English Language Standards for Internationally Educated Nurses Seeking to Work in the United Kingdom: Recommendations from Standard Setting Exercises

Gad S. Lim

Head of Assessment and Research

Cambridge Boxhill Language Assessment

April 2022

Introduction

The Nursing and Midwifery Council (NMC) are planning a review of English language requirements for nursing and midwifery professionals who are trained outside the United Kingdom. This review is expected to take place between April and October 2022. In support of this review, in March 2022, the Occupational English Test (OET) conducted a series of standard setting exercises with panels of UK nurses, nurse educators, and nurse recruiters to identify what in their view the passing standard ought to be on the Listening, Reading, and Speaking components of the OET. This report provides the details of those exercises. A similar exercise had been conducted relatively recently for the skill of Writing (Seguis and Lopez, 2019) and was therefore not included.

Occupational English Test

The OET is a specific-purpose language test designed to assess the English language ability of healthcare professionals who seek to register and practise in an English-speaking environment. It is among the tests that the NMC accepts at present. There are versions of the test for 12 health professions, with nursing being one of them. Test takers are assessed on their Listening, Reading, Writing, and Speaking ability, with separate scores reported for each of those skills.

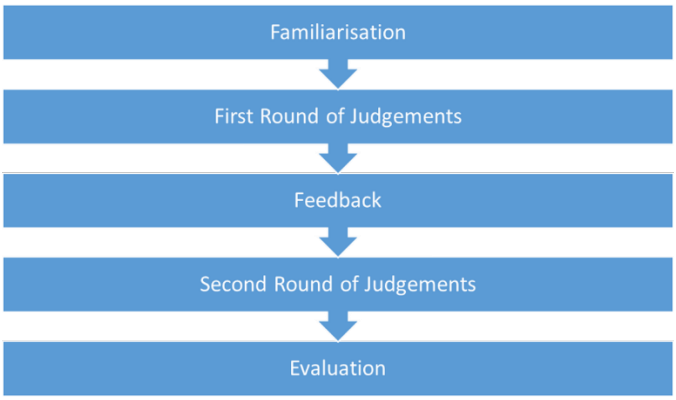
The Listening and Reading sections have 42 questions each and involve a variety of tasks. Listening includes taking notes based on a consultation, understanding short workplace interactions, and understanding a talk or interview about a health-related topic. Reading includes expeditious/fast reading of multiple texts, understanding short workplace texts, and understanding a long passage on topics of interest to healthcare professionals. The Speaking section has the test taker interacting with an interlocutor in two simulated consultations. It is scored on a number of language-related and clinical communication criteria.

Standard Setting

Standard setting is a process for separating test takers into a number of levels or categories (Cizek and Bunch, 2007). The aim is to identify the boundary between acceptable and unacceptable performances, and the score associated with that boundary. The recommended cut score and other information gathered from the standard setting exercise provide policy makers with a base of evidence to inform their decisions.

Central to standard setting is a meeting where a panel of suitably qualified individuals are led through the process of identifying cut scores by one or more facilitators (Figure 1). A familiarisation stage ensures that panellists know about the test they are working on, have adequately thought about what an acceptable person or performance ought to be like, and understand the standard setting method that they are employing. Multiple judgement rounds provide panellists with the opportunity to make and fine tune score recommendations. In between rounds panellists get to examine relevant information and discuss things that might help them with the judgement activity. After the meeting, the recommended cut scores are examined for evidence of procedural, internal, and external validity (Council of Europe, 2009).

Figure 1. The Standard Setting Process



While standard setting has traditionally involved face-to-face meetings, there is evidence that standard setting exercises carried out online can produce comparable and equally valid outcomes (e.g., Katz and Tannenbaum, 2014). Because of the continuing effects of the COVID-19 pandemic, the standard setting exercises described here were carried out online via Zoom. This modality made it easier to involve panellists from across the United Kingdom, making them more broadly representative.

Methodology

Standard Setting Panels

Standard setting exercises depend crucially on having participants who have relevant knowledge and expertise: of the test, of the test taking population, and of the performance context. For these standard setting exercises, each panel had a final size of 11 judges. (Each panel started with 12 judges but was reduced due to a late arrival, an early departure, and a withdrawal of consent.) The size of the panels is considered to be sufficient for producing dependable outcomes (Fowell, Fewtrell, & McLaughlin, 2008; Wu and Tzou, 2015). More importantly, the panels largely consisted of international recruitment leads and clinical educators who help onboard nurses for various NHS trusts, and several of the judges are themselves internationally educated nurses (Table 1). The panel members clearly have knowledge of international nurses and of the contexts in which they will be working, making them an appropriate group of expert informants for this activity.

Table 1. Demographic Information for the Panels

		Listening	Reading	Speaking
Age	**21-30**	1	0	2
	31-40	4	3	4
	41-50	3	4	4
	51-	3	4	1
Education	**Bachelors**	2	3	5
	Masters	9	6	5
	Doctorate	0	2	1
Experience	**1-10**	7	5	8
	11-20	3	3	2
	21-	1	4	1
Background	**UK**	7	8	6
	International	4	3	5
Current Role	**Clinical Educator**	5	7	6
	Int'l Recruitment	5	2	3
	Lecturer	0	1	1
	Nurse	1	1	1

Standard Setting Methods

In the field of educational measurement, different standard setting methods have been devised to deal with particular test types and contextual features.

For the Listening and Reading components, these exercises employed the Modified Angoff method (Angoff, 1971), which is the most frequently used method for discrete item tests (Council of Europe, 2009). In this method, the panellist is asked to conceive of the minimally acceptable person as concretely as possible. Then, keeping this borderline person in mind, they go through each of the test questions/items and decide: what is the probability that a minimally competent person will answer this question correctly? The sum of all those judgements becomes the judge's recommendation on the items that a minimally acceptable person will need to answer correctly.

For the Speaking component, the Paper Selection method (Plake, 1998) was used. As with the other method, the panellist is also asked to conceive of the minimally acceptable person as concretely as possible. They are then presented with a series of papers (in this case, the "papers" are recordings of the Speaking roleplays) which have already been scored by assessors, but whose scores are not made known to them. The recordings are played in order from the one with the lowest score to the one with the highest score, and the panellist selects the first performance that in their judgement reflect the level of the minimally acceptable person. The score received by the candidate in the recording becomes the judge's cut score recommendation.

Familiarisation

Prior to the standard setting meeting, participants were required to complete several activities to familiarise them with the OET. They were asked to review relevant pages on the OET website, and then complete a module that introduced various aspects of the test: what the test is, who the test taking population are, what the test looks like, how it is delivered, and how it is scored. They also had to take a sample test of the skill area for which they are serving as judge. For those working on the Speaking test, they were also given copies of the assessment criteria and assessment glossary.

At the standard setting meeting, the panellists were reminded of the purpose of the exercise, presented with a refresher on the test, and introduced to the standard setting process and the method being used. Panellists were then trained in conceptualising the minimally acceptable person; in this case, the nurse whose English is just good enough to practise safely in the UK healthcare context. They had the opportunity to discuss, first in small groups, then as a large group, those characteristics that best helped them conceptualise that borderline person. They were encouraged to think very specifically about the target skill. This section concluded only after all panellists indicated that they had a good idea of the minimally acceptable person and had no further questions or points for discussion.

For the Listening and Reading components, with the borderline person fresh in their minds, the panellists were asked to take the relevant test component under timed conditions. Afterwards, they were also provided the answer key to mark their own work. Taking the test introduces them to the contents of the specific test form they are to render judgments on, and also helps them consider what the experience of taking the test and answering those questions might be like for the borderline person. Having the answer key and knowing how many items they themselves answered correctly can help judges develop a better sense of what a reasonable level of performance to require might be.

Judgement

For the standard setting proper, the panellists were asked to enter their judgements into a shared online document. Each panellist was assigned an anonymous judge number, to ensure that they felt comfortable expressing their true opinions without fear of being judged by other participants. They were reminded of the task as well as relevant things to take into account (e.g. test day stress for actual test takers; test takers needing to concentrate over the entire test rather than just one component; their now encountering the test questions a second or third time; probability may be affected by guessing behaviour on multiple choice items). For Listening and Reading, because making exact probability judgements is cognitively quite challenging, they were asked to only render judgements in increments of 5 (e.g., 0.05, 0.10, 0.15…). For Speaking, it was to indicate the first performance they listened to that they thought was minimally acceptable. Throughout, panellists had access to the question papers, their answers, and their notes.

There were two judgement rounds for each skill. In between judgement rounds the panellists were provided with feedback (i.e., an overview of the groups' judgements), reviewed questions they found unusual or difficult to judge, and given opportunity to ask any questions they had or other topics they wanted to discuss. The feedback and discussion are meant to provide panellists with relevant information on which to base their final judgements in the second round.

At the conclusion of the meeting, the panellists were asked to complete an online evaluation form, which provides another source of evidence for the validity and reliability of the recommended cut scores arrived at.

Analysis of Validity Evidence

The documentation above of the steps followed in the standard setting meeting forms part of the evidence for procedural validity. Another source of procedural validity evidence is the feedback provided by the panellists in the evaluation forms. The evaluation forms asked to rate a series of statements on 4-point Likert-type scales, where 1 is strongly disagree and 4 is strongly agree. There was also a free text field for any other feedback they wished to provide.

On the basis of the evaluation, it would appear that the panels are satisfied with the process and the timing of the exercises, as well as with the cut score recommendations they arrived at (Table 2). The free text comments were also largely positive, apart from some feeling that Part C of the Reading test appeared to be difficult. (In that regard, the facilitator explained that putting a lower probability figure for those items would bring the recommended cut score down, which would account for the items being more difficult.) The free text comments can be found in Appendix A.

Table 2. Evaluation of the Standard Setting Process

	Listening	Reading	Speaking
I understood the judgement process.	3.50	3.55	3.64
I had enough time to complete my individual tasks.	3.75	3.64	3.82
I had enough time to participate in the discussions.	3.67	3.55	3.82
I understood the group discussion of our judgements.	3.58	3.45	3.45
I am confident in the decisions I have made.	3.42	3.27	3.64
I am confident our cut score recommendations are appropriate.	3.33	3.36	3.55

(1=strongly disagree … 4=strongly agree)

In terms of internal validity, it is usual practice to estimate the standard error of judgement (SEj) of the cut score recommendations. It is desired that the SEj be less than half the test's standard error of measurement (Cohen, Kane, and Crooks, 1999). Table 3 shows that all three panels judgements met that criteria, which suggests that the cut score recommendations would in all likelihood be replicated in another standard setting study.

Table 3. Standard Error of Measurement and Standard Error of Judgement, raw scores

	Listening	Reading	Speaking
SEM	2.54	2.62	1.82
SEM/2	1.27	1.31	0.91
SEj	0.96	0.87	0.35

Taken together, the validity evidence presented above indicates that the results obtained are worth giving credence to.

Results

The judges' cut score recommendations were aggregated, and a summary of the different panels' recommendations is presented in Table 4. (The full individual judgements can be found in Appendix B.) It can be seen that the mean and the median is quite close on all three skills, which indicates that the recommendations are not significantly affected by outliers, and the standard deviation of the panels' judgements went down from Round 1 to Round 2, which reflects greater convergence among judges in each panel. The recommendations also remained largely unchanged between Round 1 and Round 2, which could be taken as a sign that judges have high confidence in their judgements. All of these are positive evidence in support of the validity of the outcomes.

In addition, because of the NMC's interest in fairness and bias concerns, also listed in Table 4 are the average judgements of panellists from the UK and panellists originally from outside the UK, though it should be noted that the n size of these cells is such that strong conclusions probably shouldn't be drawn based on them. The only significant difference, albeit not a substantial difference, is in Speaking, and then only in Round 2. The overall picture presented is that panellists from different backgrounds do arrive at substantially the same recommendations.

Table 4. Cut Score Recommendations, raw scores

	Listening (out of 42)		Reading (out of 42)		Speaking (out of 78)	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Mean	28.29	28.28	26.23	25.69	50.55	50.64
Median	28	28	26	27	51	51
Minimum	22	23	20	19	49	49
Maximum	35	35	36	29	54	53
SD	3.5	3.3	4.1	2.9	1.5	1.1
Mean (UK)	28.01	27.99	26.53	27.06	50.86	51.33*
Mean (International)	28.78	28.79	23.47	24.02	50.20	49.80*

**p<.05*

Because different test forms contain different sets of items which differ somewhat in difficulty, it is necessary to convert raw scores into a standardized metric to facilitate comparability and interpretability. Table 5 shows the cut score recommendations from the March 2022 panels, as well as those from exercises conducted in 2017 and 2019, also with nursing profession-related informants (Seguis and Lim, 2020). Comparing the results of a standard setting exercise with other sources regarding the standard is evidence for the exercise's external validity.

For Listening and Reading, the current panels' recommendations are remarkably consistent with those of the 2017 panels. Some additional information may help to make this more evident. The current Listening panel recommended a standard of 28 on this test form. Had their recommendation been 27, that would have equated to a scaled score of 350. Similarly, had the 2017 Listening and Reading panels' recommendations been 1 point lower, they would also have equated to 350. That is to say, there is now even more evidence that the appropriate Listening and Reading ability level for nurses seeking to practise in an English-speaking healthcare environment is extremely close to the current NMC standard of B.

The panel for Speaking provides a somewhat different story, arriving at a cut score recommendation of 320, in the middle of the C+ range of 300-340, which is lower than what previous standard setting exercises and lower than the current NMC standard. This is therefore the skill where further consideration may be warranted and further evidence sought.

Table 5. Cut Score Recommendations, Current and Previous

	Listening	Reading	Speaking
Recommended Cut Score	28	27	51
Scaled Score Equivalent	360	350	320
2017/2019 Standard Setting	360	360	350

Conclusion

This report provided the details of standard setting exercises conducted with groups of UK-based nurses, nurse educators, and international nurse recruiters to determine what in their judgement the minimum acceptable standard ought to be on the OET Nursing test's Listening, Reading, and Speaking components. Details of the process as well as the results were presented, as well as evidence in support of the exercises' procedural, internal, and external validity.

For Listening and Reading, the recommended cut scores are essentially identical to those recommended by panels in 2017, which in part informed the standard currently accepted. While the recommendation is a hair more than the current standard, it is so close to the current B / 350 that it probably does not warrant a policy change in response.

For Speaking, the recommended cut score of 320 is lower than the current standard of B / 350. Seguis and Lim (2020) showed that a lower cut score for the Writing component is warranted by a change to the amount and type of writing that nurses do nowadays compared to the past. If a similar explanation were available for speaking, then it would help to support a policy change for this skill area. (Should the NMC wish to implement a change, they would need to articulate the standards in terms of scaled scores, i.e. 320, as a grade of C+ encompasses performances between 300 and 340, some of which would be below the minimally acceptable level.)

The recommended cut scores represent one important source of evidence that the NMC can consult, but they can of course set the standard higher or lower taking into account other relevant factors and considerations (e.g. other stakeholders not represented in these panels). Setting the standard too low and setting the standard too high can both have a negative impact on patient welfare in the UK, and the task is to set the standard at just the right place.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*, *2nd Ed.* Washington, DC: American Council on Education.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*, 343-366.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A manual.* Strasbourg, France: Council of Europe.

Fowell, S. L., Fewtrell, R., & McLaughlin, P. J. (2008). Estimating the number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? *Advances in Health Sciences Education*, *13*, 11-24.

Katz, I. R., & Tannenbaum, R. J. (2014). Comparison of web-based and face-to-face standard setting using the Angoff method. *Journal of Applied Testing Technology*, *15*(1).

Plake, B. (1998). Setting performance standards for professional licensure and certification: Implications for National Assessment of Educational Progress. *Applied Measurement in Education*, *11*, 65-80.

Seguis, B., & Lim, G. S. (2020). The writing that nurses do: Investigating changes to standards over time. *Assessing Writing*, *46*.

Seguis, B., & Lopez, S. (2019). *OET Writing Standard Setting Workshop Final Technical Report.* Cambridge, UK: Cambridge Assessment English.

Wu, Y. F., & Tzou, H. (2015). A multivariate generalizability theory approach to standard setting. *Applied Psychological Measurement*, *39*, 507-524.

Appendix A: Panellists' Evaluation Form Comments

Listening

This was a very informative session and it was great to be part of this process
Great workshop. Happy to be a part and contribute.
Very interesting to understand the process.
Very interesting experience. Knowledgeable facilitator.
Thank you for organising this session. It was informative and able to identify the pros and cons in answering the questions.
Interesting variation in the scores submitted by the group and while these were anonymous for us I would be interested to see how they linked to the different experiences of the people involved. The group had representation from internationally trained nurses who themselves had completed language assessments and nurses engaged in supporting internationally educated nurses and I wonder if their perceptions of probability varied? I have even greater respect for the nurses who pass this test as maintaining concentration and focus is a real challenge as the time goes on and we only sat one part of the test! Thanks Gad for your direction
It was my pleasure engaging in this all important exercise, I enjoyed every bit of it. I am extremely proud of the Nursing profession but I believe we can be more than what we currently represent and can do absolutely much better.
whilst I felt the session went well and I understood the process and expectations, 12 people is a very small group of people looking at this so I hope this has been a useful process. In terms of the OET I think it is a good exam and am satisfied with the level of nursing it involves. however there are still more elements to consider when recruiting IENs. The OET involves a mix of English speakers with different variations on pronunciation for example consideration for British, American, Australian etc versions could be a consideration.
The excel spread sheet used started from number 3 which related to question 1 which was a little confusing. Thank you, it has been a very informative and enjoyable day.

Reading

This was really useful and very interesting to hear the experiences from other trusts and also to note the similarities in our problems. Very interesting to see our session came up with the same standard that is currently set but think that the 2 final texts are difficult to answer and may be more suited to medical staff rather than nursing staff. Would it be appropriate to have different texts for different staff groups? I do feel the timings were appropriate however.
I felt section C was very difficult I got a number of them wrong I felt the time length was difficult also. Some of the questions are very testing in terms of analysis and deep comprehension. Some questions I felt more than one answer could apply. Q11 one such question. Section C is a text which a nurse would

not find in a clinical context they would when reading research papers but surely this test is to check day to day real text that they would see. Section C could be a care plan which could be altered regularly and this would make much more sense to the nurse.
Though it looks like the group has ascertained that the b score is correct, this is not what the discussions signposted to. If previously there were 30 items to be passed to score a B, our results showed this was 26 items? again in the discussions: the level of English proficiency for 3rd year nursing student at the point of qualification/entry to register is 6.5 why is this not for overseas nurses? These topics which we encountered do not depict day to day nursing texts/documents. The level of language is quite high More than cognitive processing it was cognitive stacking? There was not taking into account the participant skill in undertaking this exercise. In the group where is the British born newly qualified nurse? or 3rd year nursing students? where is the patient/service users?
Really challenging test, particularly part C and I share the concerns expressed by the group about the complexity of the text and the question as to whether in real life this would be something you would read at speed as needed in the test. I do appreciate the need to evaluate the ability for individuals to read and interpret different forms of text and the levels you need to understand but a tough test!
the discussions were useful but in my opinion some of my colleagues were too focused on the content of the exam questions rather than the language knowledge required. the OET is not testing the nurses knowledge on a particular subject matter the same as the IELTS exam didn't and it is easy to lose focus on this more so when it is related to a specific profession. I think it would it be really difficult to find an exam that would replicate a day to day role of a nurse and from my perspective should not be a requirement. I think there could be some scope to extend the timings slightly as I struggled with this however this may have been more to do with moving between two screens than the actual questions.
Thank you for the session. The comprehensive passage can be more simplified. The part c questions was a daunting prospect to answer on time; the minimum competent person had to read the passage repeatedly. Therefore the 45mins won't is enough to answer the questions. Although the passage was exciting information, there is still more to read and understand its gist. Maybe if we consider formulating more simplified courses, there is a chance of getting a high score in part C.
Excellent session and very well facilitated by Gad. Have high levels of confidence in the information being shared with the NMC

Speaking

Thank you for the opportunity afforded. I have greater confidence in this standard setting - listening - than reading exercise. Both exercises were superbly facilitated. However, some of my criticisms are:||Minimally competent person - the definition was highly subjective.|How can a a sample of 14 panelists confirm the recommended score is correct?|In an ICU the communication style is different to the Outpatient department/acute medicine/community - it is the 'communication' that is supposed to be tested, however in some of the scenarios, we need to be conversant with the national

policies/guidelines - to delegate/advice/refer.|How are the qualitative data - that is - the discussions captured and evidence in the report?|Just my thoughts.||God bless.
Was a long time to have to concentrate- may have been helpful to have a break in the listening part for 5 mins or so
For future sessions, hopefully we have more clearer audio/recording, as the number 3 for the 1st round was poor in terms of clarity thank you
Very informative day, thank you. It's been a pleasure taking part.
Yes, I would strongly recommend, 320 will be the best score for the minimum competent person.
Great workshop!

Appendix B: Panellists' Detailed Judgements

Listening

Round 1				Round 2		
Judge	**%**	**Items**		**Judge**	**%**	**Items**
1	72.14	30		**1**	72.14	30
2	65.48	28		**2**	61.90	26
3	56.67	24		**3**	56.67	24
4	73.45	31		**4**	74.76	31
5	73.45	31		**5**	72.14	30
6				**6**		
7	66.67	28		**7**	66.90	28
8	70.71	30		**8**	70.00	29
9	53.10	22		**9**	54.40	23
10	83.57	35		**10**	82.86	35
11	59.17	25		**11**	62.74	26
12	66.43	28		**12**	66.17	28
Mean	67.35	28.29		**Mean**	67.34	28.28
Median	66.67	28		**Median**	66.90	28
Min	53.10	22		**Min**	54.40	23
Max	83.57	35		**Max**	82.86	35
SD	8.3	3.5		**SD**	7.9	3.3

Reading

Round 1		
Judge	**%**	**Items**
1	68.57	29
2	65.48	28
3	58.69	25
4	51.67	22
5	57.86	24
6		
7	48.33	20
8	55.83	23
9	65.36	27
10	86.67	36
11	61.79	26
12	66.67	28
Mean	62.45	26.23
Median	61.79	26
Min	48.33	20
Max	86.67	36
SD	9.79	4.1

Round 2		
Judge	**%**	**Items**
1	68.57	29
2	65.83	28
3	53.93	23
4	68.81	29
5	56.19	24
6		
7	46.31	19
8	55.83	23
9	65.12	27
10	66.67	28
11	61.67	26
12	63.93	27
Mean	61.17	25.69
Median	63.93	27
Min	46.31	19
Max	68.81	29
SD	6.8	2.9

Speaking

Round 1			Round 2	
Judge	**Score**		**Judge**	**Score**
1	51		**1**	51
2	49		**2**	49
3	49		**3**	49
4	52		**4**	51
5	50		**5**	49
6	51		**6**	51
7	51		**7**	51
8	51		**8**	51
9	49		**9**	51
10	49		**10**	51
11	54		**11**	53
12			**12**	
Mean	50.55		**Mean**	50.64
Median	51		**Median**	51
Min	49		**Min**	49
Max	54		**Max**	53
SD	1.5		**SD**	1.1