TEST REVIEW: OCCUPATIONAL ENGLISH TEST

MAGGIE MCGEHEE
University of Hawai'i at Mānoa

INTRODUCTION

The Occupational English Test (OET) is an English for Specific Purposes (ESP) test designed to assess the language proficiency of healthcare professionals seeking to work or study in an English-speaking setting. Developed at the University of Melbourne and originally used in Australia, it is now recognized by health boards and decision-makers in fifteen countries. This review describes the test purpose, design, and scoring methods, and explores aspects of the OET that both provide support for and question the validity of using OET scores in the life-or-death decision of who is permitted to practice medicine.

TEST PURPOSE

The OET offers twelve occupation-specific tests, targeted to dentists, dietitians, doctors, nurses, occupational therapists, optometrists, pharmacists, physical therapists, podiatrists, radiologists, speech pathologists, and veterinarians. The publisher, Cambridge Boxhill Language Assessment (CBLA), does not disclose the annual number of OET test takers, but releases percentages by demographics, occupation, and first language (CBLA, 2023). Medical licensure boards, hospitals, universities, and training programs use OET for admissions and employment decisions for applicants whose first language is not English. OET results are also used to make immigration decisions for skilled worker or student visas. In the United States (US), OET is one of two options for graduates of non-US medical schools to meet a communication skills requirement before practicing or studying advanced medicine in the US (Educational Commission for Foreign Medical Graduates (ECFMG), n.d.).

The OET advertises that selecting an ESP test focused on healthcare benefits both medical decision-makers and the test takers themselves, who can focus test preparation on using English in their intended occupational context. In addition to claiming "proof of ability to communicate

effectively" in English, the OET further purports that "successful OET learners have the English language and clinical communication skills to provide high quality and safe patient care" (CBLA, n.d.). OET emphasizes that it does not test medical knowledge, only language ability, but purposefully uses only healthcare-based scenarios and materials in order to assess English proficiency in this specific domain. In addition to evaluating linguistic competence, the OET intends to evaluate test takers' extra-linguistic skills for compassionate care, empathetic listening, and navigating power differentials between practitioner and patient.

TEST METHOD

Test takers have three options for completing the OET: on paper in person at a testing center, or on a computer either at a testing center or at home. The entire test takes just under three hours to complete. Four timed subtests focus on Listening (45 min), Reading (60 min), Writing (45 min), and Speaking (20 min). The Listening and Reading passages discuss general health-related topics and are the same for all professions for these receptive language tasks. These general passages are chosen to minimize privileging or testing any candidates' specific medical knowledge, and instead focus on language ability. They use medical vocabulary and discuss themes of concern to health practitioners, but avoid requiring specialist knowledge. By contrast, the Writing and Speaking tasks are profession-specific, presenting test takers with patient cases and clinical settings unique to their specialty for these productive tasks (CBLA, n.d.). Table 1 summarizes OET's four subtests and their length, description, and items.

Table 1

OET Subtests: Length, Task Description, and Items

Subtest	Length	Task Description	#/type of items
Same content for all professions			
Listening	~15 min	Part A. Listen to 2 patient consultations and complete case notes	24 gap-filling
	~12 min	Part B. Listen to 6 workplace conversations, briefings, or consultations, and identify the gist, details, and opinions	6 multiple choice (3-option)
	~13 min	Part C. Listen to 2 interviews or presentations/lectures, and identify the gist, details, opinions	12 multiple choice (3-option)
Reading	15 min	Part A. Scan 4 texts on the same topic, and read to find needed details	20 matching, short answer, sentence completion
	~10 min	Part B. Read 6 extracts of workplace communications (i.e., memos, policy documents, manuals), find main ideas, gist, or details	6 multiple choice (3-option)
	~35 min	Part C. Read 2 long passages on different healthcare topics, answer meaning, attitude, opinion, vocabulary questions	16 multiple choice (4-option)
Profession-specific content			
Writing	45 min	Write a letter based on case notes to refer, transfer, or discharge a patient. (Some professions advise or inform a patient.)	1 prompt, write a 180-200 word letter
Speaking	20 min	2 role plays. An interlocutor plays a patient, client, relative, or caretaker in a simulated clinical visit	2 role plays

The **Listening** subtest has three parts with 42 items in total. In each part, candidates hear recorded materials and answer questions when prompted. Each recording plays only once. In Part A, test takers complete case notes via 24 gap-filling items while listening to two different clinician consultations with patients. Responses on Part A are scored by human assessors against standardized keys (see "Scores," below). Part B and C are multiple choice sections scored by computer. Part B is comprised of six workplace conversations. One multiple choice question for

each asks test takers to identify the gist, details, or opinions in the conversations (six total items). Part C includes two longer recordings of healthcare-related presentations or podcast-style interviews, with six multiple choice questions during each recording (12 total items). Notably, "a range of accents are used to reflect the global nature of the healthcare workforce" (CBLA, n.d.), not to mention that of patients. Test takers read the written questions and instructions describing the context of each excerpt before playback begins (The Official Guide to OET, 2018).

The **Reading** subtest also includes three parts. Part A (expeditious reading task, 15 min) presents four short texts on a single topic, followed by 20 matching, sentence completion, and short answer questions. Test takers are instructed to skim the four texts to locate the needed information, then look for the detailed answers. As with Listening, Part A is scored by assessors, while Parts B and C (careful reading tasks, 45 min) are computer-scored. Part B presents six short texts that might be found in a healthcare workplace (such as a policy document, workplace guidelines, manuals, or internal memos). One multiple choice question for each text asks test takers to select a detail, the gist, or the main point. Part C asks candidates to find the meaning and the author's attitude or opinion in two longer, general healthcare texts by responding to eight, four-option multiple choice questions on each. Test takers must self-pace over the 45 minutes allotted for Parts B & C combined. (CBLA, n.d.)

The **Writing** subtest is an integrated task, combining reading and writing. Test takers are given a set of case notes or other documentation about a patient, and are directed to write a 180-200 word letter. Usually, this letter refers the patient to another practitioner for further treatment, but may also be a letter transferring the patient to another hospital or discharging them from care. Some professions may instead be directed to address the letter to a patient or relative and give advice or information. Candidates have an initial 5 minutes to read the documentation and instructions, followed by 40 minutes for writing. (CBLA, n.d.)

The **Speaking** subtest comprises two roleplays. In each, a trained interlocutor plays a patient, relative, or caregiver. The two roleplays cover two different clinical visit scenarios. Test takers have three minutes to read each roleplay card giving them initial background information on the case. Some details are omitted that they must elicit during the roleplay. In addition to assessing the patient's condition and providing medical guidance, test takers must establish rapport, show empathy, and fulfill specific tasks given on the roleplay card. For computer-administered tests, the Speaking roleplay is conducted via Zoom. Whether taking the test in person or on the

computer, the timed five-minute interactions are audio-recorded with no video images, and the audio-only recordings are scored. (CBLA, n.d.)

SCORES

The OET is primarily a criterion-referenced test, in which test takers are evaluated against set rubrics (or keys for the multiple choice questions) that aim to define what each person can do in English. Test takers receive four separate scores, one for each subtest. Subtest scores range from 0-500 and are also paired with a letter grade (A, B, C+, C, D, or E). No overall or combined score is assigned, as the OET emphasizes that candidates may display different levels of ability in different linguistic areas. Scores are also not calculated or reported in relationship to other test takers' scores.

CBLA publishes an equivalency table relating OET scores to International English Testing System (IELTS) Academic and Common European Framework of Reference for Languages (CEFR) bands (CBLA, 2020; Lim, 2016). As shown in Table 2, OET provides only one set of overall band descriptors; even though test takers do not receive an overall score, OET does not provide skill-specific band descriptors for each subtest. Further, OET does not define any specific cut scores, leaving that determination up to decision-makers who use the results. CBLA does note that many health boards set a minimum score of 350/B on each subtest, and often require candidates to achieve all four scores in a single sitting (CBLA, n.d.). For example, graduates of foreign medical schools seeking to practice or study in the US must earn a minimum score of 350/B on all four sections in a single test administration (ECFMG, n.d.).

Table 2

OET Numeric Scores, Letter Grades, Band Descriptors, and IELTS and CEFR Bands

OET numeric score	OET letter grade	OET band descriptor	IELTS band	CEFR band
450-500	A	Can communicate very fluently and effectively with patients and health professionals using appropriate register, tone and lexis. Shows complete understanding of any kind of written or spoken language.	8.0-9.0	C2
400-440	B	Can communicate effectively with patients and health professionals using appropriate register, tone and lexis, with only occasional inaccuracies and hesitations. Shows understanding in a range of clinical contexts.	7.5	C1
350-390			7.0	
300-340	C+	Can maintain the interaction in a relevant healthcare environment despite occasional errors and lapses, and follows standard spoken language normally encountered in his/her field of specialization.	6.5	B2
250-290	C		6.0	
200-240			5.5	
100-190	D	Can maintain some interaction and understand straightforward factual information in his/her field of specialization, but may ask for clarification. Frequent errors, inaccuracies and mis- or overuse of technical language can cause strain in communication.	<5.5	--
0-90	E	Can manage simple interaction on familiar topics and understand the main point in short, simple messages, provided he/she can ask for clarification. High density of errors and mis- or overuse of technical language can cause significant strain & communication breakdowns.		--

As mentioned above, the multiple choice questions in Parts B and C of the Listening and Reading subtests are computer-scored. The Writing and Speaking subtests and Part A of the Listening and Reading subtests are scored by trained assessors. At least two different, randomly-assigned assessors score each candidate's response on these subtests. OET adjusts scores based on raters' patterns of severity or leniency (McNamara et al., 2019). If the two assessors' scores

do not match, or if an assessor has questions about how to rate a response, the task is referred to at least one additional senior assessor as well.

For Listening and Reading, scoring guides detail what and how much correct information must be included in a response to receive a given score. Writing and Speaking subtests are scored against "Assessment Criteria and Level Descriptors" updated in 2019 and 2018, respectively (CBLA, n.d.). Five Writing subtest criteria are scored from 0 to 7: Content, Conciseness & clarity, Genre & style, Organization & layout, and Language. A sixth criterion (Purpose) is scored from 0 to 3. Scores on the Speaking subtest are based on four linguistic criteria rated from 0 to 6 (Intelligibility, Fluency, Appropriateness, and Grammar & expression), and on six clinical communication criteria rated from 0 to 3 (Relationship building, Understanding & incorporating the patient's perspective, Providing structure, Information gathering, and Information giving). These critera scores are then converted to the test taker's final reported subtest score.

Importantly, OET does not disclose the specific calculations used to relate critera scores to the reported numeric subtest scores or letter grades. OET does state that in order to receive a 350/B on Writing or Speaking (commonly set as the cut score by decision-making bodies), the test taker must achieve "a high level of performance on all…criteria" and that "Test-takers securing grade B will have achieved predominantly scores of 5 out of 6 on each linguistic criteria and 2 out of 3 for the clinical communication criteria" (CBLA, n.d.).

COST AND PUBLISHER

The OET registration fee is quite high, at US $455 / AU $587 as of May 2023. CBLA does not publish decision factors behind setting this high fee. All four subtests include at least one part scored by two or more human assessors, and trained interlocutors are necessary for the Speaking roleplays. These would contribute to cost; still, other English proficiency tests incorporate these same features without such high fees. OET takers may also elect to take individual subtests, at a fee of AUD $200.50 for one, AUD $399.00 for two, or AUD $477.50 for three (CBLA, n.d.). Again, OET does not dictate whether candidates should only be evaluated based on scores received on all four subtests in a single sitting, or may substitute individual section scores by

retaking a section or sections. As with setting a cut score, these determinations are left to the decision-making bodies accepting OET (CBLA, n.d.).

Cambridge Boxhill Language Assessment (CBLA) publishes the OET. The OET Centre is based in Australia, with collaboration from Cambridge Assessment English, the Australian Research Council, and the University of Melbourne. The Cambridge English Language Assessment Research and Validation team develops the listening and reading portions, while speaking and writing materials are developed by OET test writers. Contact information:

The OET Centre	AUS +61 3 8658 3963
PO Box 16136	UK +44 1202 037333
Collins St West VIC 8007	USA +1 855 585 0125
Australia	www.occupationalenglishtest.org

Validity

In part, validity asks whether a test measures what it purports to measure, in its specific context and for its intended use (Brown, 2005; Chapelle, 2012). OET states that "successful OET learners have the English language and clinical communication skills to provide high quality and safe patient care" (CBLA, n.d.), and scores are used to make decisions about who may study and work in healthcare in English-speaking settings. To support the validity of using OET scores in this way, evidence should suggest that those scores indeed reflect test takers' linguistic abilities within the target healthcare context; that OET reliably provides consistent results; and that decisions based on the scores successfully identify professionals who can deliver care in English.

The strongest factors in support of an OET validity argument are its publisher's ongoing efforts to create items that reflect language use in the target healthcare context, and to produce scores that suggest how successfully test takers will communicate in English in the healthcare workplace. Questions remain about statistical evidence for reliability and benchmarking with other established measures. Additional data are needed to evaluate the effects of OET's 2018 and 2019 Speaking and Writing rubric revisions.

Support

OET is definitively a test of English for a specific purpose, and OET designers appropriately strive to approximate real-world language use within a health domain. From its initial development, OET applied linguists collaborated with supervisors of international healthcare providers to craft criteria and scoring based on their evaluation of language levels necessary to provide safe care (McNamara, 1996). Accordingly, Speaking roleplays simulate clinical visits. Listening and Writing sections, wherein test takers review audio or written texts and compose case notes and letters, reflect real daily work responsibilities. UK NARIC notes that the OET Listening items "reflect authentic features through use of pace, emphasis, digressions and accents and, as such, are well selected to reflect real life communication within a healthcare setting" (Coleman, 2019). OET Reading tasks draw on a range of typical medical-domain texts, ranging from technical reports and journal articles to hospital memos and emails. These materials are well suited to OET's specialized purpose, supporting content validity because test items are similar to real-world tasks. Test takers can expect to be familiar with the situations and documents from work experience, and decision-makers can expect that scores reflect candidates' use of English in healthcare scenarios, albeit simulated.

Equally important for validity, OET emphasizes that it does not test medical competency outside of language. OET designers take care to avoid questions that rely on understanding a certain disease or treatment plan, or that would privilege test takers who happen to have that expertise. This supports OET's claim to assess language proficiency and not something else. For example, Listening and Reading sections are not specialty-specific, but use general medical and health-related topics and materials "to test language knowledge and ability over and above the candidate's knowledge of that field" (CBLA, n.d.). The Writing and Speaking sections purposefully present targeted materials based on the test taker's profession, because different professionals "engage with…patients about different issues in different contexts." Even in these profession-specific subtests, scores do not consider the quality of the test taker's medical assessments or diagnostics, but only their use of English to convey their messages. This is appropriate to an ESP test's use for evaluating language ability only, within a target domain.

Validity also considers whether a test positively impacts those who take it (Chapelle, 2012). Prospective OET test takers are positively spurred to acquire domain-specific vocabulary, linguistic tools for structuring clinical visits, and letter- and memo-writing conventions in

English. OET linguistic criteria require successful test takers to command English grammar, pronunciation, and register, and its newer clinical communication criteria evaluate extra-linguistic skills such as demonstrating empathy and explaining medical terminology in lay terms for patients. These tasks are specific to successfully using English while providing care. Further supporting positive impact on test takers for this ESP test, Carr found that prospective medical practitioners "overwhelmingly prefer the OET" to IELTS, "rating it more achievable, more relevant, and more motivational than the IELTS" (Carr, 2021).

The OET uses a variety of item types to assess candidates' language proficiency. This variety balances the ease and practicality of scoring selected response items, with the greater authenticity of extended production tasks, supporting validity. Item types including selected response (multiple choice, matching), limited production (gap-filling, short answer, sentence completion), and extended production tasks (letter writing, role plays) (Qian & Pan, 2013). Test takers also must demonstrate their ability in a variety of linguistic arenas even within a subtest (i.e., test takers must read in the Writing section, write in the Listening section, and Listen in the Speaking section). OET validity is supported by inclusion of these tasks that require candidates to interact with another speaker in real time, and write a document reflective of one they would write when working in healthcare.

OET's rigorous scoring approach and attention to inter-rater reliability also lends support for validity. Other than the multiple-choice sections, each item is scored by two assessors who are randomly selected. If their scores do not match, the item is referred for re-scoring to a third rater. Further, OET assessors are "monitored for accuracy and consistency, and the scores they award are adjusted to take into account any leniency or severity" (CBLA, n.d.). Although OET does not release details on how these adjustments are calculated, these practices support inter-rater reliability, so that individual test takers will receive comparable scores regardless of which assessors are assigned to score them.

Beyond the test, for predictive validity, it is important to ask how well scores actually predict test takers' successful use of English to perform real-life language tasks. OET's ongoing approach to test development and revising rubrics based on what is important to actual health practitioners also supports its validity as an ESP test. Over time, OET received feedback that test takers who scored high on the Speaking subtest often struggled with communication in the workplace (Vidaković & Khalifa, 2013). OET launched a research project in the early 2000s to

evaluate and refine its own evaluation criteria, demonstrating active response to criticism and a commitment to improvement.

OET's research program explored the test's validity by comparing evaluations from applied linguists and OET assessors with those of and healthcare practitioners with domain knowledge (Manias & McNamara, 2016; T. McNamara et al., 2018; Pill, 2016; Pill & McNamara, 2016; Séguis & McElwee, 2019). As part of that project, Elder et al. (2012) asked practicing doctors, nurses, and physical therapists to comment on what they found relevant about language use in clinical roleplays. The researchers identified themes such as logical question structuring, asking open-ended questions, explaining medical terminology in lay terms, and showing sensitivity to patients (Elder et al., 2012). These concepts were absent from the then-current OET Speaking criteria, which were at that time limited to linguistic skills only. Using only linguistic scoring criteria might be more appropriate for an overall language proficiency test, but left an evaluation gap in this ESP test.

Based in part on these studies, OET added new Clinical Competence criteria to the Speaking subtest in 2018 with the goal of reflecting what health professionals and supervisors perceive as important to effective workplace communication. In order to evaluate whether language experts and health professionals perceived these criteria comparably, OET asked senior assessors and eight healthcare domain experts to use them to evaluate the same OET Speaking test recordings from test takers who had received a range of proficiency scores (Séguis & McElwee, 2019). Participants gave individual critera scores, an overall score, and chose a threshold recording of the "worst" performance that they deemed "minimally acceptable." The senior OET assessors and the healthcare professionals agreed strongly on all of these, lending initial support to Speaking scoring validity under the revised criteria (Séguis & McElwee, 2019).

Separately, Davidson recruited 18 medical professionals for standard-setting workshops and think aloud protocols to elicit what they find important in evaluating OET Writing samples (2022). Overall, their responses aligned with the OET Writing subtest criteria at the time of the research in early 2019. OET has since revised those five Writing criteria into a new set of six, based on CBLA's separate research program. OET's 2019 revision seeks to improve these Writing criteria that Davidson found already showed strong evidence of reflecting professionals' priorities when evaluating medical writing.

Questions

OET's efforts to continually review, revise, and improve belong in the "Support" section, above. However, because few studies are yet available assessing the 2018 and 2019 revisions to the Speaking and Writing criteria, the success of these revisions remains a question. Overall, having access to additional scoring and results details that OET keeps confidential would provide a further basis for discussing OET validity.

As described under "*Scores*," OET does not publish calculations for how it translates a test taker's individual criteria ratings into a final subtest score. This makes it difficult to evaluate these procedures. Nor does OET define a cut score for any of its sections, leaving this up to decision-makers who use the test. In 2023, OET published a table of reliability calculations for 2021 scores. Replicated in Table 3 below, the OET website report notes, "The reliability of the Listening and Reading sub-tests is reported using Cronbach Alpha, and the reliability of the Writing and speaking sub-tests is reported using Spearman Reliability" (CBLA, 2023). No further details are presented of how these reliability coefficients were developed, including methodology, sampling, or calculations.

Table 3

OET-reported subtest reliability for 2021 test results

Subtest	Overall Reliability
Listening	0.81
Reading	0.83
Writing	0.77
Speaking	0.79

In general, target reliability coefficients are 0.8 or higher for high stakes tests, and 0.9 for professionally-developed high stakes standardized tests such as OET. These OET-reported reliability coefficients for 2021 are on the low side of acceptability for a high stakes test, with Writing and Speaking scores falling below the minimum 0.8 threshold. Given that reliability is a prerequisite for validity, this raises important questions about OET as a professionally-developed standardized test.

In the Speaking subtest, there is also a question of score comparability because "Different role-plays are used for different candidates at the same test administration" (CBLA, n.d.). Though this is likely due to test security, no explanation is provided of why this choice has been

made or how scenarios are distributed. This differential application of prompts would be especially concerning if the test were more strictly norm-referenced, if candidates were scored in comparison to how others performed on the same day. Still, different candidates may receive Speaking prompts of varying difficulties, which would affect score comparability.

Regarding linguistic-specific skills only, CBLA publishes benchmarks of OET against CEFR and IELTS Academic bands as a reflection of general English language proficiency (see Table 1) (CBLA, n.d.). However, whether these are entirely comparable is unclear. In 2019, UK NARIC evaluated OET assessments and test taker samples against CEFR bands and benchmarked A/B/C OET ratings with CEFR levels C2/C1/B2, respectively (Coleman, 2019). CEFR may provide a more useful comparison than IELTS. Lim compared the scores of test takers who took both the OET and IELTS Academic in 2013, to benchmark the scores and calculate correlation. Correlations were only around 0.50 for most subtests, and even lower for Writing at 0.36 ($p<0.01$). Lim wrote in partial explanation, "As this data shows, the two are not entirely comparable, for entirely expected reasons" (2016), namely, that IELTS Academic and OET are targeted to different target language use contexts (general academia vs. working in a medical setting). Still, Lim's findings in part led OET to revise its scoring system to better align Reading and Writing B scores with IELTS, and to include another score option (inserting C+) to improve granularity of information about test takers in the B-C range (Lim, 2016). An updated correlation study would be instructive, especially after the 2018 and 2019 OET revisions.

While OET Writing and Speaking criteria were revised, there remains a risk of negative washback during test preparation for the Writing and Speaking sections especially. Test takers may achieve high scores by memorizing key phrases and structures that fit the criteria, which they may use during the test without complete understanding or command of the target ability (Séguis & McElwee, 2019). OET weighted the new Speaking criteria in part to minimize this, but analysis is needed to understand the effects and success of the weighting.

Further, there may remain opportunities to improve prompts and interlocutor training for Speaking. In 2016, Woodward-Kron & Elder compared OET Speaking roleplay recordings with the same test takers' performances on a different Australian test designed to evaluate medical competence only (the Objective Structured Clinical Examination, or OSCE). These researchers discussed threats to OET authenticity and reliability, such as an overly short time limit (five minutes), too little use of lay terms and over-use of formal language not reflective of real patient

speech, and the "unnaturally cooperative behavior" of the trained interlocutors (Woodward-Kron & Elder, 2016). They suggested revising the roleplay prompts and interlocutor training to bring the OET Speaking test more in line with observed OSCE performances. This study was part of OET's revision research program described under "*Support*," but it is unclear whether OET revised prompts or interlocutor training, and the Speaking roleplays still comprise five minutes.

Finally, one question raised by the Listening subtest is that materials are read in a range of accents, and test takers may encounter any of them, including British, American, Australian, or any other global English. It could also be argued (and OET does) that this supports validity, as it is intended to "reflect the global nature of the healthcare workforce" (CBLA, n.d.). Indeed, it is a healthcare reality that practitioners will encounter patients who speak any number of Englishes (or no English). Still, for the purposes of OET applicability, it might be relevant for test takers to hear selected accents depending on where they intend to practice, so decision-makers understand on which accent(s) a given candidate's Listening scores are based. Further, OET currently only includes accents representing these "inner circle" Englishes, which does not truly represent the full breadth of language that health care providers will encounter in practice. An open question is how to appropriately design a language test for caregivers who will need skills to communicate and negotiate meaning with patients from a range of English backgrounds and proficiency levels.

CONCLUSION

It is critically important that healthcare professionals are able to use language to safely perform their work. OET's emphasis on developing authentic materials and scoring methods in close collaboration with practicing healthcare professionals support its use as an ESP test in this high-stakes decision. OET's variety of item types, each based on real daily tasks of a working healthcare professional, mean test takers and decision-makers alike can expect to be familiar with the situations, texts, and scenarios being tested. From the perspective of decision-makers, an ESP test such as OET is assumed to be a better predictor of domain-specific English usage than a general English proficiency test because "neither command nor fluency of a standard language guarantees success in specific contexts such as medicine" (Hull, 2016).

While many of the published studies on OET were conducted by researchers at the University of Melbourne, the birthplace and ongoing primary contributor to the OET, their outright goal for conducting these research programs is to improve the test (rather than "simply" validating or supporting it). In general, OET's constant efforts to evaluate the test suggest its designers sincerely strive for continuous improvement. Evidence in support of OET validity include this continuous review to ensure the materials and tasks reflect the target language use domain. OET's research-based revisions focused on how test takers eventually perform and are perceived by others working in the real-world healthcare context. Still, in the absence of publicly available data, questions remain about reliability of OET scores across test takers and administrations. Specifically, new data and updated analyses are needed in order to evaluate whether and how well 2018 and 2019 revisions support the validity of using OET scores to predict test takers' communicative ability in English in real clinical workplaces.

REFERENCES

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (1st ed.). McGraw-Hill.

Carr, A. (2021). OET vs IELTS: Finding the Most Appropriate Way to Test Language Skills for Medicine. *ESP Today*, *9*(1), 89–106. https://doi.org/10.18485/esptoday.2021.9.1.5

CBLA. (n.d.). *OET - The English Language Test for Healthcare Professionals.* OET - Occupational English Test. Retrieved February 5, 2022, from https://www.occupationalenglishtest.org/

CBLA. (2020). *Factsheet: Information for OET candidates (V4.0).* Retrieved July 12, 2023 from https://prod-wp-content.occupationalenglishtest.org/resources/uploads/2020/01/24145449/FAC_INF_060819.pdf

CBLA. (2023). *Test statistics: Test performance 2021*. Retrieved July 11, 2023 from https://www.oet.com/discover/about-oet/test-statistics

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21–33). Routledge.

Coleman, M. (2019, July 17). *New report compares OET to international standard of language ability*. OET - Occupational English Test. https://www.occupationalenglishtest.org/oet-benchmarked-to-cefr/

Davidson, S. (2022). The domain expert perspective: A qualitative study into the views expressed in a standard-setting exercise on a language for specific purposes (LSP) test for health professionals. *Language Testing, 39*(1), 117–141. https://doi.org/10.1177/02655322211010737

Educational Commission for Foreign Medical Graduates. (n.d.). *ECFMG Requirements For Certification*. Retrieved February 13, 2022, from https://www.ecfmg.org/certification/requirements-for-certification.html

Elder, C., Pill, J., Woodward-Kron, R., McNamara, T., Manias, E., Webb, G., & McColl, G. (2012). Health Professionals' Views of Communication: Implications for Assessing Performance on a Health-Specific English Language Test. *TESOL Quarterly, 46*(2), 409–419. https://doi.org/10.1002/tesq.26

Hull, M. (2016). Medical language proficiency: A discussion of interprofessional language competencies and potential for patient risk. *International Journal of Nursing Studies, 54*, 158–172. https://doi.org/10.1016/j.ijnurstu.2015.02.015

Lim, G. S. (2016). *The Occupational English Test and IELTS: A Benchmarking Report* (p. 9). Cambridge Assessment English. occupationalenglishtest.org/research

Manias, E., & McNamara, T. (2016). Standard setting in specific-purpose language testing: What can a qualitative study add? *Language Testing*, *33*(2), 235–249. https://doi.org/10.1177/0265532215608411

McNamara, T. (1996). *Measuring second language performance*. Longman.

McNamara, T., Elder, C., Flynn, E., Knoch, U., Manias, E., Woodward-Kron, R., & Yahalom, S. (2018). Cross-disciplinary collaboration in research on a specific-purpose language test in the healthcare setting. *Journal of Applied Linguistics and Professional Practice, 13*(1–3), 189–210. https://doi.org/10.1558/japl.31857

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: The role of measurement*. Oxford University Press.

Pill, J. (2016). Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing, 33*(2), 175–193. https://doi.org/10.1177/0265532215607400

Pill, J., & McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing, 33*(2), 217–234. https://doi.org/10.1177/0265532215607402

Qian, D. D., & Pan, M. (2013). Response Formats. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 860–875). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla090

Séguis, B., & McElwee, S. (2019). Assessing Clinical Communication On The Occupational English Test. In S. Papageorgiou & K. M. Bailey (Eds.), *Global Perspectives on Language Assessment* (1st ed., pp. 63–79). Routledge. https://doi.org/10.4324/9780429437922-5

The Official Guide to OET (1st ed.). (2018). Kaplan Publishing.

Vidaković, I., & Khalifa, H. (2013). Stakeholders' perceptions of the Occupational English Test (OET): An exploratory study. *R E S E A R C H, 54*, 4.

Woodward-Kron, R., & Elder, C. (2016). A comparative discourse study of simulated clinical roleplays in two assessment contexts: Validating a specific-purpose language test. *Language Testing, 33*(2), 251–270. https://doi.org/10.1177/0265532215607399