



Programmatic assessment for learning, meaningful personalisation of assessment

23d Thai Medical Education Conference

Flinders University

Acknowledgements

- Dr Gillian Kette
- Dr Nyoli Valentine
- Prof Cees van der Vleuten



Directions for assessment reform

All Australian higher education to submit credible and concrete action plans by June 2024

Assessment reform for the age of artificial intelligence September 2023 **TEQSA**



Why the urgency?

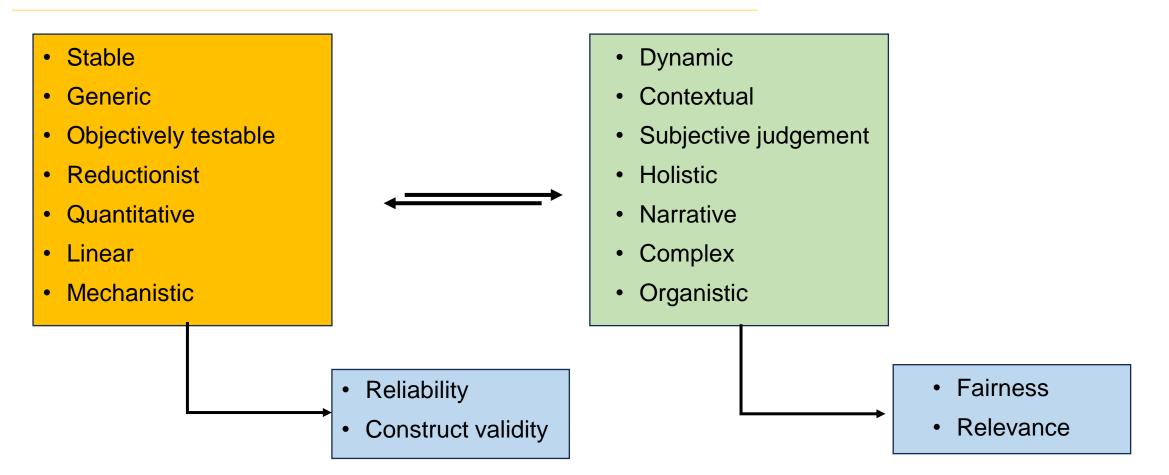




1 Changes in medical education



The nature of competence





My study

PBL groups Lectures Practicums Clerkships/clinical rotations MCQs/OEQs Oral examinations OSCEs Essays

Current students

TBL groups (Recorded/streamed) lectures Practicums Clerkships/clinical rotations MCQs/OEQs Oral examinations OSCEs Essays Portfolio



My study

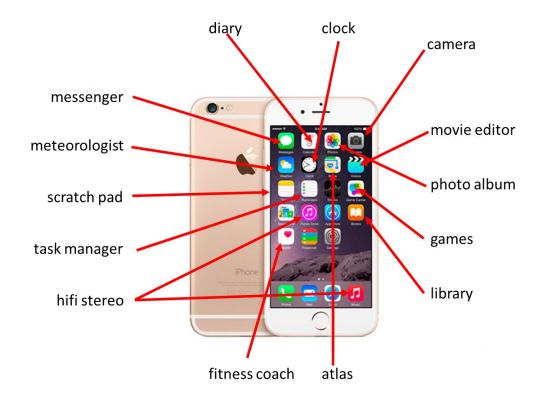






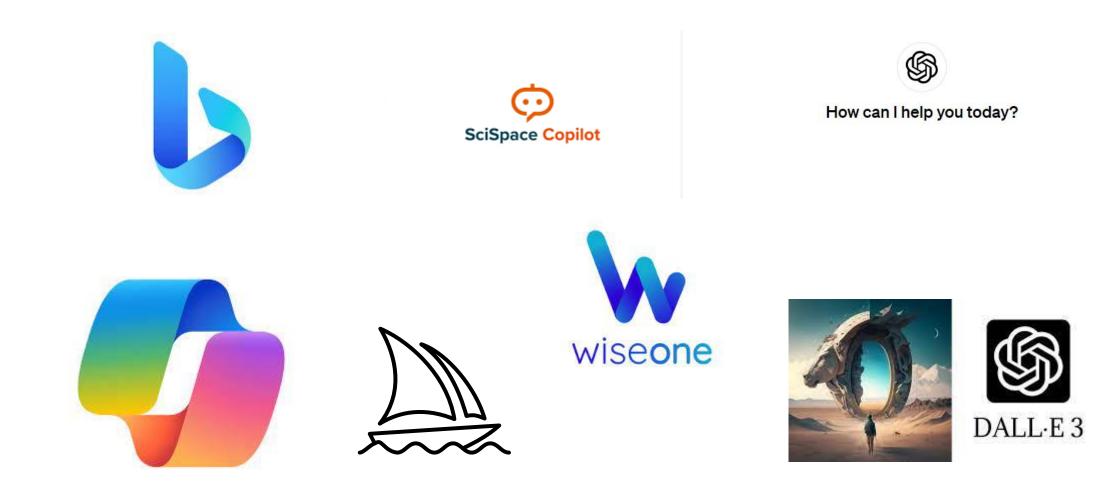


Current students





And now....generative AI





2 Changes in technology



Changes in the world

- Democratisation of knowledge
- Democratisation of transactions/trust
- Democratisation of teaching
- Democratisation of communication
- Democratisation of availability
- Democratisation of intelligence/cognition











modern student 'affordances'

Communication/communities Collaborations Creations Convergence

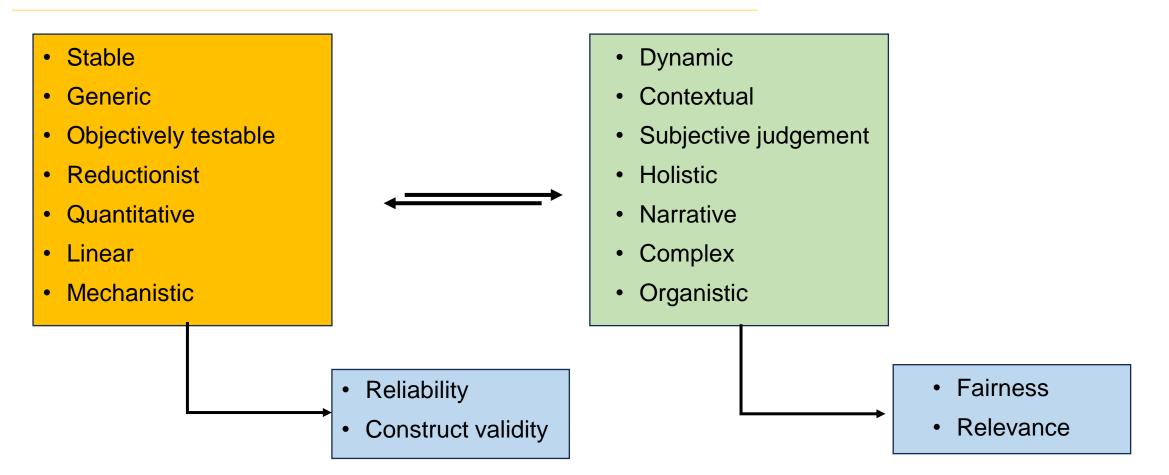




3 Problems with current assessment



The nature of competence





Let's explore the current situation

- Traits are stable and generic characteristics
- Individual items are meaningless in themselves
- Only sum scores determine what the assessment measures
- Statistics are based on elimination of information
- One best assessment method for each trait

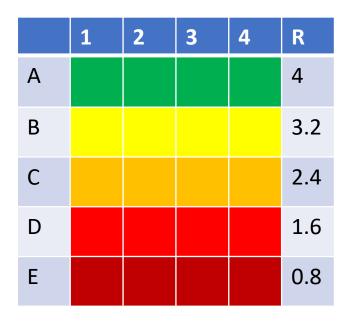


Traits are stable and generic characteristics

knowledge	skills	problem solving	attitude	
TEST	TEST	TEST	TEST	
TEST	TEST	TEST	TEST	

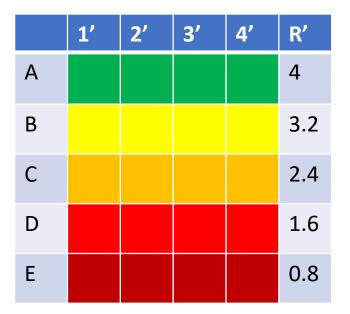


testing: theoretical situation



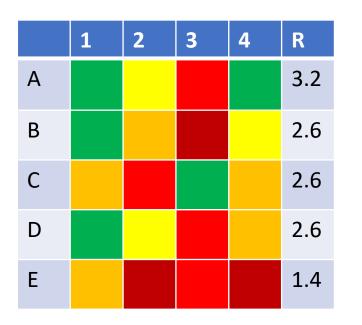
TEST

PARALLEL TEST

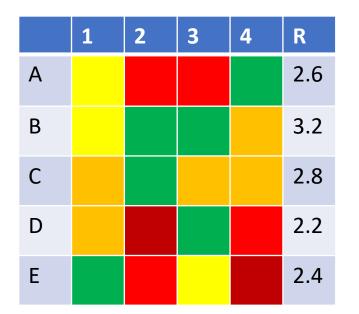




testing: real situation



TEST



PARALLEL TEST



Let's explore the current situation

- Traits are stable and generic characteristics
- Individual items are meaningless in themselves
- Only sum scores determine what the assessment measures
- Statistics are based on elimination of information
- One best assessment method for each trait



Individual items are meaningless in themselves

Mr Jones is 35 years old. He visits his family physician because of chest pain. The pain comes and goes and can be improved by deep breathing or a change of posture. What is at this moment the most likely origin of his pain??

- a the chest wall;
- b the lungs;
- c the heart;
- d the oesophagus.



Ms. Smit is 72 years old. She has chest pains. Several times her blood pressure is taken and found to be 170/100 mmHg. Which antihypertensive drug is most indicated for her??

- a captopril.
- b chlorthalidone.
- c metoprolol.



Individual items are meaningless in themselves

resuscitation 'station' in an OSCE





Individual items are meaningless in themselves

communication 'station' in an OSCE



Flinders University

Let's explore the current situation

- Traits are stable and generic characteristics
- Individual items are meaningless in themselves
- Only sum scores determine what the assessment measures
- Statistics are based on elimination of information
- One best assessment method for each trait



Total scores

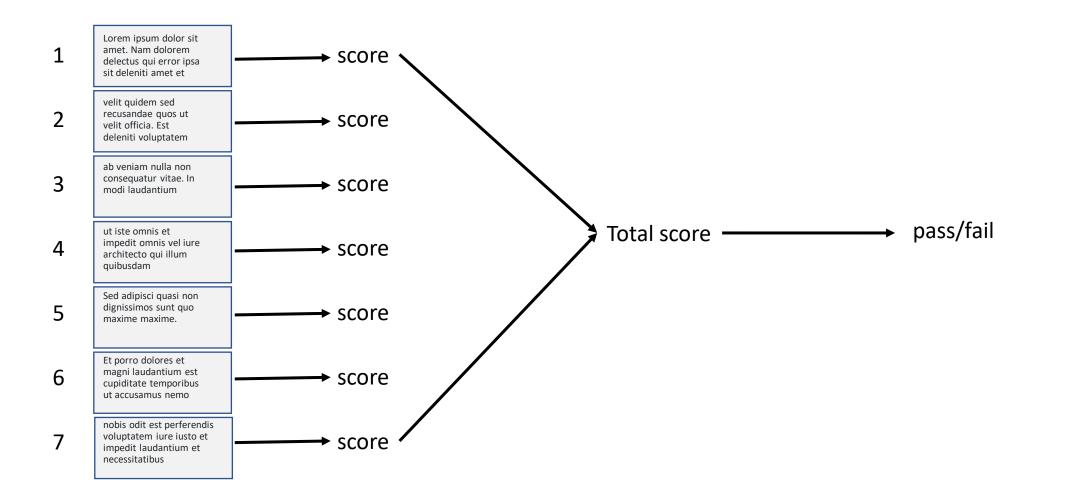


Let's explore the current situation

- Traits are stable and generic characteristics
- Individual items are meaningless in themselves
- Only sum scores determine what the assessment measures
- Statistics are based on elimination of information
- One best assessment method for each trait



elimination of information



Flinders University

Let's explore the current situation

- Traits are stable and generic characteristics
- Individual items are meaningless in themselves
- Only sum scores determine what the assessment measures
- Statistics are based on elimination of information
- One best assessment method for each trait



knowledge	skils	problem solving	attitudes
TEST > TEST	TEST > TEST	TEST > TEST	TEST > TEST



And now.....generative AI

Gilson et al

JMIR MEDICAL EDUCATION

Original Paper

How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment

Aidan Gilson¹², BS, Courad W Saffanel¹, BS; Thomas Huang², BS, Vinnig Socrates¹³, MS, Ling Chi¹, BSE; Richard Andrew Taylor¹², MD, MHS; David Chartsah¹², PhD ¹Section for Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, CT, United States ¹Doparator of Emergency Medicine, Yue University School of Medicine, New Haven, CT, United States

-Department of sameginer, you can be able of the same set o

Corresponding Author: David Chartash, PhD

Sector for Biomedical Informatics and Data Science Yale University School of Medicine 300 George Sheet Suite 501 New Haven, CT, 06511 United States Phone: 1203 737 7379 Phone: 1203 737 7379

Related Articles: Comment in: http://mhealth.jmir.org/2023/1/e46876/ Comment in: http://mhealth.jmir.org/2023/1/e46885/

Comment in: http://mheaith.jmir.org/2023/1/e46885 Comment in: https://mededu.jmir.org/2023/1/e4830 Comment in: https://mededu.jmir.org/2023/1/e50336

Abstract

Background: Chat Generative Pre-trained Transformer (ChatGPT) is a 175-billion-parameter natural language processing model that can generate conversation-style responses to user imput. Objective: This tanky aimed to evaluate the performance of ChatGPT on questions within the scope of the United States Medical

Locenize Examination Step 1 and Step 2 exam, is well as to analyze segments for user interpretability. Methods: We used 2 as of a multiple-tokine question to evolution CutoPT's production setup of the method industry with a loss Step 1 and Step 2. The first set was derived from AMEOS's, a commonly used question built for method industry, which also provide statistics on question difficulty, and the performance on emerivativity to the nucleon 1 and for all varies of Medical Examiner (MMEOS) and the set of the set

Readim: Of the 4 data set, ALROS5.569; ALROS5.569; TAME.Free-Sop], and NBLE.Free-Sop]. CharGT hadreed associated at 4% 14400, 024 5% (2100), 024 5% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910), 024 5\% (5910),

EUROPEAN JOURNAL OF ENGINEERING EDUCATION 2023, VOL. 48, NO. 4, 559-614 https://doi.org/10.1080/03043797.2023.2213169 Taylor & Francis Taylor & Francis Group

OPEN ACCESS

ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity

Sasha Nikolic $^{\circ}$ a, Scott Daniel $^{\circ}$ b, Rezwanul Haque $^{\circ}$ c, Marina Belkina^d, Ghulam M. Hassan $^{\circ}$ e, Sarah Grundy^f, Sarah Lyden $^{\circ}$ 9, Peter Neal $^{\circ}$ f and Caz Sandison $^{\circ}$ a

⁸Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, Australia; ⁸School of Professional Practice & Leadership, University of Technology Sydney, Sydney, Australia; ⁵School of Science, Technology and Engineering, University of the Sunshine Coast, Stopp Downs, Australia; ⁶Colege, Western Sydney University, Sydney, Australia; ⁵School of Computer Science & Software Engineering, University of Western Australia; Perth, Australia; ⁵School of Chemical Engineering, University of New South Wales, Sydney, Australia; ⁹School of Engineering, University of Tasmania, Hobart, Australia

ABSTRACT

ChatGPT, a sophisticated online chatbot, sent shockwaves through many sectors once reports filtered through that it could pass exams. In higher education, it has raised many questions about the authenticity of assessment and challenges in detecting plagiarism. Amongst the resulting frenetic hubbub, hints of potential opportunities in how ChatGPT could support learning and the development of critical thinking have also emerged. In this paper, we examine how ChatGPT may affect assessment in engineering education by exploring ChatGPT responses to existing assessment prompts from ten subjects across seven Australian universities. We explore the strengths and weaknesses of current assessment practice and discuss opportunities on how ChatGPT can be used to facilitate learning. As artificial intelligence is rapidly improving, this analysis sets a benchmark for ChatGPT's performance as of early 2023 in responding to engineering education assessment prompts. ChatGPT did pass some subjects and excelled with some assessment types. Findings suggest that changes in current practice are needed, as typically with little modification to the input prompts, ChatGPT could generate passable responses to many of the assessments, and it is only going to get better as future versions are trained on larger data sets.

ARTICLE HISTORY Received 28 March 2023 Accepted 8 May 2023

Artificial intelligence (AI); assessment; ChatGPT; gengineering education; ct GPT-3; integrity

JAMA | Special Communication | AI IN MEDICINE

Creation and Adoption of Large Language Models in Medicine

Nigam H. Shah, MBBS, PhD; David Entwistle, BS, MHSA; Michael A. Pfeffer, MD

IMPORTANCE There is increased interest in and potential benefits from using large language models (LLMs) in medicine. However, by simply wondering how the LLMs and the applications powered by them will reshape medicine instead of getting actively involved, the agency in shaping how these tools can be used in medicine is lost.

Viewpoint Related article CME at jamacmelookup.com

OBSERVATIONS Applications powered by LLMs are increasingly used to perform medical tasks without the underlying language model being trained on medical records and without verifying their purported benefit in performing those tasks.

CONCLUSIONS AND RELEVANCE The creation and use of LLMs in medicine need to be actively shaped by provisioning relevant training data, specifying the desired benefits, and evaluating the benefits via testing in real-world deployments.

ABOUT

SEARCH

818

Save

advanced search

743

Citation

264

Share

JAMA. doi:10.1001/jama.2023.14217 Published online August 7, 2023. Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Nigam H. Shah, MBBS, PhD, Center for Biomedical Informatics Research, Stanford University, 3180 Porter Dr, Palo Atto, CA 94305 (nigam@) stanford.edu).

PLOS DIGITAL HEALTH

GOPEN ACCESS 💋 PEER-REVIEWED

Performance of ChatGPT on USMLE: Potential for Al-assisted medical education using large language models

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, Victor Tseng @

Published: February 9, 2023 • https://doi.org/10.1371/journal.pdig.0000198

Article	Authors	Metrics	Comments	Media Coverage	Download PDF
					Print Sh
Abstract	Abstract				
uthor summary	We evaluated	Check for upd			
ntroduction	Medical Licens	sing Exam (USMLE),	which consists of three exams	s: Step 1, Step 2CK, and Step	Related PLOS Artic
lethods			e passing threshold for all thre t. Additionally, ChatGPT demo		Related PLUS Artic
Results	concordance a	has COMPANION			
Discussion	may have the making.	potential to assist with	n medical education, and pote	ntially, clinical decision-	ChatGPT passing US shines a spotlight on
Supporting information					flaws of medical edu



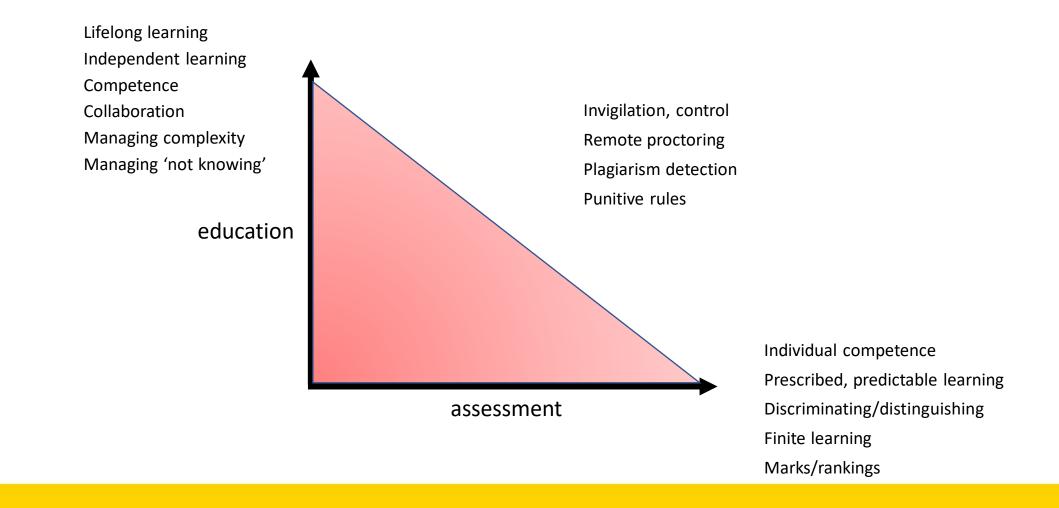
4 Value proposition and process



Value proposition and processes

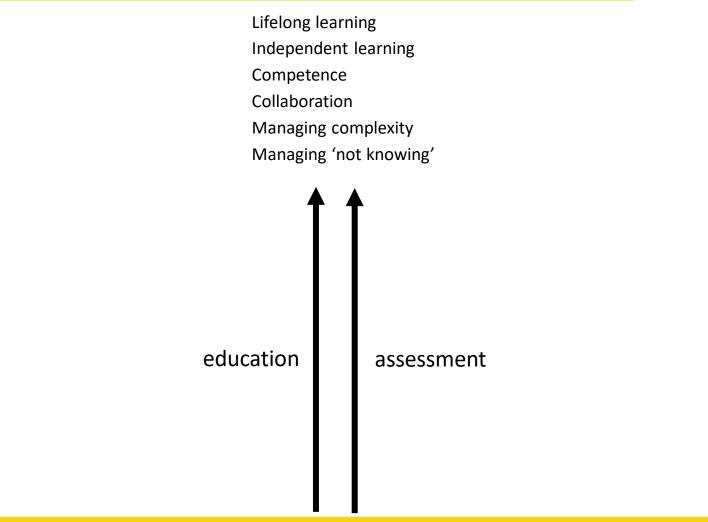


The crux of the problem?





The solution?

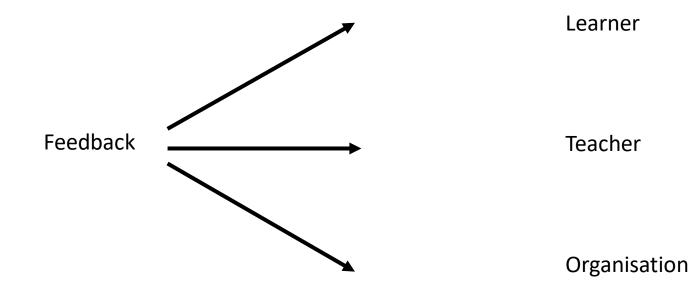




5 Assessment for learning



Assessment for learning ≠ just formative assessment





Assessment for learning

- Content
 - meaningful

• active

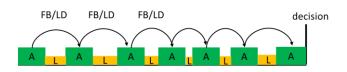
• collaborative



Assessment for learning

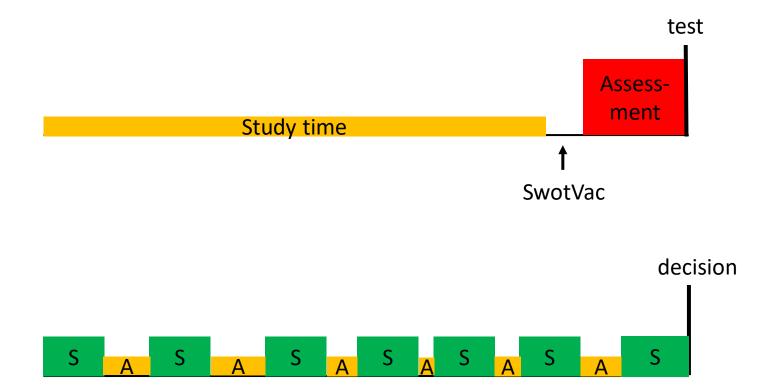
- Process:
 - Distributed
 - Interleaved
 - Connected





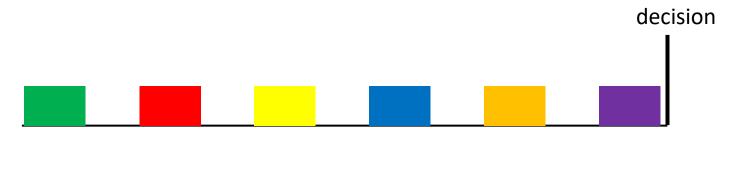


Assessment for learning: distributed





Assessment for learning: interleaved

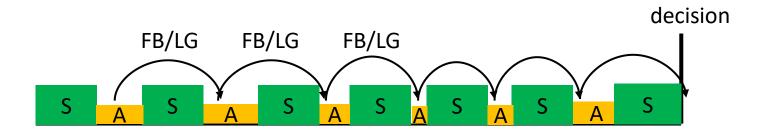








Assessment for learning: connected





Assessment of learning

- Score 1 🔶 Mini-CEX 1
- Score 2 Mini-CEX 2
- Score 3 Mini-CEX 3
- Score 4 🔶 Mini-CEX 4
- Score 5 Mini-CEX 5
- Score 6 Mini-CEX 6

- → Feedback + LG 1
- → Feedback + LG 2
- → Feedback + LG 3
- → Feedback + LG 4
- \rightarrow Feedback + LG 5
- \rightarrow Feedback + LG 6



Assessment for learning

- Score 1 Mini-CEX 1
- Score 2 🔶 Mini-CEX 2
- Score 3 Mini-CEX 3
- Score 4 Mini-CEX 4
- Score 5 🔶 Mini-CEX 5
- Score 6 Mini-CEX 6

- Feedback + LG 1
- → 🛛 Feedback + LG 2 🟅
- Feedback + LG 3
- 🛶 🛛 Feedback + LG 4 🔮
- 🛶 🛛 Feedback + LG 5 🔮
- → Feedback + LG 6 ◆



But why programmatic assessment?

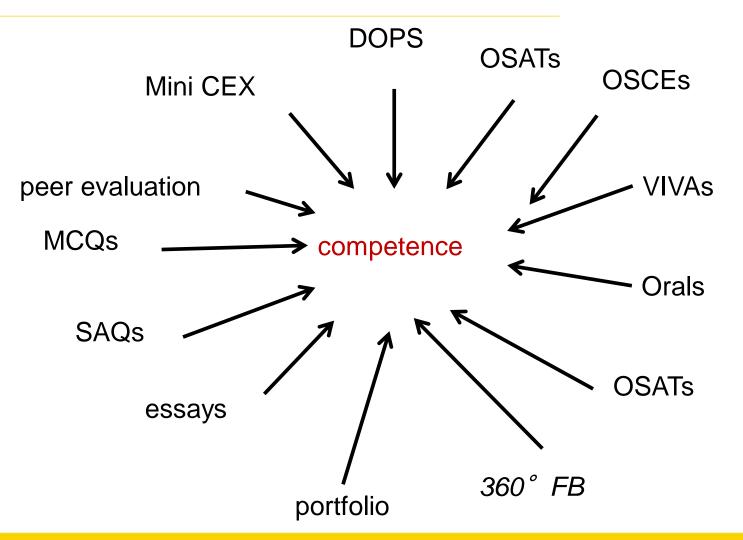




• Multiple instruments with meaningful synthesis

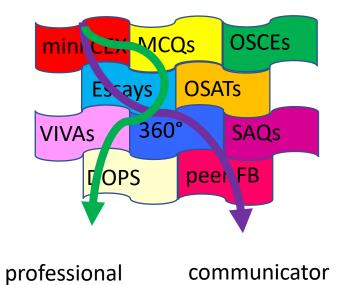


Multiple instruments





Mulitple instruments



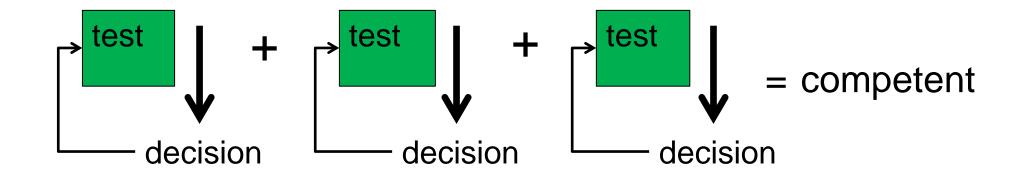
professional



- Multiple instruments with meaningful synthesis
 - Interleaving
 - Meaningfulness
 - complexity

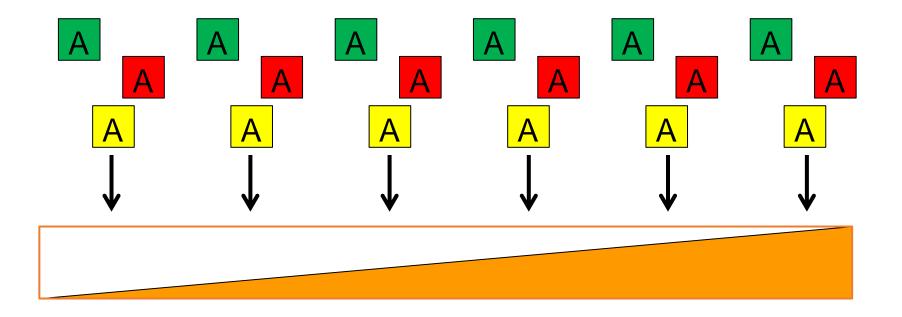


Every assessment moment = decision moment





Longitudinal assessment

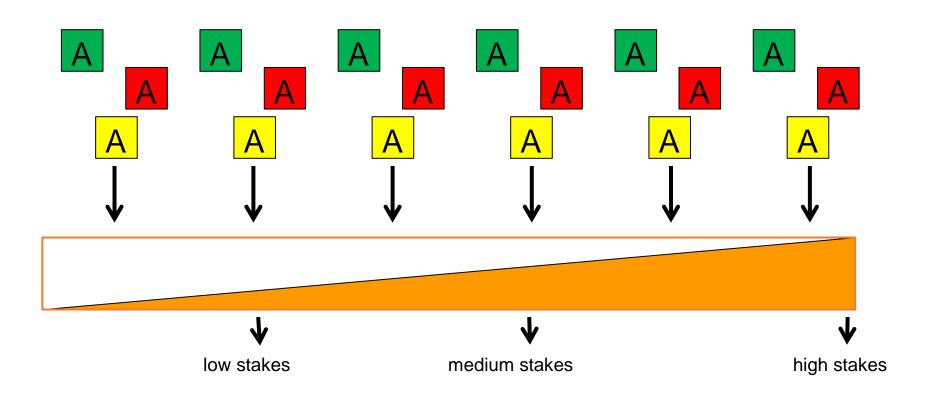




- Continuous and longitudinal assessment
 - Distributed learning
 - Connected assessment



Proportionality



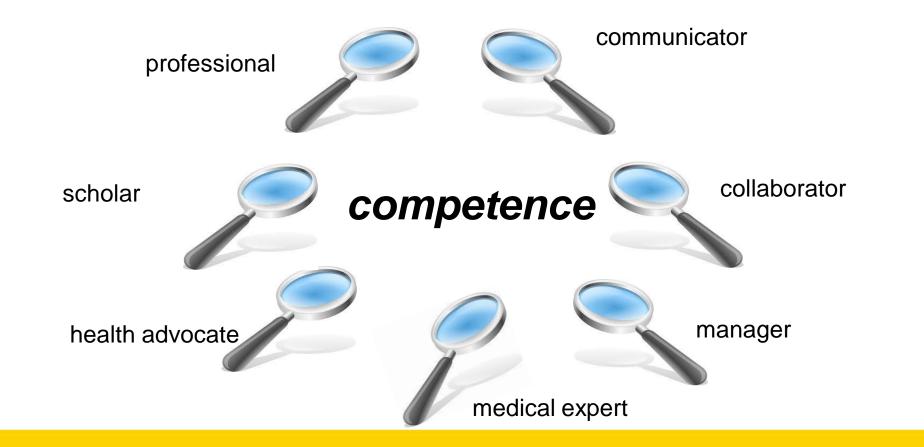


- Proportionality of decisions
 - Goodhart's law
 - Fairness
 - Meaningfulness/Relevance
 - Prevention of leniency bias



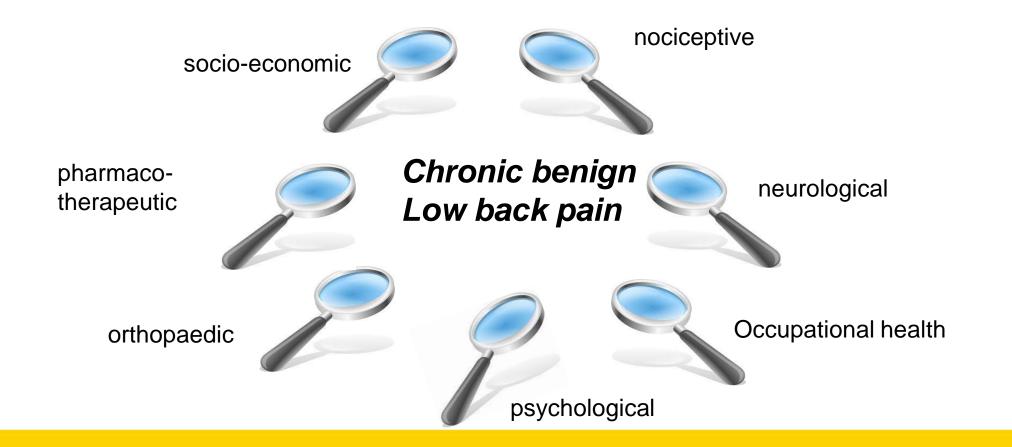
- Multiple instruments with meaningful synthesis
- Continuous and longitudinal assessment
- Proportionality of decisions
- Variety on quality approaches
- Assessment as a diagnostic process (instead of measurement)



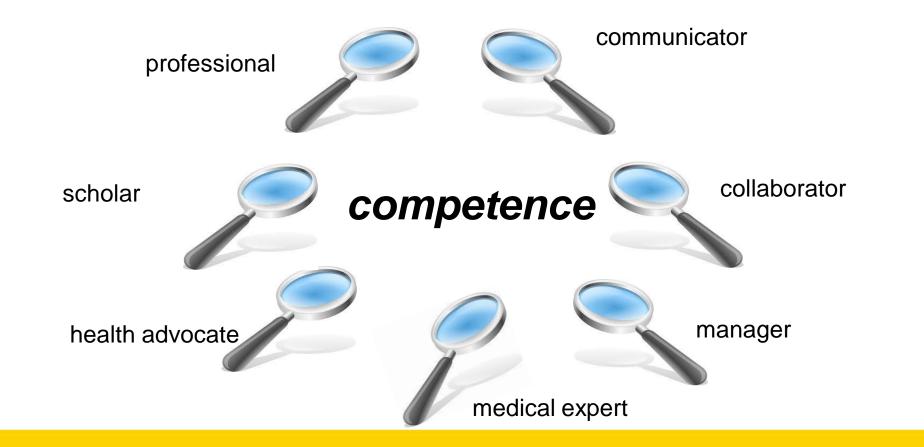




Domains as lenses













background reading

- 1. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological bulletin 1955;52(4):281 302.
- 2. Ebel RL. The practical validation of tests of ability. Educational Measurement: Issues and Practice 1983;2(2):7 10.
- 3. Cronbach LJ. What price simplicity? Educational Measurement: Issues and Practice 1983;2(2):11-12.
- 4. Kane MT. Validation. In: Brennan RL, ed. Educational Measurement. Westport: ACE/Praeger 2006:17 64.
- 5. Norman GR, Smith EKM, Powles AC, et al. Factors underlying performance on written tests of knowledge. Medical Education 1987;21:297-304.
- 6. Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement 1982;6(1):1-11.
- 7. Van der Vleuten CPM, Van Luyk SJ, Beckers HJM. A written test as an alternative to performance testing. Medical Education 1988;22:97-107.
- 8. Norman G, Swanson D, Case S. Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. Teaching and Learning in Medicine 1996;8(4):208-16.
- 9. Norman GR, Van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Medical Education 1991;25(2):119-26.
- 10. Boud D. Assessment and the promotion of academic values. Studies in Higher Education 1990;15(1):101-11.
- 11. Govaerts MJB, Van der Vleuten CPM, Schuwirth LWT, et al. Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. Advances in health sciences education 2007;12(2):239-60.
- 12. Canmeds. http://www.royalcollege.ca/portal/page/portal/rc/canmeds Ottawa2005 [accessed 26 August 2018.
- 13. ACGME. http://www.acgme.org/What-We-Do/Accreditation/Milestones/Overview Chicago2007 [accessed August 2018].
- 14. Albanese MA, Mejicano G, Mullan P, et al. Defining characteristics of educational competencies. Medical Education 2008;42(3):248-55.



background reading

- 15. Van der Vleuten CPM. The assessment of Professional Competence: Developments, Research and Practical Implications. Advances in Health Science Education 1996;1(1):41-67.
- 16. Kane M. Current concerns in validity theory. Journal of Educational Measurement 2001;38(4):319-42.
- 17. Schuwirth LWT, Van der Vleuten CPM. A plea for new psychometrical models in educational assessment. Medical Education 2006;40(4):296-300.
- 18. Driessen E, Van der Vleuten CPM, Schuwirth LWT, et al. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. Medical Education 2005;39(2):214-20.
- 19. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Medical Education 2005;39(3):309-17.
- 20. Dannefer E, Henson L. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. Academic Medicine 2007;82(5):493-502.
- 21. Van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. Medical Teacher 2012;34:205-14. doi: 10.3109/0142159X.2012.652239
- 22. Schuwirth LWT, Van der Vleuten CPM, Durning SJ. What programmatic assessment for learning in medical education can learn from healthcare. Perspectives on Medical Education 2017:1-5. doi: 10.1007/s40037-017-0345-1
- 23. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. Medical Education 2012;46(1):38-48. doi: 10.1111/j.1365-2923.2011.04098.x
- 24. Schuwirth LWT, Van der Vleuten CPM. Assessing Competence: Extending the Approaches to Reliability. In: Hodges BD, Lingard L, eds. The Question of Competence. Ithaka New York US: Cornell University Press 2012.

Prideaux Discipline of

Clinical Education

25. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Medical Teacher 2011;33(6):478-85.



background reading

- 26. Delandshere G, Petrosky AR. Assessment of Complex Performances: Limitations of Key Measurement Assumptions. Educational Researcher 1998;27(2):14-24.
- 27. Shute V. Focus on formative feedback. Review of educational research 2008;78(n):153-89.
- 28. Ericsson KA. An expert-performance perspective of research on medical expertise: the study of clinical performance. *Medical Education* 2007;41:1124-30. doi: 10.1111/j.1365-2923.2007.02946.x
- 29. Cook DA, Kuper A, Hatala R, et al. When assessment data are words: validity evidence for qualitative educational assessments. Academic Medicine 2016;91(10):1359-69.
- 30. Popham WJ. Assessment Literacy for Teachers: Faddish or Fundamental? Theory Into Practice 2009;48:4-11.
- 31. Watling C, Driessen E, Van der Vleuten CPM, et al. Beyond individualism: professional culture and its influence on feedback. Medical Education 2013;47(6):585-94.
- 32. Harrison CJ, Könings KD, Schuwirth L, et al. Barriers to the uptake and use of feedback in the context of summative assessment. Advances in health sciences education 2015;20(1):229–45. doi: 10.1007/s10459-014-9524-6
- 33. Cilliers FJ, Schuwirth LWT, Adendorff HJ, et al. The mechanisms of impact of summative assessment on medical students' learning. Advances in health sciences education 2010;15:695-715. doi: 10.1007/s10459-010-9232-9
- 34. Valentine N, Durning S, Shanahan EM, et al. Fairness in human judgement in assessment: A hermeneutic literature review and conceptual framework. Advances in Health Sciences Education 2021;26(2):713-38.
- 35. Valentine N, Shanahan EM, Durning SJ, et al. Making it fair: Learners' and assessors' perspectives of the attributes of fair judgement. Medical Education 2021;55(9):1056-66. doi: https://doi.org/10.1111/medu.14574
- 36. Valentine N, Durning SJ, Shanahan EM, et al. The pursuit of fairness in assessment: Looking beyond the objective. Medical Teacher 2022:1-7.

