading or listening) and are focused on checking whether the candidate understands that particular item. Comprehension beyond the gap can only be inferred.
- A neglect of integrated skills, though Pearson is making moves in this direction. Testers often attempt to test skills (reading, listening, writing and speaking) separately even though this does not reflect the real-life integrated nature of language in action. If we read a text for professional purposes, we are likely to take notes and then use these notes to prepare speech or a piece of writing. Even if we read for just pleasure then we do not answer comprehension questions based on it, so the test task itself is inauthentic.
- A requirement to infer the learners' wider ability from the tested sample of the language. The proficiency test provides samples of language, and the test takers are tested on these. On the basis of this performance, their wider ability is inferred as being C1, for example. If they can write one kind of formal letter at C1 it is inferred that they can write others. While this *may* be true, to a certain extent, it does not tell us whether they can write a report at C1, or an academic article. There is a limit to how much we can infer from the samples provided by these tests. A properly designed ESP test would give us much more confidence in our inferences than these GE tests.

An institution could decide to write their own specific tests, but this is a large undertaking and for anything more than informal progress test type situations needs to be taken seriously. This means taking a proper project approach to the task, with training and quality control measures. The testing team would need to be trained in test writing, and the tests produced would need to be moderated and pre-tested. Raters of spoken and written tests would need to be trained and standardised and monitored. Teachers could be trained to do this as part of their duties, but it is time consuming and proper allowance for this would have to be made. Teachers should not just be loaded up with testing duties on top of their existing official teaching duties.

2 Quality of Information

We test to get information. The key issue is the quality and quantity of information.

Multiple-choice tests give us specific information about discrete items. The information is very focused on that particular item. The learner might get that question correct, using, say, the present simple correctly. They might then misuse *another* example of the present simple. How many examples do we need to have to say with confidence that they can use the present simple? Information from multiple-choice questions is highly context specific and difficult to generalise from. Ideally, we would need a lot of such highly specific data in order to safely infer anything from such questions. A multiple-choice question gives us information about *that* multiple choice question and extremely limited information beyond that specific question. Any inferences are dangerous because the information *quality* is limited.

For Placement testing, a test of 60 – 100 multiple-choice items makes sense because we are looking for broad precision. We can always adjust our finding and move a learner up or down a level after a couple of lessons, though moving up a level is always psychologically easier than moving down, so, if in doubt, place lower. For anything requiring more precision we would need at least 100 such questions *at that level* in order to probe their abilities; not 100 questions covering *all* levels.

Face Validity

Multiple-choice item tests have *face validity* because test-takers are used to seeing such tests. They look like tests are supposed to look like. So, they are accepted as valid tests. However, this is just an argument for tradition. It would be relatively easy to set up a similar norm for testing, say by testing oral skills while hopping on one leg. After some time, it would become the accepted norm to test oral skills in this way, and other tests would not have the face validity of the 'speaking-test-while-hopping-on-one-leg' tests.

Face validity is the weakest form of test validity.

An essay writing task on the other hand, for example, gives us a huge amount of information about a lot of language points. A well-written, well structured, grammatically correct, lexically varied essay tells us a great deal about a learner's ability to write essays. While we cannot say with full confidence that they can write a good essay on *another* topic, we can be reasonably confident that they will make a good effort. We have more and better information - and our inferences about their wider ability can be made with more confidence.

Similarly, an oral interview provides us with a wealth of data about their speaking abilities *within* an interview format. While we cannot say with full confidence that they could make a good presentation, for instance, we can say that a good speaker in an interview *could probably* make a reasonable presentation. Our inferences are safer because we have more and better data.

The downside of these ways of testing like writing tasks and speaking tasks are the difficulties in rating. Everything is a trade-off, and we will examine these now.

3 Trade-offs: Test Costs and Benefits

As with everything, testing is a series of trade-offs. These include the costs and benefits of particular tests, which should be measured against each other.

Costs and benefits can be measured in three particular areas: in test preparation, test administration and test rating. These should be judged against the global outcome of information obtained from the test task – measured both in terms of quantity and quality. The final factor to consider is the backwash or washback effects of the text type; this can be positive or negative.

Let us examine these concepts through five common task types and see how they differ.

1. Multiple-choice items
2. Cloze test with multiple-choice options
3. Cloze test without multiple-choice options
4. Oral interview task
5. Essay writing task

1 Single Sentence Stem Multiple-choice Items

	Test Preparation	Test Administration	Test Rating
Costs	Relatively high cost; difficult to write good items; need to trial and revise; item writers should be trained in writing such tests.	Test security required; costs of printing tests.	Minimal cost of time to rate using provided key. The multiple-choice nature of test has a rating cost – the candidates can guess and this affects the quality of the information we obtain from the test – the information is not as reliable.
Benefits	Test items can be kept in a test item bank and reused, given good security.	Easy to administer; straightforward; most candidates understand task.	Easy to rate and check rating.

Information obtained from the task: high quality specific information about individual task items (and nothing else). But very low *quantity* of information – one question, one piece of information. Inferences about wider language ability can be made but with low confidence.

Washback: Negative; given a final test with multiple-choice items, a rational decision by teachers and students would be to practice such tasks. This is an opportunity cost – time taken to practice such tasks is time not spent on other, more productive learning tasks.

2 Cloze Test with Multiple-choice Options

	Test Preparation	Test Administration	Test Rating
Costs	Text sourcing costs (time to find text etc); cost of preparing multiple-choice items; if text adapted/edited then costs increase; item writers need to be trained in writing such tests; tasks should be moderated and pre-tested.	Test security required; costs of printing tests.	Minimal cost of time to rate using provided key. The multiple-choice nature of test has a rating cost – the candidates can guess and this affects the quality of the information we obtain from the test – the information is not as reliable.
Benefits	Single text provides a number of items; no need to think of item stems. Test items can be kept in a test item bank and reused, given good security.	Easy to administer; straightforward; most candidates understand task.	Easy to rate and check rating.

Information obtained from the task: this is comparable to single sentence stem multiple-choice item tasks but multiplied because there are a number of items per text. The quantity increases but the quality remains – the information we get is focused on the particular items tested. This kind of task tells us little beyond the items tested, despite the contextualisation. Inferences about wider language ability can be made but with low confidence

Washback: Negative; given a final test with a cloze + multiple-choice items task, a rational decision by teachers and students would be to practice such tasks. Again, this is an opportunity cost – time taken to practice such tasks is time not spent on other, more productive tasks.

3 Cloze Test without Multiple-choice Options

	Test Preparation	**Test Administration**	**Test Rating**
Costs	Text sourcing costs (time to find text etc); Low cost of preparing tasks if n-word item used; cost of preparing alternative item key; if text adapted/edited then costs increase; tasks should be moderated and pre-tested.	Test security required; costs of printing tests.	Minimal cost of time to rate using provided key.
Benefits	Easiest contextualised test task to prepare. Single text provides a number of items; no need to think of item stems. No special training required to prepare this test. Task can be kept in a test item bank and reused, given good security.	Easy to administer; straightforward; most candidates understand task.	Easy to rate and check rating.

Information obtained from the task: As this kind of text omits the multiple-choice items the *quality* of information in total is greater and more reliable than in Tasks 1 and 2. The candidate demonstrates the ability to recall an appropriate word from their memory and apply it to the gap provided. This suggests that they might be able to use that particular word appropriately in other gaps in other contexts. A cloze test gap-fill is a better task than single sentence multiple-choice items or cloze tests with multiple-choice items. However, the *quantity* of information is still limited to the number of items tested, and so is comparable with Task 2. Inferences about wider language ability can be made but with low confidence.

Washback: Negative; given a final test with a cloze task, a rational decision by teachers and students would be to practice such tasks. Yet again, this is an opportunity cost – time taken to practice such tasks is time not spent on other, more productive tasks.

4 Oral Interview Task

	Test Preparation	**Test Administration**	**Test Rating**
Costs	Test task writers need training in such tasks, though less than for 1 and 2; tasks should (ideally) be moderated and pre-tested.	Test security required; costs of printing tests. Test interlocutor(s) needs to be trained and standardised in test delivery. Monitoring of test delivery to ensure fairness etc. Test security required; costs of printing tests. Costs of recording and storage of recordings etc. if done; much more difficult to administer than pen and paper test tasks like 1, 2 and 3.	Test rater needs to be trained and standardised, and ratings need to be monitored.

Benefits	Less time consuming than previous tasks to prepare.	Speaking tests are shorter than standard written tests as the information content of a 10-20 minute spoken test is more than a pen and paper written test lasting a few hours.	Rating can be done quickly – at very end of test; even less time consuming than marking tasks 1, 2 and 3. Recorded interviews can provide a record of test-taker performance.

Information obtained from the task: A huge amount – about grammar, vocabulary, pronunciation etc.

Washback: Potentially mostly positive; such a test type emphasises productive skills speaking, and if the interview tasks are useful then the washback is positive. If the interview tasks are not relevant to the learners' needs, then the task declines in value and the washback becomes less positive.; however, speaking skills for interviews are transferrable skills so some utility remains.

5 Essay Writing Task

	Test Preparation	**Test Administration**	**Test Rating**
Costs	Test task writers need training in such tasks, though less than for 1 and 2; tasks should (ideally) be moderated and pre-tested.	Test security required; costs of printing tests.	Test rater needs to be trained and standardised, and ratings need to be monitored.
Benefits	Less time consuming than all previous tasks to prepare.	?	The completed scripts provide a record of test-taker performance.

Information obtained from the task: An enormous amount, easily accessible as being on paper (or electronic).

Washback: Potentially mostly positive; such a test type emphasises productive written speaking, and if the essay tasks are useful then the washback is positive. If the essay tasks are not relevant to the learners' needs, then the task declines in value and the washback becomes less positive. An ability to write good essays should probably positively correlate with the ability to write other tasks.

It seems clear that oral interview and essay writing tasks produce a huge amount of information about learner performance, with positive washback. There are costs – mainly in rater and interlocutor training and ongoing standardisation and monitoring of examiners but these are balanced by lower preparation costs. The other kinds of tasks have higher preparation costs and lower test administration and rating costs but provide much less information and much lower quality of information, as well as having seriously negative backwash effects.

Indeed, multiple-choice and gap-fill tasks tell us very little beyond giving us information on the learners' ability to do such tasks. They tell us nothing about the learners' ability to use the language, to actually do things with the language, except do multiple-choice and gap-fill tasks.

Direct testing of productive skills, on the other hand, gives us real insights into the test-takers' abilities to do such tasks. Given the quantity and quality of information obtained, any inferences about wider language ability can also be made with more confidence than with the other test types. These, then, are the kinds of test types which we should be focusing on in our testing. We need the best kind of information we can obtain, and as much as we can reasonably obtain.

Multiple-choice items give us too narrowly-focused a view on the test-takers' language ability. The trade-offs here are too great. Direct testing of productive skills involves the fewest trade-offs between costs and benefits when compared to the quantity and quality of information we obtain from such tests.

4 Playing to our Strengths

In internationally recognised testing systems like IELTS or Cambridge or Pearson tests the main problems they have are test security and scale. Their tests need to be scalable to tens thousands of candidates worldwide - often hundreds in one day in one location. The reading, listening and use of English test components (if there is such a paper) are designed to cope with scale through the task types – multiple-choice questions and gap-fills of various kinds. If there is a big enough room with enough desks, a relatively small number of invigilators and markers can process a lot of papers. The scale problems really come with rating the writing papers and conducting the oral interviews. This is why the IELTS writing rating is now done online and IELTS are offering computer-delivered speaking tests (from January 2022). Standardised, objective tests using multiple-choice and gap-fill questions are an *industrial response* to the testing problem. And with Pearson's attempts at computer-based rating of writing papers the response is becoming even more industrial.

If your institution has thousands of candidates per year, then it makes sense to adopt an industrialised response. You should appoint a Testing Team and task them with producing good industrial tests. Just remember that you need a lot of multiple-choice questions in order to say something useful about a person's ability at a particular level. They give you precise but limited information about one particular language point.

If you do not have such large numbers of candidates, it makes more sense to adopt a different *craft-based approach* to assessment which makes use of existing teacher skills. Instead of training teachers to produce good multiple-choice questions and having to go through all the steps this involves (drafting the item, moderating the item, pre-testing the item and so on), which is very time-consuming, build tasks so the teachers can use their existing skills of evaluating written and spoken English. All teachers have skills at evaluating their learners' output and giving feedback on it; it's all part of teaching.

5 Integrated Scenario-based Competency Tests

If we want to be able to say with confidence that this candidate can do X, Y and Z, then, ideally, we need to be able to show directly that the candidate can do X, Y and Z. Put simply – we should ask the learners to do something and evaluate them on their ability to do it.

Multiple choice questions based on a reading text tell us something about their comprehension of the text (of those parts of the text which are tested) but nothing about what they might want to do *with their understanding of the text*. It is at best a partial picture of textual comprehension.

A speaking interview does not necessarily tell us about the candidate's ability to make a presentation or give a briefing, or make small talk.

Direct Testing is Better Testing

We should test a wider range of competencies in order to give us a better picture of their abilities in the language. This would mean that we would have more direct evidence rather than inferences that this candidate is proficient at this particular level of English.

We should produce a list of speaking and writing tasks which we think are important for our learners. We should then prioritise these as 'essential', 'useful', and 'non-essential'. Once the essential tasks are identified we should rank these; these tasks are the core tasks of our course and ones we should be spending most of our time and effort on.

We should actively test the essential ones in order of importance and consider testing the 'useful' category as portfolio tasks tested through continuous assessment.

Integrated Scenario-based Competency Tests

Integrated Scenario-based Competency Tests are an approach to testing the candidate's actual competency through direct tests of their ability to understand language and then actually do something meaningful with the language.

The key points are that the tests are (as the name suggests):

Integrated: For example, the candidate reads a text and then presents on it. This reflects real life. Although we read for pleasure, we don't test our comprehension on what we have read for pleasure; it's a personal learning experience, not an externally validated one, unless we discuss what we have read in a book club, for example. We might listen to something and make up our minds on the topic, but again this is an internal process which we do not submit to a comprehension check. It is only when we use the language to do things (talk, present, persuade, inform etc.) that there is any kind of external examination of the language we use. We might read the cinema listings, read reviews and then decide which film we want to see; we might then discuss this with others or attempt to persuade them to see the film we have chosen. We might do some research and the present on a topic in a seminar or a meeting, or write up a report. In each case, it is our language output which is subject to scrutiny in real life, and this should be reflected in our testing. In integrated testing, reading and listening texts are used as input, and comprehension of the input is tested through an evaluation of the output we produce based on that input.

Scenario-based: the tests are properly contextualised in a meaningful (though not necessarily serious) scenario which gives the test verisimilitude and thus more validity.

Competency-based: the tests test the candidate's ability to actually do things with the language so we can safely say that they have demonstrated the competency to do X rather than inferring that as they are '*at*' level C1 they can be assumed to be able to do X with a certain level of competency.

The Inspiration for the Concept

The idea for Integrated Scenario-based Competency Tests comes from the sport of competitive tactical pistol shooting. At a standard shooting range there will be lanes with a shooting bench at one end where the shooter stands with their pistol. At various distances down the lane there will be targets. The shooter stands at the bench and shoots at the targets, trying to group shots in as small a circle as possible in the centre of the target, for example. This activity, both practice and competitive, is analogous to completing sentence and text gap-fills in language learning and testing. The activity practices and tests a set of subskills in order, in shooting, to demonstrate the ability to group a number of shots on target. If this was all there was to shooting (and language learning) then that would be fine and that would be all that we would need to do. We could claim that a good shooter can group shots closely at 5/10/25 m ranges. And a good language user can successfully fill the gaps in the sentence(s) or texts. Yet that is not all there is to pistol shooting and being able to complete a gap-fill is not all there is to using language [and testers in their hearts know this; they use the gap-fill (for example) in order to infer language ability].

In competitive tactical pistol shooting competitors, good target shooters with many hours on the standard range, enter a building or area which contains a number of rooms or zones and move through the rooms or zones and deal with the situations they find there. In some rooms they will have to shoot several bad guys; in other rooms they will have to refrain from shooting civilians or hostages. This is the 'language in use' phase of pistol shooting. In these rooms they demonstrate their ability to double tap bad guys and not shoot good guys in order to show their competency in real-life scenarios. Their speed and accuracy skills (fast and accurate shooting) are deployed in varying situations to show they have the competency to deploy these skills in the real world. Such practice is not a one-off; the range is practiced on intensively with different scenarios. There is continual training of the basics at the standard range and intensive pre-competition training at the tactical range.

Similarly, in language learning and testing, learners and test candidate's need continual skills practice in quickly and accurately recalling the language (the target shooting practice stage), and intensive practice in using the language to do things like make presentations, write emails and reports etc. according to varying scenarios (the tactical range practice stage).

Testing Receptive Skills

While a focus on directly testing productive skills seems appropriate, can we include other elements so that reading and speaking can also be tested as well? Testing of the so-called receptive skills – listening and reading – is, of course, necessary. The term 'receptive' is useful short-hand for these skills but completely misleading in its idea of passivity. Reading and listening are hardly passive activities; we don't just receive the message – we have to work at it to extract it. It makes more sense to term these tests (*meaning-attitude) message* extraction tasks, where the test taker extracts meaning from the stream of speech or words on the page, deduces the attitude of the speaker-writer from their choices of words, and assembles a mental representation of the message from the information in the words used (given that <u>all</u> words contain information), and then decides on their *attitude* to the message which has been understood: belief/disbelief; agreement/disagreement; interest/indifference; credulity/incredulity; action/inaction and so on. In short, people read and listen to understand a multi-layered message of the text in order for them to take a position of what they have learned. And their position is something which we should be testing them on.

Reading and listening tasks can and should be incorporated into tests – where the candidates read or listen to a test with a conscious purpose – to use the information in the test for a follow-up speaking or writing task. How well they read the text, or understood the listening, will be clear through their speaking or written performance.

A Controlling Factor

Any test type should be rejected if the backwash effects are negative. This should be our controlling factor in test choice. Thus, any test task has to have a real-world analogue. It must be something they would have to do in real life.

Examples of Integrated Scenario-based Competency Test Tasks for ESP Situations

The individual tasks can range from relatively simple conceptualisations to long simulations.

1. You meet a colleague in the corridor. You have not seen him/her for a couple of days. Catch up with him/her. You start the conversation.
2. You are a police officer. You are going to interview a witness to a murder. Interview the witness, write up the statement and ask the witness to sign it. [Preparation: information for the witness; basic information for the police officer about the murder]
3. You are the Company XO. You have been asked to plan a field training exercise. Using the materials provided (map etc.) plan an appropriate timetable for the exercises, including list of equipment required etc. Prepare to brief the company commander and platoon leaders on the exercise. Your briefing should take 15 mins. You have 30 mins to prepare. [Preparation required: minimal – a suitable map.]
4. You are the unit quartermaster. As part of a reequipment exercise you have been asked to comment on possible replacements for camping equipment. Look at the specifications sheets of the equipment being considered and write a short memo on your thoughts on the suitability of the equipment. [Preparation required: specifications sheets of the selected equipment]
5. You are an air traffic controller. You have just arrived on duty. Take over from your colleague…… [There follows an hour-long simulation of air traffic control]. [Preparation required: a suitable chart; radio call signs, flight plans; cards with instructions; multiple players playing different roles.]

Preparation for Testing and Rating

Once the essential tasks have been identified and prioritised then test tasks should be written and moderated. This means that the task writers should present their tasks to their peers, which should do the tasks. Then the tasks are reviewed and revised.

The task evaluation criteria should also be drafted and used to see if they work. Ideally you would have some performances to use the rating scales on, so that you can standardise the rating between teacher-examiners.

Once you have agreed the tasks and rating scales for the testing session then you need to plan how the candidates move through the tasks. If you have one essential task – like the air traffic control simulation above, then you would need one room to do it in. If you have a number of tasks then you would have to decide whether to have the candidates in one room, doing task after task in that room, or whether you are going to go for the full tactical experience with the candidates moving from room to room and doing a different task in each room. This is much to be preferred because it builds motion into the testing – real life language use involves movement. We go to meetings. See people in passing. Stand up to talk. Pace the room as we think. This solution though will involve careful planning and staggered timings.

The Test Day

There are a number of different ways the test day could be organised. This is one, based on five tasks, five rooms, and each task being allowed 20 minutes.

Test Registration Room

The candidates arrive for testing at a set time and are briefed on the test. The first candidate arrives at 9 am, for example; Candidate 2 arrives at 9.25 and so on.

0905

Room 1

Candidate 1 enters the first room and does the first task, which is to read a text and write a summary memo of it. A teacher explains the task and watches the candidates write the task. This takes 20 mins.

0925- 0930

Waiting Area

Candidate 1 then leaves the room and moves to a waiting area.

Test Registration Room

Candidate 2 arrives and is briefed on the test.

0930 -0950

Room 2

Candidate 1 goes to Room 2. The candidate then listens to a short lecture and makes notes on the lecture. These notes will be evaluated later.

This candidate then goes back to the waiting area.

Room 1

Candidate 2 meanwhile enters Room 1 and does the task there. Then goes to the waiting area.

0950-0955

Test Registration Room

Candidate 3 arrives and is briefed on the test.

0955-1015

Room 3

Candidate 1 goes to Room 3. The candidate does a roleplay with the teacher-examiner in the room. A second teacher-examiner rates the interaction as it happens. This candidate then goes back to the waiting area.

Room 2

Candidate 2 goes to Room 2 to do the task. Then goes to the waiting area.

Room 1

Candidate 3 meanwhile enters Room 1 and does the task there. Then goes to the waiting area.

1020-1040

Room 4

Candidate 1 goes to Room 4. The candidate prepares a mini-presentation on a topic using some written prompts. The teacher-examiner rates the presentation. It is also recorded. This candidate then goes back to the waiting area.

Room 3

Candidate 2 goes to Room 3 to do the task. Then goes to the waiting area.

Room 2

Candidate 3 meanwhile enters Room 2 and does the task there. Then goes to the waiting area.

1045-1105

Room 5

Candidate 1 goes to Room 5. The candidate answers a series of questions from the examiner about a variety of topics. A second teacher-examiner rates the interaction as it happens. It is also recorded. This candidate then goes back to the waiting area. The test is over for this candidate.

Room 4

Candidate 2 goes to Room 4 to do the task. Then goes to the waiting area.

Room 3

Candidate 3 meanwhile enters Room 3 and does the task there. Then goes to the waiting area.

1110-1130

Room 5

Candidate 2 finally goes to Room 5. Then goes to the waiting area. The test is over for this candidate.

Room 4

Candidate 3 meanwhile enters Room 4 and does the task there. Then goes to the waiting area.

1135-1155

Room 5

Candidate 3 then enters Room 5 and does the task there. Then goes to the waiting area. The test is over for this candidate.

This could be shown on a table, as on the next page.

Costs and Benefits

Now, obviously there are costs with this approach – the greater number of rooms required, the increase in the time required to process a large number of candidates, and the increase in the number of invigilator-examiners required. Compared to a written examination conducted in one room this kind of system is much less *efficient*, but it is more *effective* – because we get much better quality of information from these tasks. The candidates also have a novel experience from this kind of testing. They enter a new room and have a new task to do. They move through the test in time and space and have a feeling of physical progress through the test.

	0900-0905	0905-0925	0925- 0930	0930 -0950	0950-0955	0955-1015	1015-1020	1020-1040	1040-1045	1045-1105	1105-1110	1110-1130	1130-1135	1135-1155
Test Registration Room	Candidate 1 arrives and is briefed on the test.		Candidate 2 arrives and is briefed on the test.		Candidate 3 arrives and is briefed on the test.									
Room 1		Candidate 1		Candidate 2		Candidate 3								
Waiting Area			Candidate 1		Candidate 1 & 2		Candidate 1,2 & 3		Candidate 1,2 & 3		Candidate 2 & 3		Candidate 3	
Room 2				Candidate 1		Candidate 2		Candidate 3						
Room 3						Candidate 1		Candidate 2		Candidate 3				
Room 4								Candidate 1		Candidate 2		Candidate 3		
Room 5										Candidate 1		Candidate 2		Candidate 3

6 Conclusions

This article has been written in the spirit of helping you to reconsider testing. In some situations, the traditional industrial approach to testing will be appropriate – in placement testing, for example, where multiple-choice questions or gap-fills will give us good enough information about certain specific language items and this information will enable us to, within the broad precision required, place our learners in an appropriate class.

These kinds of tests though do not give us *good enough* information to be able to comment meaningfully on what our learners can do with the language. We need to test their speaking and writing to do this because the *quality* and *quantity* of information from such tasks is so much more. We should also build in meaningful and purposeful reading and listening tasks into these speaking and writing tasks to test all skills. Our test should reflect real language use. They should give us good quality information. And they should be biased towards what we are good at – evaluating writing and speaking performance.

The backwash effect of such tasks will also be positive. If we are going to test army officers on giving orders, then we will practice giving orders. If police officers need to give briefings, then that is what we should practice and test. If we are training air traffic controllers then we should test them on this language in a realistic simulation.

We should prioritise what our learners need to do with the language and give them the language they need to do this. Then practice it again and again. And then test them on it in **Integrated Scenario-based Competency Tests.** Think of them as the live-firing exercises which mark the end of some training. Our testing should be realistic and reflect real world language use. It should be serious but not academic. It should be focused on language in use or language in action. And focused on the professional language competencies our learners need.

The original version of this article was published in Issue Nine (January 2022) of Teaching Uniformed Personnel, a Free Quarterly Magazine for Military, Police and Border Guard Teachers.