Title of Project:
How Valid are Domain Experts' Judgments of Workplace Communication? Implications for Setting Standards on the Occupational English Test (OET) Writing Sub-Test

Researcher:
Simon Davidson
University of Melbourne
s.davidson3@student.unimelb.edu.au

Research Supervisors:
Dr. Ute Knoch
University of British Columbia

Assoc. Prof. Cathie Elder
University of Melbourne


Simon Davidson

Project Summary

As part of the requirement to attain professional registration and be employed in Australia, International Medical Graduates (IMGs) need to satisfy a number of prerequisites that include demonstrating satisfactory English language proficiency and competent clinical skills/knowledge. Some concern has been noted that the minimum standards on tests used to evaluate language proficiency may be inadequate for proving effective workplace readiness. To better comprehend the validity of these apprehensions, this study examined the use of domain expert judges in determining performance standards for a specific purpose language (LSP) test: the Occupational English Test (OET). The OET is used to evaluate the English language writing, speaking, reading and listening abilities of health professionals; in this research project they are medical doctors who received their qualifications outside of Australia in non-English speaking contexts. The objective of the OET test is to establish health professionals' abilities to communicate in English with other health professionals and patients in specific occupational related contexts. The writing task on the OET test requires test-takers to write is a letter of referral that is based on a set of case notes that have been provided. The OET writing sub-test's minimum standards were set via a procedure that elicited insights from suitable medical practitioners who had experience with relevant workplace communication demands and could determine levels of competency regarded as acceptable for this specific use. Setting suitable performance standards on such an assessment is of vital importance. If the standards are set too low, the well-being and safety of patients may be endangered by health professionals who aren't able to communicate effectively. However, if the standards are set too high, communicatively competent test candidates could be deprived of certification and, consequently, not be allowed to work in regions where their skills/knowledge are greatly needed.

The purpose of the study was to ascertain the minimum levels of English language written proficiency that were deemed suitable by domain experts for effective performance in the workplace. Another consideration was to take into account the extent that health professionals, without linguistic training, are able to attend to issues of language and communication independently of clinical competence,

which is their area of expertise. What was the basis for the decisions made by subject-matter experts (i.e., medical doctors) and how strongly did these underlying reasons parallel the communication construct that the OET is intended to measure? A previous study (Manias & McNamara, 2016; Pill & McNamara, 2016) investigated these issues in regard to the OET speaking subtest, while the current study concentrated on the OET writing subtest, which is to date underexplored area. This doctoral research is an independent part of a larger validation study of the OET.

The study was designed to answer four research questions:

1. What criteria do standard-setting participants use as a basis for their decisions in judging writing responses and to what extent are these decisions language-based?
2. Is there any variability between judges in what they attend to while setting standards?
3. How do standard-setting judges view the process and outcome of the standard-setting procedure?
4. What occupational specific standards (cut scores) do doctors set on the Occupational English Test (OET) writing sub-test?

To explore answers to the research questions, 18 health professionals (GPs, specialists and medical educators) were recruited to participate in standard-setting workshops, which were intended to elicit judgements about particular performance level allocations for the task. In addition, reasons and features for why participants deemed a writing sample to be eligible for a passing grade were investigated. The study utilized a mixed approach to data collection, using both qualitative and quantitative methods. A pilot study was also performed to choose a suitable standard setting procedure from two potential methods. To obtain additional understanding of the foundation of the standards being established, a sample of participants completed verbal reports in the form of a think-aloud protocol (TAP) was elicited. The health professionals' observations from the workshops and verbal reports were coded thematically and inter-coder reliability checks were carried out. The qualitative analysis revealed that the participants took into account seven main features of the writing samples in deciding on their judgments of performance levels: task fulfillment, content, organization, expression, presentation, professionalism, audience recognition, and a separate category of 'other' for aspects that could not be accommodated in the above themes. The study focused on not only the establishment of new standards for the OET writing sub-test but also on (1) the effectiveness and validity of the method itself, (2) the participants' comprehension of and confidence in the procedure, (3) panellists' consistency in the method's used, and (4) the "indigenous assessment criteria" (i.e., the values that underlie everyday judgments of performance by domain experts in actual workplace situations (Jacoby & McNamara, 1999) that participants used in assessing the characteristics of the writing samples.

Prior to new passing standards and the "cut scores" being calculated, a many facet Rasch analysis using FACETS software (Linacre, 2017) was undertaken to consider any variation in subject-matter experts' performance level judgments regarding their being overly severe, lenient, or inconsistent. The subsequent quantitative analysis generated a slightly more severe passing standard than the existing one. This result paralleled the findings of similar standard-setting studies on the OET using the Analytic Judgment method (AJM) (e.g., Knoch, Elder, Woodward-Kron, Flynn, Manias, McNamara, Zhang Ying & Huisman, 2017; Pill & McNamara, 2016). The new standards were contrasted with present OET cut scores, which showed a higher "fail" proportion than the existing data set. The more stringent passing standard founded by subject-matter experts in this study could be interpreted as supporting suggestions

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

and views that the current standard is not high enough and that some IMGs who are not yet communicatively proficient are, nevertheless, being employed in Australian work environments with inadequate written communication skills.

The qualitative analysis additionally examined whether subject-matter experts are capable of judging language ability independently from other professional skills/knowledge as is required by Australian federal government conditions. Some domain experts' decisions (a minority on the whole) were swayed by assessments of test candidates' clinical competency, which is outside the construct of communicative competence as circumscribed by the OET. Yet overall, the qualitative results indicated that domain experts were undeniably focusing on textual features associated with what the OET is meant to assess. The fundamental consideration of whether subject-matter experts, without language assessment training, are well placed for establishing the standards in a language for specific purposes (LSP) test such as the OET was taken into account. Despite the fact that some participants' decisions departed to some degree from the present OET writing sub-test criteria, validity confirmation gathered in this study confirmed, for the most part, that the ensuing new standards, resulting from subject-matter expert involvement, were acceptable. The validity outcomes of this study's findings for the OET writing sub-test, and for LSP testing more generally, were considered by means of a new argument-based validity framework created by Knoch and Macqueen (in preparation).

The most noteworthy features of this study are that first, a formal standard-setting procedure has been carried out with domain expert involvement on the OET writing sub-test, giving rise to a methodically developed range of cut scores for explaining test candidate performance. In addition, further evidence has been gathered concerning how well non-linguistically trained subject-matter experts are able to assess language performance and address facets of language and communication independently (or not) of their evaluations of professional competency. In this regard, the findings have offered new confirmation for the case that some language testers have proposed that language knowledge and content knowledge in LSP performances are inseparable. Furthermore, in keeping with contemporary approaches to test validation, the study has framed the standard-setting discussion in the context of a validity argument which has not been performed before with an LSP test. Another practical contribution of this study is that it has suggested more justifiable and defensible performance standards for the OET writing sub-test than those that are presently being used as the origin of the current OET standards has not been publicly available or been accessible to scrutiny for some time.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

References

Alderson, J. C., Candlin, C. N., Clapham, C. M., Martin, D. J., & Weir, C. J. (1986). *Language proficiency testing for migrant professionals: New directions for the Occupational English Test. A report submitted to the Council on Overseas Professional Qualifications*. Lancaster: Institute for English Language Education, University of Lancaster.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*(1), 1-16.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1988). *Validity: An evolving concept*. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Australian Government Department of Health. (2016). *District of Workforce Shortage*. Canberra: Commonwealth of Australia.

Australian Government Department of Health and Ageing. (2008). *Report on the Audit of Health Workforce in Rural and Regional Australia*. Canberra, Australia: Commonwealth of Australia.

Australian Government Department of Health and Ageing. (2011). *Submission to the Inquiry into Registration Procedures and Support for Overseas Trained Doctors.* Canberra, Australia: Commonwealth of Australia.

Australian Health Practitioner Registration Agency (AHPRA). (2017). Overseas practitioners. Retrieved from https://www.ahpra.gov.au/Registration/Registration-Process/Overseas-Practitioners.aspx.

Australia Medical Association (AMA) (2004). AMA position statement - overseas trained doctors. Retrieved from https://ama.com.au/sites/default/files/documents/OTDPositionStatement2004.pdf.

Australian Medical Council (AMC). (2017). Assessment pathways to registration for international medical graduates. Retrieved from http://www.amc.org.au/assessment/pathways.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1-34.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Purpura, J. E. (2008). 32 language assessments: Gate-keepers or door-openers?. *The Handbook of Educational Linguistics*, *456-468*.

Bahry, L., Hambleton, R.K., Gotzmann, A., De Champlain, A., & Roy, M. (2012). *National Assessment Collaboration Standard Setting Study Report - Technical Report for the Technical Advisory Committee and National Assessment Collaboration Examination Committee*. Retrieved from http://mcc.ca/wp-content/uploads/Technical-Reports-Bahry-2012.pdf.

Baker, D., & Robson, J. (2012). Communication training for international graduates. *The Clinical Teacher*, *9*(5), 325-329.

Barton, D., Hawthorne, L., Singh, B., & Little, J. (2003). Victoria's dependence on overseas trained doctors in psychiatry. *People and Place*, *11*(1), 54-64.

Basturkmen, H., & Elder, C. (2004). The practice of LSP. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp. 672-694). Malden, MA: Blackwell.

Bejar, I. I., Braun, H., & Tannenbaum, R. J. (2007). A prospective, predictive and progressive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1-30). Maple Grove, MN: Jam Press.

Berendsen, A. J., Kuiken, A., Benneker, W. H., Meyboom-de Jong, B., Voorn, T. B., & Schuling, J. (2009). How do general practitioners and specialists value their mutual communication? A survey. *BMC Health Services Research*, *9*(1), 143.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, *15*, 4-9.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*, 137-172.

Berry, V., O'Sullivan, B., & Rugea, S. (2013). *Identifying the appropriate IELTS score levels for IMG applicants to the GMC register - Report submitted to the General Medical Council*. Centre for Language Assessment Research (CLARe), The University of Roehampton, London. Retrieved from http://www.gmc-uk.org/Identifying_the_appropriate_IELTS_score_levels_for_IMG_applicants_to_the_GMC_register.pdf_55197989.pdf.

Birrell, B. (2011). *Australia's new health crisis - too many doctors.* Melbourne, Australia: Centre for Population and Urban Research, Monash University.

Birrell, B., & Schwartz, A. (2006). Accreditation of overseas trained doctors: The continuing crisis. *People and Place*, *14*(3), 37.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Abingdon, VA: Routledge.

Borello, E. (2016, August 9). Calls for Government to stop giving visas to overseas-trained doctors to address rural shortage. *ABC News*. Retrieved from http://www.abc.net.au/news/2016-08-09/calls-to-stop-giving-overseas-trained-doctors-visas/7706612.

Bracewell, R. J., & Breuleux, A. (1994). Substance and romance in analysing think-aloud protocols. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 55-88). Thousand Oaks, CA: Sage.

Brady, H., Collier, D. & Seawright, J. (2010), Refocusing the discussion of methodology. In H. Brady & Collier D. (Eds.), *Rethinking social inquiry: Diverse tools, shared standards* (pp. 16-31). New York, NY: Rowman and Littlefield.

Brewer, J., & Hunter, A. (1989). *Multimethod research: A synthesis of styles*. Thousand Oaks, CA: Sage.

Brice, C. (2005). Coding data in qualitative research on L2 writing: Issues and implications. In P. K. Matsuda & T. Silva (Eds.), *Second language writing research: Perspectives on the process of knowledge construction* (pp. 159-175). Mahwah, NJ: Lawrence Erlbaum.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation and understanding* (pp. 65-116). Hillsdale, NJ: Lawrence Erlbaum.

Brown, A. (1993). LSP testing: The role of linguistic and real-world criteria. *Melbourne Papers in Language Testing*, *2*(2), 35-54.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Brown, J. D. (2013). Cut Scores on Language Tests. *The encyclopedia of applied linguistics*, Oxford, UK: Blackwell Publishing.

Buckendahl, C. W. (2005). Guest editor's introduction: Qualitative inquiries of participants' experiences with standard setting. *Applied Measurement in Education*, *18*(3), 219-221.

Canadian English Language Benchmark Assessment for Nurses (CELBAN). (2015). CELBAN Overview. Retrieved from http://www.celban.org/celban/display_page.asp?page_id=3.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). London, UK: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47.

Carver, P. (2008). *Self-sufficiency and international medical graduates - Australia*. Melbourne, Australia: National Health Workforce Taskforce.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369-383.

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.

Chapelle, C. (2011). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.) *Building a validity argument for the test of teaching English as a foreign language* (pp. 319-352). New York, NY: Routledge.

Chapelle, C. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.). *The Routledge handbook of language testing* (pp. 21-33). New York, NY: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference?. *Educational Measurement: Issues and Practice*, *29*(1), 3-13.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chur-Hansen, A. (1997). Language background, proficiency in English, and selection for language development. *Medical Education*, *31*(5), 312-319.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106.

Cizek, G. J. (1996). An NCME instructional module on: Setting passing scores. *Educational Measurement: Issues and Practice*, *15*(2), 20-31.

Cizek, G. J. (Ed.), (2001). *Standard setting: Concepts, methods and perspectives*. Mahwah, NJ: Erlbaum.

Cizek, G. J. (2012). An introduction to contemporary standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 3-14). New York, NY: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). The body of work and other holistic methods. In G. J. Cizek & M. B. Bunch (Eds.), *Standard setting: A guide to establishing and evaluating performance standards on tests* (pp. 117-153). London, UK: Sage.

Cohen, A. D. (1987). Using verbal reports in research on language learning. In C. Faerch & G. Kasper (Eds.), *Introspection in second language learning research* (pp. 82-95). Clevedon, UK: Multilingual Matters.

Cohen, A. D. (1988). The use of verbal report data for a better understanding of test taking processes. *ARAL, 11*(2), 30-42.

Cohen, A. D. (1996). Verbal reports as a source of insight into second language learner strategies. *Applied Language Learning, 7*(1), 7-26.

Cohen, A. D. (2000). Exploring strategies in test-taking: Fine-tuning verbal reports from respondents. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. 127-150). Mahwah, NJ: Lawrence Erlbaum.

Collier, D., Brady, H., & Seawright, J. (2010). Introduction to the second edition: A sea change in political methodology. In H. Brady & D. Collier (Eds.), *Rethinking social inquiry: Diverse tools, shared standards* (pp. 1-10). New York, NY: Rowman & Littlefield.

Cooper, M., & Holtzman, M. (1983). Talking about protocols. *College Composition and Communication, 34*, 284-293.

Council of Europe (2011). *Common European Framework of Reference for Languages: Learning, teaching and assessment (CEFR).* Retrieved from http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf.

Couser, G. (2007). Twelve tips for developing training programs for international medical graduates. *Medical Teacher*, *29*(5), 427-430.

Crawford, T., & Candlin, S. (2013). A literature review of the language needs of nursing students who have English as a second/other language and the effectiveness of English language support programmes. *Nurse Education in Practice*, *13*(3), 181-185.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. *Handbook of mixed methods in social and behavioral research*. Los Angeles, CA: Sage.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Los Angeles, CA: Sage.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement*. (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I Braun (Eds.), *Test validity*. (pp. 3-19). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education*, *3*(3), 265-286.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*, 67-96.

Cusimano, M. D. (1996). Standard setting in medical education. *Academic Medicine*, *71*(10), S112-20.

Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing*, *1*(1), 50-69.

Davies, A. (1995). Testing communicative language or testing language communicatively: What? How? *Melbourne Papers in Language Testing*, *4*(1), 1–20.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18*(2), 133-147.

Dawber, T., Lewis, D. M., & Rogers, W. T. (2002). *The cognitive experience of bookmark standard setting participants*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Denzin, N. K., & Lincoln, Y.S. (1998). *Strategies of qualitative enquiry*. Thousand Oaks, CA: Sage.

DeRemer, M. L. (1998). Writing assessment: raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.

Donnelly, L. (2017, August 22). Revealed: Health chiefs plan to make it easier for failing trainees and foreign medics to become GPs. *The Telegraph*. Retrieved from http://www.telegraph.co.uk/news/2017/08/22/revealed-health-chiefs-plan-make-easier-failing-trainees-foreign/.

Dorgan, K. A., Lang, F., Floyd, M., & Kemp, E. (2009). International medical graduate-patient communication: A qualitative analysis of perceived barriers. *Academic Medicine*, *84*(11), 1567-1575.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, UK: Cambridge University Press.

Douglas, D. (2001a). Language for specific purposes assessment criteria: Where do they come from?. *Language Testing*, *18*(2), 171-185.

Douglas, D. (2001b). Three problems in testing language for specific purposes: Authenticity, specificity and inseparability. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. F. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 45-52). Cambridge, UK: Cambridge University Press.

Douglas, D. (2004). Discourse domains: The cognitive context of speaking. *Studying Speaking to Inform Second Language Learning*, *8*, 25-47.

Douglas, D. (2005). Testing languages for specific purposes. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 857–868). Mahwah, NJ: Lawrence Erlbaum.

Douglas, D. (2010). This won't hurt a bit: Assessing English for nursing. *Taiwan International ESP Journal*, *2*(2), 1-16.

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria. In *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60-81). Cambridge, UK: Cambridge University Press.

Douglas, D., & Selinker, L. (1985a). Principles for language tests within the 'discourse domains' theory of interlanguage: Research, test construction and interpretation. *Language Testing*, *2*(2), 205-226.

Douglas, D., & Selinker, L. (1985b). The problem of comparing episodes in discourse domains in interlanguage studies. Retrieved from http://eric.ed.gov/?id=ED308710

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System*, *20*(3), 317-328.

Douglas, D., & Selinker, L. (1994). Research methodology in context-based second-language research. In E. E. Tarone, S. M. Gass & A. D. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 119-131). Hillsdale, NJ: Lawrence Erlbaum.

Duncan, G. F., & Gilbey, D. (2007). Cultural and communication awareness for general practice registrars who are international medical graduates: A project of Coast City Country Training. *Australian Journal of Rural Health*, *15*(1), 52-58.

Dunlea, J., & Matsudaira, T. (2009). Investigating the relationship between the EIKEN tests and the CEFR. *Linking to the CEFR levels: Research perspectives.* Arnhem, The Netherlands: CITO and EALTA.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Frankfurt, Germany: Peter Lang.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing*, *10*(3), 235-254.

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, *18*(2), 149-170.

Elder, C., & McNamara, T. (2016). The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing*, *33*(2), 153-174.

Elder, C., McNamara, T., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us. *Language & Communication*, *57*, 14-21.

Elder, C., McNamara, T., Woodward-Kron, R., Manias, E., Webb, G., McColl, G., & Pill, J. (2013). *Towards improved healthcare communication: Development and validation of language proficiency standards for non-native English-speaking health professionals – Final report for the Occupational English Test Centre*. Retrieved from

http://www.occupationalenglishtest.org/Documents/ViewDocument.aspx?club=oet&DocumentID=eff97948-f338-40fb-8817-e1ca763f9f99.

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.

Elkin, K., Spittal, M. J., & Studdert, D. M. (2012). Risks of complaints and adverse disciplinary findings against international medical graduates in Victoria and Western Australia. *Medical Journal of Australia*, *197*(8), 448.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1987). Verbal reports on thinking. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 24-53). Clevedon, UK: Multilingual Matters.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panellists in a standard-setting study. *Applied Measurement in Education*, *18*(3), 257-267.

Forman, J., & Damschroder, L. (2008). Qualitative content analysis. In L. Jacoby & L. A. Siminoff (Eds.), *Empirical methods for bioethics: A primer* (pp. 39-62). Oxford: Elsevier.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375.

Francois, J. (2012). Improving family medicine residents' written communication. *Canadian Medical Education Journal*, *3*(1), e64-e68.

Freelon, D. (2010). ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, *5*(1), 20-33.

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System*, *28*(4), 483-497.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29.

Garling, P. (2008). *Acute care services in NSW public hospitals*. Sydney, Australia: NSW Government.

Giraud, G., & Impara, J. C. (2005). Making the cut: The cut score setting process in a public school district. *Applied Measurement in Education*, *18*(3), 289-312.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education*, *18*(3), 223-232.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education, 12*(1), 13-28.

Grant-Davie, K. (1992). Coding data: issues of validity, reliability, and interpretation. In G. E. Kirsch & P. A. Sullivan (Eds.), *Methods and methodology in composition research*. (pp. 270-286). Carbondale, IL: Southern Illinois University.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook* (Vol. 5). Cambridge, UK: Cambridge University Press.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255-274.

Hall, P., Keely, E., Dojeiji, S., Byszewski, A., & Marks, M. (2004). Communication skills, cultural challenges and individual support: challenges of international medical graduates in a Canadian healthcare environment. *Medical Teacher*, *26*(2), 120-125.

Hambleton, R., Jaeger, R., Plake, B., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*(4), 355-366.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th Ed., pp. 433-470). Westport, CT: Praeger.

Hamp-Lyons, L., & Lumley, T. (2001). Assessing language for specific purposes. *Language Testing*, *18*(2), 127-132.

Harding, C., Parajuli, N., Johnston, L., & Pilotto, L. (2010). Comparing patients' perceptions of IMGs and local Australian graduates in rural general practice. *Australia Family Physician, 39*(4).

Harris, D. P. (1969). *Testing English as a second language*. New York, NY: McGraw-Hill.

Harvey, A. L., & Way, W. D. (1999). A comparison of web-based standard setting and monitored standard setting. Retrieved from http://eric.ed.gov/?id=ED463747.

Hawthorne, L. (2012). International medical migration: What is the future for Australia? Retrieved from https://www.mja.com.au/system/files/issues/001_03_230712/haw10088C_fm.pdf.

Heaton, J. B. (1975). *Writing English language tests: A practical guide for teachers of English as a second or foreign language*. London, UK: Longman.

Hein, S. F., & Skaggs, G. E. (2009). A qualitative investigation of panellists' experiences of standard setting using two variations of the bookmark method. *Applied Measurement in Education*, *22*(3), 207-228.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

Hein, S. F., & Skaggs, G. E. (2010). Conceptualizing the classroom of target students: A qualitative investigation of panellists' experiences during standard setting. *Educational Measurement: Issues and Practice*, *29*(2), 36-44.

Henn, M., Weinstein, M., & Foard, N. (2009). *A critical introduction to social research*. Los Angeles, CA: Sage.

House, E. R. (1980). *Evaluating with validity*. Los Angeles, CA: Sage.

House of Representatives Standing Committee on Health and Ageing. (2012). *Lost in the labyrinth: Report on the inquiry into registration processes and support for overseas trained doctors.* Canberra: Parliament of the Commonwealth of Australia. Retrieved from http://www.aph.gov.au/Parliamentary_Business/Committees/House_of_Representatives_Committees?url=haa/overseasdoctors/index.htm.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*(2), 219-230.

Hughes, J., & Parkes, S. (2003). Trends in the use of verbal protocol analysis in software engineering research. *Behaviour & Information Technology*, *22*(2), 127-140.

Hulstijn, J. H. (2011). Language proficiency in native and non-native speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249.

Hunt, K. (1965). *Grammatical structures written at three grade levels*. NCTE research report No. 3. Champaign, IL: NCTE.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-236). Creskill, NJ: Hampton Press.

Hycner, R. H. (1985). Some guidelines for the phenomenological analysis of interview data. *Human Studies*, *8*(3), 279-303.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth, UK: Penguin.

Ingram. D. E. (1984). *Report on the formal trialing of the Australian Second Language Proficiency Ratings (ASLPR).* Canberra, Australia: Australian Government Publishing Service.

International Civil Aviation Organization (ICAO) (2017). ICAO / Safety / Language Proficiency Requirements (LPR). Retrieved from https://www.icao.int/safety/lpr/Pages/Language-Proficiency-Requirements.aspx.

Impara, J. C., & Plake, B. S. (1998). Participants' ability to estimate item difficulty: A test of the assumptions of the Angoff standard-setting method. *Journal of Educational Measurement, 35*(1), 69-81.

Jacoby, S. (1998). *Science as performance: Socializing scientific discourse through the conference talk rehearsal*. Los Angeles, CA: University of California.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction*, *28*(3), 171-183.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, *18*(3), 213-241.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). Washington, DC: American Council on Education.

Jaeger, R. M., & Mills, C. N. (2001). An integrated judgement procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods and perspectives* (pp. 313-318). Mahwah, NJ: Erlbaum.

Jourdenais, R. (2001). Cognition, instruction and protocol analysis. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 354-375). Cambridge, UK: Cambridge University Press.

Kaftandjieva, F. (2004). *Standard setting, Section B in the Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning, teaching and assessment.* Strasbourg, France: Council of Europe. Retrieved from http://www.coe.int/T/DG4/Portfolio/?L=E&M=/documents intro/Reference supplement intr.html.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527-535.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement 38*, 319-42.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31-41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective*, *2*(3), 135-170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Greenwood Publishing.

Kane, M. T. (2012a). Articulating a validity argument. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 34-47). New York, NY: Routledge.

Kane, M. T. (2012b). Validating score interpretations and uses. *Language Testing*, *29*, 3-17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M., Crooks, T., & Cohen, A. (1999a). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*, 5-17.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999b). Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education*, *4*(3), 195-207.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4-12.

Katz, I. R., Tannenbaum, R. J., & Kannan, P. (2009). Virtual standard setting. *CLEAR Exam Review*, *20*(2), 19-27.

Katz, I. R., & Tannenbaum, R. J. (2014). Comparison of web-based and face-to-face standard setting using the Angoff method. *Association of Test Publishers*, *15*(1), 1-17.

Keely, E., Myers, K., Dojeiji, S., & Campbell, C. (2007). Peer assessment of outpatient consultation letters-feasibility and satisfaction. *BMC Medical Education*, *7*(1), 13.

Kenyon, D., & Römhild, A. (2014). Standard setting in language testing. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 1-18). Hoboken, NJ: John Wiley & Sons.

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, I. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.

Kingston, N. M., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 201-223). New York, NY: Routledge.

Knoch, U. (2009). Collaborating with ESP stakeholders in rating scale validation: The case of the ICAO rating scale. *SPAAN FELLOW*, *1001*, 21.

Knoch, U. (2014). Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes*, *33*, 77-86.

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment*, *2*(2), 48-66.

Knoch, U., Elder, C., Woodward-Kron, R., Flynn, E., Manias, E., McNamara, T., Zhang Ying, B., & Huisman, A. (2017). *Towards improved quality of written patient records: Development and validation of language proficiency standards for writing for non-native English-speaking health professionals - Final Report for Cambridge Boxhill Language Assessment*. Retrieved from

http://arts.unimelb.edu.au/__data/assets/pdf_file/0008/2546585/LP130100171-Final-Report-1-November-2017.pdf.

Knoch, U., & Macqueen, S. (In preparation). *Assessing English for professional purposes: Language and the workplace*. New York, NY: Routledge.

Knoch, U., & McNamara, T.F. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 277-304). New York, NY: Routledge.

Konno, R. (2006). Support for overseas qualified nurses in adjusting to Australian nursing practice: A systematic review. *International Journal of Evidence Based Healthcare*, *4*(2), 83-100.

Lado, R. (1961). *Language testing: The Construction and use of foreign language tests: A teacher's book.* Inglaterra, UK: Longmans, Green and Company.

Lee, Y-W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT Essays: Scores from humans and e-rater (TOEFL Research Report RR-81)*. Princeton, NJ: Educational Testing Service.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). The bookmark standard-setting procedure: Methodology and recent implementations. In *Annual meeting of the National Council on Measurement in Education, San Diego, CA*.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In *DR Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment. Phoenix, AZ*.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard-setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225-253). New York, NY: Routledge.

Linacre, J. M. (2017). Facets computer program for many-facet Rasch measurement, version 3.8.0. Beaverton, OR: Winsteps.com.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2010). Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Retrieved from https://www.researchgate.net/profile/Cheryl_Bracken/publication/242785900_Practical_Resources_for_Assessing_and_Reporting_Intercoder_Reliability_in_Content_Analysis_Research_Projects/links/0deec52e14791a0d6f000000.pdf.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and* perspectives (2nd ed., pp. 175-218). Mahwah NJ: Erlbaum.

Louis, W. R., Lalonde, R. N., & Esses, V. M. (2010). Bias against foreign-born or foreign-trained doctors: Experimental evidence. *Medical Education*, *44*(12), 1241-1247.

Lumley, T. (1995). The judgements of language-trained raters and doctors in a test of English for health professionals. *Melbourne Papers in Language Testing*, *4*(1), 74-98.

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes*, *17*(4), 347-367.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York, NY: Peter Lang.

Lumley, T., & Brown, A. (1996). Specific-purpose language performance tests. *Australian Review of Applied Linguistics, Supplement Series*, *13*(1), 105-136.

Lumley, T., Lynch, B. K., & McNamara, T. (1994). A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing*, *3*(2), 19-40.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*, 425-44.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Macqueen, S., Yahalom, S., Kim, H., & Knoch, U. (2012). *Exploring writing demands in healthcare settings.* Melbourne, Australia: Language Testing Research Centre, University of Melbourne.

Manias, E., Jorm, C., & White, S. (2008). How is patient care transferred safely? In C. Jorm (Ed.), *Windows into safety and quality in health care* (pp. 37-48). Sydney, Australia: Australian Commission on Safety & Quality in Healthcare.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher*, *42*, 78-88.

McDonnell, L., & Usherwood, T. (2008). International medical graduates – challenges faced in the Australian training program. *Australian Family Physician, 37*, 481-484.

McGinty, D. (2005). Illuminating the" black box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education*, *18*(3), 269-287.

McGrath, B. P. (2004). Integration of overseas-trained doctors into the Australian medical workforce. *Medical Journal of Australia*, *181*(11/12), 640-642.

McGrath, P., Henderson, D., & Holewa, H. (2012). Language issues: an important professional practice dimension for Australian International Medical Graduates. *Communication & Medicine*, *10*(3), 191-200.

McNamara, T. F. (1990). *Assessing the second language proficiency of health professionals* (Unpublished doctoral dissertation). University of Melbourne, Melbourne, Australia.

McNamara, T. F. (1996). *Measuring second language performance.* London, UK: Addison Wesley Longman.

McNamara, T. F. (1997a). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446–466.

McNamara, T. F. (1997b). Problematising content validity: The Occupational English Test (OET) as a measure of medical communication. *Melbourne Papers in Language Testing*, *6*(1), 19-43.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

Medical Board of Australia. (2014). Good medical practice: A code of conduct for doctors in Australia. Retrieved from http://www.medicalboard.gov.au/Codes-Guidelines- Policies.aspx.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: American Council on Education and Macmillan.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 92-114). Cambridge, UK: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229-258.

Move to swap foreign doctors for local (2017, February 18). *SBS News*. Retrieved from https://www.sbs.com.au/news/move-to-swap-foreign-doctors-for-local

Mullan, F. (2005). The metrics of the physician brain drain. *The New England Journal of Medicine*, *353*(17), 1810-1818.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*(1), 3-19.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Occupational English Test (OET). (2009). OET Info Booklet 2009. Retrieved from https://www.ahpra.gov.au/documents/default.aspx?record=WD10%2F4003&dbid=AP&chksum=axNx1hDhgzMvaMcq8mAtZA%3D%3D

Occupational English Test (OET). (2017). Occupational English Test. Retrieved from http://www.occupationalenglishtest.org/.

Oller, J. W. (2012). Grounding the argument-based framework for validating score interpretations and uses. *Language Testing*, *29*(1), 29-36.

O'Loughlin, K. (2008). Assessment at the workplace. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7: Language testing and assessment, pp. 69-80). New York, NY: Springer.

O'Neill, T. R. (2004). The minimum proficiency in English for entry-level nurses. Retrieved from http://www.ncsbn.org/pdfs/TOEFL_Research_Brief_vol_1.pdf.

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly, 4*(4), 295-317.

O'Neill, T. R., Marks, C., & Wendt, A. (2005). Recommending a minimum English proficiency standard for entry-level nursing. *JONA's Healthcare Law, Ethics, and Regulation, 7*(2), 56-58.

O'Sullivan, B. (2012). Assessment issues in languages for specific purposes. *The Modern Language Journal*, *96*(s1), 71-88.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation*, *35*(2-3), 95-101.

Papageorgiou, S. (2007). *Relating the Trinity College London GESE and ISE exams to the Common European Framework of Reference: Piloting of the Council of Europe draft Manual (Final project report*). Lancaster: Lancaster University.

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly*, *13*(2), 109-123.

Parnell, S. (2016, August 9). Visa plan to stop foreign doctor influx. *The Australian*. Retrieved from http://www.theaustralian.com.au/national-affairs/health/visa-plan-to-stop-foreign-doctor-influx/news-story/67a9915f4c258f360875785499d3975c.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

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.

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.

Pill, J., & Woodward-Kron, R. (2012). How professionally relevant can language tests be? A response to Wette (2011). *Language Assessment Quarterly*, *9*(1), 105-108.

Pill, T. J. H. (2013). What doctors value in consultations and the implications for specific-purpose language testing (Unpublished doctoral dissertation). University of Melbourne, Melbourne, Australia.

Pilotto, L. S., Duncan, G. F., & Anderson-Wurf, J. (2007). Issues for clinicians training international medical graduates: A systematic review. *Medical Journal of Australia, 187*(4), 225-228.

Piterman, L., & Koritsas, S. (2005). Part II. General practitioner-specialist referral process. *Internal Medicine Journal*, *35*(8), 491-496.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181-199). New York, NY: Routledge.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgement method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.

Plake, B. S., & Impara, J. C. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*, 353–366.

Playford, D. E., & Maley, M. A. L. (2008). Medical teaching in rural Australia: Should we be concerned about the international medical graduate connection? *Medical Journal of Australia*, *189*(2), 125-127.

Purpura, J. E. (2008). Assessing communicative language ability: Models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7: Language Testing and Assessment, pp. 53-68). New York, NY: Springer.

Qian, H., Woo, A., & Banerjee, J. (2014). *Setting an English Language Proficiency Passing Standard for Entry-Level Nursing Practice Using the Michigan English Language Assessment Battery*. Retrieved from https://www.ncsbn.org/14_NCLEX_technicalbrief_SettinganEnglishLanguageProficiencyPassing.pdf.

QSR International Pty Ltd. (2014). NVivo qualitative data analysis software, Version 11.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Read, J., & Wette, R. (2009). Achieving English proficiency for professional registration: The experience of overseas-qualified health professionals in the New Zealand context. *IELTS Research Reports*, *10*, 181-222.

Rea-Dickins, P. (1987). Testing doctors' written communicative competence: An experimental technique in English for specialist purposes. *Quantitative Linguistics*, *34*, 185-218.

Reckase, M. (2000). A survey and evaluation of recently developed procedures for setting standards on educational tests. *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements*, 41-70.

Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching*, *47*(2), 135-173.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, *14*(2), 159-176.

Royal Australasian College of Physicians (RACP). (2017). International Medical Graduates - Australia. Retrieved from https://www.racp.edu.au/become-a-physician/overseas-trained-physicians-and-international-medical-graduates.

Royal Australian College of General Practitioners (RACGP). (2016). Referring to other medical specialists: A guide for ensuring good referral outcomes for your patients. Retrieved from http://www.racgp.org.au/your-practice/business/tools/support/referring/.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: how raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 129-152). Cambridge, UK: Cambridge University Press.

Savignon, S. (1983). *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley.

Saxena, S., Dennis, S., Vagholkar, S., & Zwar, N. (2006). Assessment of the learning needs of International Medical Graduates. *Focus on Health Professional Education*, *8*(2), 49-57.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*, 1-30.

Selinker, L., & Douglas, D. (1985). Wrestling with 'context' in interlanguage theory. *Applied Linguistics*, *6*(2), 190-204.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Shin, S. Y., & Lidster, R. (2016). Evaluating different standard-setting methods in an ESL placement testing context. *Language Testing*, *34*(3), 357-381.

Short, S., Hawthorne, L., Sampford, C., Marcus, K., & Ransome, W. (2012). 'Filipino nurses down under': Filipino nurses in Australia. *Asia Pacific Journal of Health Management*, *7*(1), 7.

Skaggs, G., Hein, S. F., & Awuor, R. (2007). Setting passing scores on passage-based tests: A comparison of traditional and single-passage bookmark methods. *Applied Measurement in Education*, *20*(4), 405-426.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panellists thinking when they participate in standard-setting studies?. *Applied Measurement in Education*, *18*(3), 233-256.

Someren, M. V., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: a practical approach to modelling cognitive processes*. London, UK: Academic Press.

Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing*, *18*(4), 463-481.

Spike, N. A. (2006). International medical graduates: The Australian perspective. *Academic Medicine*, *81*(9), 842-846.

Srivastava, R. (2008). A bridge to nowhere - The troubled trek of foreign medical graduates. *The New England Journal of Medicine*, *358*(3), 216-219.

Stansfield, C. W., & Wu, W. (2001). Towards authenticity of task in test development. *Language Testing, 18(2),* 187-206.

Strauss, A. L., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded*

theory procedures and techniques. Newbury Park, CA: Sage.

Tannenbaum R. J., & Katz I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology, Vol 3: Testing and assessment in school psychology and education* (pp. 455-477). Washington, DC: American Psychological Association.

Tannenbaum, R. J., & Wylie, E. C. (2004). *Mapping test scores onto the common European framework (ETS RR-05-18).* Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English-language proficiency test scores onto the Common European Framework (TOEFL Research Report No. RR-80).* Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard setting methodology (TOEFL iBT Series Report No. 06).* Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Wylie, E. C. (2013). *Mapping TOEIC® and TOEIC Bridge™ Test Scores to the Common European Framework of Reference*. Princeton, NJ: Educational Testing Service.

Thorndike, E. L. (1918). Individual differences. *Psychological Bulletin*, *15*, 148-159.

Tjia, J., Mazor, K. M., Field, T., Meterko, V., Spenard, A., & Gurwitz, J. H. (2009). Nurse-physician communication in the long-term care setting: Perceived barriers and impact on patient safety. *Journal of Patient Safety*, *5*(3), 145.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Valette, R. M. (1977). *Modern language testing*. New York, NY: Harcourt, Brace and World.

Vann, R. J., Lorenz, F. O., & Meyer, D. M. (1991). Error gravity: Faculty response to errors in written discourse of non-native speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181-19). Norwood, NJ: Ablex Publishing.

Vaus, D. D. (2002). *Surveys in social research* (5th ed.). New South Wales, Australia: Allen & Unwin.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Hampshire, UK: Palgrave Macmillan.

Wendt, A., & Woo, E. A. (2009). A minimum English proficiency standard for The Test of English as a Foreign Language Internet-Based Test (TOEFL iBT). *RN*, *86*, 19.

Wette, R. (2011). English proficiency tests and communication skills training for overseas-qualified health professionals in Australia and New Zealand. *Language Assessment Quarterly*, *8*(2), 200-210.

Widdowson, H. (1979). *Explorations in applied linguistics*. Oxford, UK: Oxford University Press.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive difference in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465-492.

Woo, A., Dickinson, P., & de Jong, E. J. (2010). Setting an English language proficiency passing standard for entry-level nursing practice using the Pearson Test of English Academic. Retrieved from https://www.ncsbn.org/NCLEX_technicalbrief_PTE_2010.pdf.

Woodward-Kron, R., Stevens, M., & Flynn, E. (2011). The medical educator, the discourse analyst, and the phonetician: A collaborative feedback methodology for clinical communication. *Academic Medicine*, *86*(5), 565-570.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370-371.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Xi, X. (2008). Methods of test validation. *Encyclopedia of language and education*, *7*, 177-96.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?. *Language Testing*, *28*(1), 31-50.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cut scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org