Author/s:
Chiavaroli, N;Le, L;Parker-Newlyn, L;Pywell, S

Title:
Should we extend the currency of cognitive ability test scores? Considerations from construct, equity, and psychometric perspectives in medical selection

Date:
2024-03

Citation:
Chiavaroli, N., Le, L., Parker-Newlyn, L. & Pywell, S. (2024). Should we extend the currency of cognitive ability test scores? Considerations from construct, equity, and psychometric perspectives in medical selection. International Journal of Selection and Assessment, 32 (1), pp.138-148. https://doi.org/10.1111/ijsa.12453.

Persistent Link:
https://hdl.handle.net/11343/355705

Neville Chiavaroli ORCID iD: 0000-0003-1488-9747

Lyndal Parker-Newlyn ORCID iD: 0000-0001-6644-3515

Full Title: Should we extend the currency of cognitive ability test scores? Considerations from construct, equity and psychometric perspectives in medical selection

Running Title: Currency of cognitive ability test scores

Authors:

Neville Chiavaroli, Australian Council for Educational Research

Luc Le, Australian Council for Educational Research

Lyndal Parker-Newlyn, Graduate School of Medicine, University of Wollongong, Australia

Sean Pywell, Australian Council for Educational Research

Correspondence:

Neville Chiavaroli, Australian Council for Educational Research, Camberwell VIC 3124, Australia Email: neville.chiavaroli@acer.org

Acknowledgements:

The authors would like to acknowledge the valuable feedback provided by Professor Paul Garrud (University of Nottingham, UK) and important support provided by Lisa Norris (ACER) in the development of this paper.

Funding

Authors NC, LL and SP received financial support from the Australian Council for Educational Research for research and writing related to this paper.

Title: Should we extend the currency of cognitive ability test scores? Considerations from construct, equity and psychometric perspectives in medical selection

Abstract

Scores on cognitive ability tests used in tertiary admissions contexts generally have very limited currency. This can have significant implications for prospective applicants to high-demand courses which use cognitive tests as part of the selection process. In this paper we present both psychometric and non-psychometric considerations regarding the score currency of ability tests, using GAMSAT as an example. We found that GAMSAT scores showed sufficient stability at the cohort level for institutions to be reasonably confident that a test score would continue to provide a valid representation of cognitive ability for up to a five-year period. However, candidates' pre-test preparation will influence whether it is in their interest to re-sit a test even within an extended currency period.

Practitioner points:

- Cognitive ability tests used in tertiary admissions contexts generally have very limited score currency, typically one or two years.
- This limited period of currency means that many candidates may need to re-sit the test if not initially successful, which can have significant and wide-ranging implications for both candidates and institutions, especially in terms of diversity and equity.
- Our research on GAMSAT re-sit scores suggest that cognitive ability tests with robust equating procedures can have sufficient stability at the cohort level for institutions to be reasonably confident that a test score will continue to provide a valid representation of cognitive ability for up to a five-year period.
- Nonetheless, the variability observed in the scores of some individual candidates means that it may sometimes be in the interests to re-sit the GAMSAT despite the extended

currency; this will be influenced by the extent of personal preparation with which candidates attempt their first sitting.

Keywords: cognitive ability testing, equity, medical school selection, practice effects, score currency, widening participation, test resits

Introduction

Scores on cognitive ability tests used for tertiary selection generally have limited currency, typically only one or two years.[1] Candidates who are unsuccessful in obtaining a place in their preferred course therefore usually need to re-sit the required admissions test in order to be re-considered for a place. This limited period of currency has significant implications for prospective applicants to high-demand courses which utilize selection tests. Yet there appears to be little or no discussion in the literature justifying the relatively short currency of such tests.

From a pragmatic perspective, short currency periods may be understandable. For successful candidates, their test score is unlikely to ever be needed again, while for unsuccessful candidates, a non-qualifying score is also redundant, and commonly acts as a stimulus to re-sit the test, potentially with better preparation. However, as we discuss below, not all candidates have the resources to re-sit a selection test. Further, in some circumstances, even nominally successful test scores may expire before the student gains admission to the course, especially where there is intense competition for

[1] Examples include: BMAT – 1 year (www.admissionstesting.org); IELTS – 2 years (www.britishcouncil.org); MCAT – 2-3 years (depending on medical school; students-residents.aamc.org); TOEFL – 2 years (www.ets.org/toefl); UCAT – 1 year (www.ucat.edu.au). Notable exceptions to this pattern include GMAT – 5 years (www.mba.com); LSAT, up to 6 years (but with limits on test attempts; https://www.lsac.org/lsat/lsat-scoring) and SAT, which technically has no expiry date, although the developers advise colleges that scores more than five years old 'may be less valid predictors of college academic performance' (satsuite.collegeboard.org/sat). GAMSAT scores, discussed in this paper, currently have 2 years' currency.

places and the selection process involves multiple sequential steps. Finally, the recent COVID pandemic has disrupted educational practices at all levels and across the globe, raising new challenges for testing agencies, medical schools and student applicants (Camara & Mattern, 2022; Castro et al., 2022). Apart from requiring rapid shifts from pencil-and-paper to online testing (Corridon, 2021), the pandemic also disrupted admissions timelines, both for selection decisions and for applicants to take up their offered places. In some cases, this also meant that potentially eligible test scores may have expired before completion of the selection process.

It is primarily this latter context that prompted a re-consideration of the currency period of scores on the GAMSAT (Graduate Medical School Admissions Test) by the governing consortium, although the other abovementioned considerations had also been discussed for some time. In this paper, we present and discuss the factors that were considered in re-evaluating the score currency period, including examples of data that were used to explore the issue in relation to GAMSAT scores. We present this specific example as a way of exploring the question of test score currency more generally, and to discuss the data and factors that we consider relevant when evaluating the appropriateness of the currency of cognitive ability tests, especially in the tertiary admissions context.

GAMSAT is developed by the Australian Council for Educational Research (ACER) in conjunction with GEMPASS Australia Limited to assist in the selection of applicants for graduate-entry medical and health professional programs in Australia, Ireland, and the UK (ACER, 2021). The test is divided into three sections (Reasoning in Humanities and Social Sciences; Written Communication; and Reasoning in Biological and Physical Sciences) and is designed to assess 'the capacity [of applicants] to undertake

high-level intellectual studies in a demanding course' (ACER, 2021). It includes knowledge and application of concepts in the relevant sciences, as well as broader skills such as textual and data interpretation, logico-deductive reasoning, critical thinking, and written communication. GAMSAT has had a two-year score currency since its inception in 1995 (Aldous et al., 1997). However, as mentioned above, the impact of COVID and other considerations have led the governing consortium of medical schools to explore whether the currency of GAMSAT scores might justifiably be extended, to accommodate similar circumstances and to address other important considerations for medical selection.

Relevant factors for extending the currency of GAMSAT scores

Apart from the practical issues of currency due to COVID-related impacts on administration of the test, three other key factors have formed part of the deliberations of the GAMSAT consortium in relation to the currency of test scores: the nature of the construct, the potential impact on diversity and equity of successful applicants, and psychometric data relating to score stability.

The nature of cognitive ability tests for tertiary selection

Cognitive ability tests generally represent a mix of verbal reasoning, quantitative reasoning, and discipline-specific skills (Kuncel & Hezlett, 2010). While there is considerable overlap between academic achievement and cognitive ability, achievement tests generally aim to assess prior learning and emphasize knowledge recall (Kelly et al., 2018; Stemler, 2012; Stemler & Sternberg, 2013), whereas cognitive ability tests target reasoning ability, although typically in relation to specific

disciplinary domains (Kuncel & Hezlett, 2010).[2] Such reasoning ability includes understanding of texts, critical thinking, problem-solving, and comprehension and processing of complex information (ACER, 2002; Kuncel & Hezlett, 2010). Meta-analyses across disciplines provide robust evidence of the validity of cognitive ability tests for predicting relevant performance and academic outcomes in graduate courses (Kreiter & Kreiter, 2007; Kuncel & Hezlett, 2007; Kuncel & Hezlett, 2010; Ross et al., 2013; Stemler, 2012). Performance on such tests has also been shown to be relatively stable over time (Huber et al., 2015; Kuncel & Hezlett, 2007), notwithstanding the potential for (on average) small improvement with retesting (as discussed below). Scores from these tests are frequently used for university admission purposes, especially for high-demand professional courses such as law, medicine, and management. In many cases, the use of these tests also reflects a desire by the institutions to broaden selection criteria beyond solely prior academic achievement (Kelly et al., 2018).

A key feature of such cognitive ability tests is the inclusion of contextual material or 'prompts' to enable assessment of *application* of knowledge, rather than factual knowledge per se. GAMSAT uses this approach for its disciplinary-oriented sections 1

[2] In many selection contexts, cognitive ability tests are referred to as aptitude tests. Here too, there is variation and imprecision in usage (Lohman, 2004), but, conceptually, 'aptitude' is usually taken to refer to future potential for learning or acquiring of a specific skill, whereas 'ability' refers to present capacity. However, as many scholars have argued, the distinction between aptitude and ability is not always clear or defensible, as poignantly suggested by changes over time to the designation of the 'A' of the SAT test. In this paper, we use the term 'cognitive ability test' to refer to reasoning-based selection tests, even when in certain contexts (such as medical/health professional selection) they are frequently referred to as 'aptitude tests' (e.g. Kelly et al., 2018; Nicholson, 2005).

and 3; examples from each section are provided in the boxes in the Appendix.[3] The nature of such contextualized questions in selection tests is fundamental to any discussion of construct and score currency. A key consequence of this approach is that the underlying construct is only *partially* dependent on disciplinary knowledge, since much if not all of the necessary base information is supplied within the prompt material. In addition to the predictive validity research cited above, the construct validity of this contextual approach to assessing reasoning is perhaps best demonstrated (in the medical school graduate selection context, at least) by the fact that, globally, many non-Science graduates perform sufficiently well on the GAMSAT (and other medical entry tests) to gain entry into medical courses in significant numbers every year (Elliott & Epstein, 2005; Lam et al., 2020; MDANZ, 2021). Likewise, Science graduates need to be able to reason in Humanities contexts in the GAMSAT sufficiently well to be competitive for entry into graduate medical courses.

Diversity and equity considerations

Educational research points to the influence of many external factors on student academic performance, such as quality of teachers and/or schools, parental background, and socioeconomic status (SES) (Perry et al., 2016; Thomson, 2018). One of the outcomes of such research is a growing interest in and commitment to the notion of 'widening participation' in tertiary education, with a focus on how applicants from underrepresented and/or minoritized groups can be better supported in the selection process (Cleland et al., 2018; Fyfe et al., 2022; Rees & Woolf, 2020; Younger et al., 2019). The argument that a broader basis for selection, including assessment of

[3] Other tests using a similar prompt-based approach are BMAT, MCAT, and SAT.

reasoning ability beyond school-based academic performance, should enable a more diverse range of applicants to compete for tertiary places is one of the main justifications for the use of cognitive ability tests in the medical school selection process (Kelly et al., 2016; McDonald et al., 2000; Nicholson, 2005).

While assessing general reasoning ability is one way in which universities might attempt to minimize the advantages conferred by the abovementioned socioeconomic factors, it does not, of course, eliminate them. Other strategies are required to enable successful applications from traditionally underrepresented students, such as targeted outreach to certain schools and student groups (Martin et al., 2018; McLachlan, 200), enabling 'aspirational capacity' in less advantaged students (Ho et al., 2022), and appreciating and responding to the wider social and political context of medical school selection (Cleland et al., 2018; Fielding et al., 2018). Nevertheless, it is vital to ensure that the use of a cognitive ability test does not itself provide an additional barrier to such applicants. Some have argued that the availability of (expensive) commercial coaching for certain applicants is a significant advantage and, therefore, a validity threat to the use of such tests for selection (e.g. Stringer, 2008). On the other hand, while basic *familiarity* with the test format and type of item is strongly encouraged and typically supported by test development agencies (for example, through descriptions of the tested constructs and availability of practice tests), the idea that content-specific *coaching* can significantly increase the likelihood of success on selection tests appears overstated. Despite anecdotal (and marketing) claims of improved scores for candidates who choose to enrol in coaching courses, and similar perceptions among many candidates (Kumar et al., 2018), researchers in the area are generally sceptical of their actual impact on scores, concluding that coaching provides little benefit beyond that achievable through appropriate familiarisation with the form of questions and content

of the test (Griffin, 2018; Griffin et al., 2008; McGaghie et al., 2004; Wilkinson and Wilkinson, 2013).

Another potential threat to equity is the cost of sitting the selection test itself. The current costing models of medical selection processes means that a significant proportion of the costs of the testing process is passed on to candidates. If sitting a selection test once is difficult financially for some candidates, then re-sitting the test in the event of non-success will be even more so, and recent research has established this as a realistic deterrent (Griffin et al., 2019; Kumar et al., 2018). It is in this context that the issue of score currency is particularly relevant. On equity grounds, at least, extending the currency of selection test scores may offer a small but potentially effective way of reducing the obstacles for initially unsuccessful candidates who may not have the means for re-sitting the test.

Psychometric considerations

Extending the currency of a selection test requires that the test construct be relatively stable over the intended time frame. In the case of GAMSAT, the relative stability of GAMSAT scores has previously been established in relation to the results of successive cohorts over the period 2005−2014 (Mercer et al., 2015). This is perhaps not surprising for a high-stakes test with a large cohort, consistent high reliability, relatively unchanged test construct, and a robust equating design incorporating many link items across successive test sittings (ACER, 2022; Aldous et al., 1999). However, such data does not provide information about the potential for improvement on successive sittings, or whether a GAMSAT score remains a relatively valid measure of a candidate's reasoning ability several years after the initial sitting. If this were shown to

be the case, this could indicate that the present two-year currency is unnecessarily conservative.

GAMSAT scoring and post-test analysis is based on the Rasch model (Rasch, 1960), a version of Item Response Theory (Lord & Novick, 1968) used in educational measurement. This model enables equating of scores from all test versions onto the same scale by accounting for differences in test difficulty, allowing scaled scores across GAMSAT tests to be comparable. However, over time there may be a degree of 'drift' in score equivalence. Drift can arise in two ways: changes in what or who is being tested, or as the result of stochastic measurement error inherent in score equating. Large and demographically stable cohorts, stable constructs, and rigorous test and equating design mitigate against both types of drift, but the possibility must be considered in discussions of score currency (Wells et al., 2002).

A key source of evidence for score stability is candidate re-sit data, that is, the performance of candidates who sit the same or equivalent version of the test on more than one occasion. Thousands of unsuccessful applicants choose to re-sit a medical selection test every year (Griffin et al., 2019), thus allowing an estimation of the potential for improvement on such ability tests. Previous studies on other medical selection tests have documented small improvements in candidate scores on subsequent attempts (Andrich et al., 2017; Griffin et al., 2019; Lievens et al., 2007; Puddey et al., 2014), a finding which is consistent with the broader literature on cognitive ability tests (e.g. Hausknecht et al., 2007; Scharfen et al., 2018; Schleicher et al., 2010). We therefore analysed GAMSAT re-sit data (for the period 2015−2020) to investigate the relative stability of scores for candidates who chose to resit the GAMSAT test (including multiple re-sits). Each subsequent attempt involved a different set of items

equated with previous versions using Rasch analysis. We investigated this using three forms of analyses of cohort performance, namely: changes in GAMSAT score relative to previous sittings; range of GAMSAT scores over multiple attempts; and mean scores in the most recent sitting (May 2020) according to testing occasion (see Tables 1 & 2, and Figures 1 & 2).

Data and candidates

We collated and analysed the scores of 7324 candidates from the May 2020 GAMSAT administration as the prime reference point for analysis. Table 1 shows the total number of times that these candidates had ever sat GAMSAT in the five-year period up to and including May 2020, along with the year(s) in which they had previously taken the test.

Table 1: Distribution of GAMSAT May 2020 candidates by number of test sittings in 2015-2020

No of testing occasions	May 2020	Per cent	Sep 2019	Mar 2019	Sep 2018	Mar 2018	Sep 2017	Mar 2017	Sep 2016	Mar 2016	Sep 2015	Mar 2015
1	2788	38.1										
2	1943	26.5	1092	599	87	72	13	24	5	20	2	29
3	1308	17.9	647	1061	336	305	51	99	12	58	8	39
4	599	8.2	261	496	268	386	87	147	16	86	7	43
5	334	4.6	166	275	157	264	98	163	32	110	14	57
6	185	2.5	91	163	95	161	67	141	34	96	14	63
7	89	1.2	45	87	52	85	39	82	34	61	13	36
8	42	0.6	31	40	32	41	25	39	27	33	8	18

9	26	0.4	17	25	22	25	19	26	17	25	11	21
10	6	0.1	6	6	4	6	5	6	6	6	4	5
11	4	0.1	4	4	4	4	4	4	4	4	4	4
Total	**7324**	**100.0**	**2360**	**2756**	**1057**	**1349**	**408**	**731**	**187**	**499**	**85**	**315**

The scores comprised those of 4314 females (58.9%) and 3010 males (41.1%), and 4495 (61.4%) candidates who spoke primarily English at home and 2929 (38.6%) who primarily spoke a language other than English. The corresponding proportions for first-time candidates and resitting candidates are very similar: 59.9% female and 40.1% male for first-time candidates versus 58.3% and 41.7% for resitting candidates, and 60.8% English-speaking background and 39.2% other language for first-time candidates versus 61.7% and 38.3% for resitting candidates. Given that the number of candidates who took GAMSAT May 2020 as their fifth or further attempt accounts for less than 10% of the GAMSAT May 2020 cohort, in the following analyses we combine these candidates into the category '5 or more' testing occasions.

Results

Our initial analysis looked at the changes in GAMSAT scores compared to any sitting in the previous five years. Figure 1 shows the distribution of score changes in the overall GAMSAT score[4] of candidates who sat in May 2020 relative to any previous sitting in the March administrations of 2015-2019 (regardless of the number of sittings), in the form of box and whisker plots. The number of candidates meeting this criterion for each

[4] The 'GAMSAT overall score' is calculated as a weighted mean of the three component scores (where section 3 is double weighted relative to sections 1 and 2).

sitting in March of 2015-2019 is shown along the X-axis. This data indicates that while individual scores differed significantly between sittings (as illustrated by the 'outliers' and the whiskers in each boxplot), the improvement for the majority of re-sitting cohorts is relatively small. The improvement relative to previous test occasion is also relatively constant (generally between -1 and 7 marks for those candidates within the $25^{th}/75^{th}$ percentile boxes), suggesting that the construct and its measurement is relatively stable.

This picture is further supported by Pearson test-retest correlations for resits between May 2020 and back to March 2015; values (on GAMSAT overall score) were: $r_{2019} = 0.77$ (i.e. May 2020 sitting compared with March 2019), $r_{2018} = 0.73$, $r_{2017} = 0.73$, $r_{2016} = 0.70$ and $r_{2015} = 0.69$ respectively. The differences between each of these correlations can be considered minor, given the minimal or small Cohen's effect sizes ($q = 0.09$, 0.09, 0.15, and 0.17 respectively; reference range: $q < 0.1$, no effect; $0.1 \leq q < 0.3$, small effect; $0.3 \leq q < 0.5$, medium effect; $q \geq 0.5$, large effect; Cohen 1988). Further, the only significant difference between the correlations is between r_{2019} and the remaining correlations, which we would attribute to the shorter timespan for 2020/2019. Thus, these high and relatively uniform correlations indicate overall limited variation in GAMSAT scores with successive sittings, consistent with the relatively small changes shown in Figure 1.

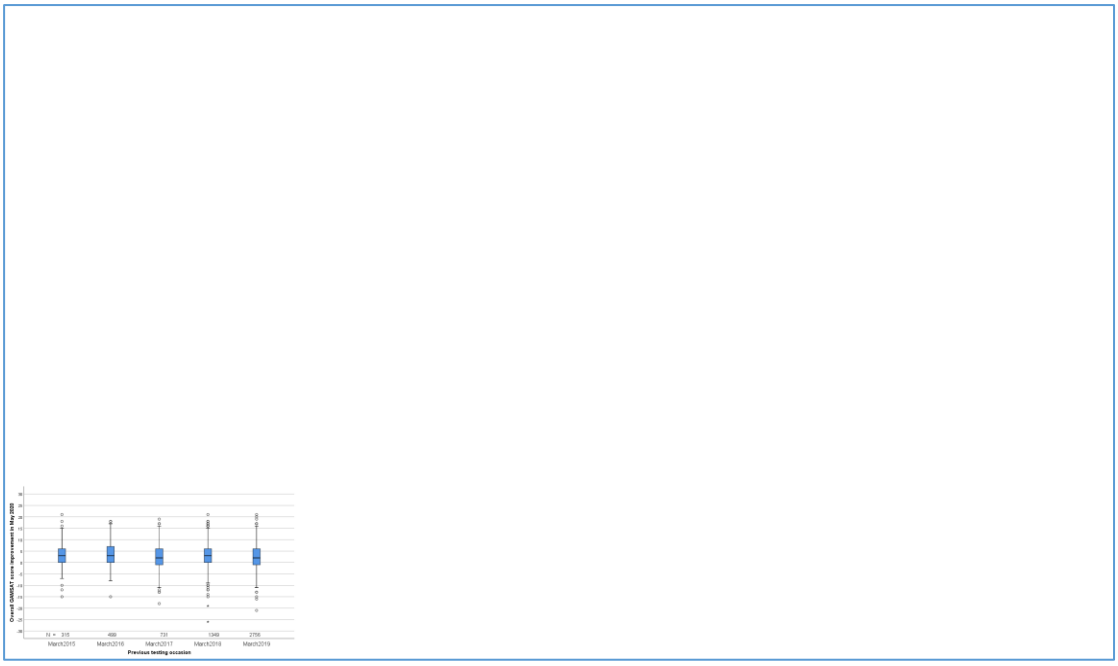

Figure 1: Box and whisker plot showing changes in GAMSAT overall score in May 2020 relative to previous sitting (2015-2019)

Note: In each box and whisker plot, the horizontal middle lines denote median values; the boxes extend from the 25th to the 75th percentile of each group's distribution of scores; vertical extending lines denote adjacent values (i.e., the most extreme values within 1.5 interquartile range of the 25th and 75th percentile of each group); circles denote outliers (observations outside the range of adjacent values); and stars denote extreme outliers more than three interquartile range of the 25th and 75th percentile of each group.

We further investigated the mean scores of candidates in the May 2020 administration, according to two different categorisations: by the overall number of times candidates sat GAMSAT (Table 2, with effect size in Table 3), and by testing occasion (Figure 2). In Table 2, the most recent mean test score for each group confirms the expectation that lower performers in the first sitting tend to attempt the test more times than the higher performers (compare, for example, the Overall mean score on their first sitting for the '5 or more' group to the mean score for first time candidates). Additionally, this data demonstrates that while each group tends to increase their mean score in their next test attempt (columns headed 2^{nd} to 5^{th}), the average of the mean scores for each group (final column) tends to generally decrease with each successive attempt. This is consistent with the data displayed as box and whisker plots in Figure 2, showing more clearly the relatively limited increase in scores with successive resits, both in terms of median

overall GAMSAT score as well as the range of scores. In other words, successive attempts do not, on average, lead to higher scores compared with single sitting candidates.

Table 2: GAMSAT mean scores in subsequent attempts (relative to May 2020)

Domain	No of attempts	N	%	1st	2nd	3rd	4th	5th	Mean Score
Overall score	1	2788	38.1%	58.82					58.82
	2	1943	26.5%	57.58	59.70				58.64
	3	1308	17.9%	56.66	58.61	60.22			58.50
	4	599	8.2%	55.35	56.67	58.18	59.67		57.47
	5 or more	686	9.4%	55.18	56.21	57.00	57.94	59.16	57.10
	All	**7324**	**100.0%**	**57.48**	**58.45**	**58.90**	**58.75**	**59.16**	
Section 1	1	2788	38.1%	57.27					57.27
	2	1943	26.5%	55.02	57.72				56.37
	3	1308	17.9%	54.43	55.25	57.67			55.78
	4	599	8.2%	53.99	54.55	55.49	57.48		55.38
	5 or more	686	9.4%	53.27	54.27	54.58	55.02	56.44	54.72
	All	**7324**	**100.0%**	**55.52**	**56.07**	**56.35**	**56.17**	**56.44**	
Section 2	1	2788	38.1%	60.78					60.78
	2	1943	26.5%	59.93	61.76				60.85

	3	1308	17.9%	59.80	61.27	62.34			61.14
	4	599	8.2%	59.41	60.35	61.50	62.00		60.82
	5 or more	686	9.4%	59.17	60.59	60.84	61.45	62.20	60.85
	All	**7324**	**100.0%**	**60.11**	**61.25**	**61.75**	**61.71**	**62.20**	
Section 3	1	2788	38.1%	58.60					58.60
	2	1943	26.5%	57.66	59.67				58.67
	3	1308	17.9%	56.19	58.96	60.44			58.53
	4	599	8.2%	54.07	55.88	57.85	59.56		56.84
	5 or more	686	9.4%	54.18	54.99	56.29	57.67	59.03	56.43
	All	**7324**	**100.0%**	**57.14**	**58.25**	**58.74**	**58.55**	**59.03**	

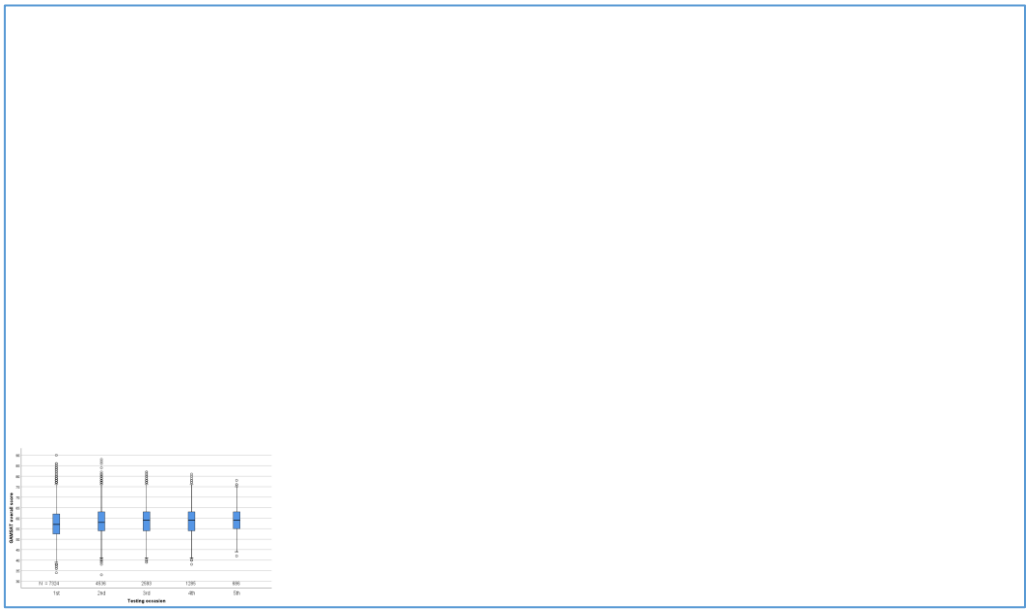

Figure 2: Box and whisker plot showing distribution of GAMSAT overall scores in May 2020 based on testing occasion. (See Figure 1 for explanation of the components of the box and whisker plots.)

The effect sizes tell a similar story. Table 3 presents effect size values (using Cohen's *d*) for score improvement in subsequent attempts for each testing occasion group. (These effect size values can be interpreted as small = 0.2, moderate = 0.5, or large = 0.8 and above; (Cohen, 1988). The data confirm that, in general, the nature of score changes on subsequent testing occasions is relatively small.

Table 3: Summary of effect size of score improvement in subsequent attempts

Domain	No of attempts	2nd-1st	3rd-2nd	4th-3rd	5th-4th
Overall score	2	0.44			
	3	0.56	0.34		
	4	0.32	0.41	0.31	
	5 or more	0.25	0.20	0.26	0.30
	All	**0.42**	**0.31**	**0.29**	**0.30**
Section 1	2	0.52			
	3	0.18	0.47		
	4	0.13	0.21	0.41	
	5 or more	0.23	0.07	0.10	0.29
	All	**0.33**	**0.31**	**0.25**	**0.29**
Section 2	2	0.23			
	3	0.22	0.13		
	4	0.14	0.17	0.06	

	5 or more	0.20	0.04	0.09	0.11
	All	**0.21**	**0.12**	**0.08**	**0.11**
Section 3	2	0.26			
	3	0.48	0.20		
	4	0.27	0.33	0.23	
	5 or more	0.11	0.19	0.23	0.20
	All	**0.29**	**0.22**	**0.22**	**0.20**

Finally, the test-retest correlation between two consecutive attempts is highest for Overall scores (r = 0.77 -0.79) and for Section 3 (r = 0.73 - 0.75), followed by Section 1 (r = 0.67 - 0.69). The correlations are lowest for Section 2 (Written Communication; r = 0.51 - 0.54). All correlations are statistically significant at the 0.01 level. The lower values for Section 2 can be explained by the nature of the construct (written communication), the format (open-ended responses) and the potential variability in topics between versions of the test.

Notwithstanding the relatively small improvements reflected in the above data, it is important to appreciate that even minor score differences may still change the outcome for applicants. This can be more fully appreciated by considering typical scores required for selection into a medicine course. Successful entry scores vary from year to year and between medical schools, but an overall GAMSAT score of 60 usually places candidates in contention for the next phase of the selection process in most medical schools. In May 2020, the mean score of first-time sitters was a little less than this score (Table 2), and while the mean score of those who re-took GAMSAT showed only

slightly improved scores that rarely reached a mean of 60, obviously some resitting candidate will have reached and surpassed this threshold. The implications of extending the score currency are most relevant for candidates with scores around this zone, as we discuss below.

Discussion

Our analysis of GAMSAT re-sit data over a five-year period, pointing to a small degree of improvement in mean scores of re-sitting candidates, concords with previous studies on other high-stakes cognitive ability tests, both medical selection tests (Andrich et al., 2017; Griffin et al., 2019; Lievens et al., 2007; Puddey et al., 2014) and in broader testing contexts (as summarised by Hausknecht et al., 2007 and Scharfen et al., 2018). As Hausknecht and colleagues (2007) note, the magnitude of this effect varies across studies, but the finding of relatively modest improvement with subsequent sittings, in terms of average cohort performance, seems to be a well-established phenomenon. Psychometric data for GAMSAT over a five-year period yielded similar results, notwithstanding some individual large increases in score. Importantly, the improvement appears to be 'asymptotic' (Andrich et al., 2007), with successive sittings generating smaller increases. In the context of GAMSAT and its potential use over a relatively limited period (typically two to five years), the data point to a relatively stable underlying construct, suggesting that extending the currency period will have minimal impact on the relevance and applicability of a GAMSAT score over the usual period of use of GAMSAT scores. Furthermore, given the typical entry score for most medical schools who use GAMSAT as part of their selection process, extending the currency of the test will enable candidates with borderline scores to consider applying to different medical schools with lower entry thresholds, without having to necessarily re-sit the test (and incur further costs).

Nonetheless, for some candidates, even the relatively small mean increases seen in our data may prove to be of practical significance, and obviously this also applies to the less common but more dramatic improvements in scores seen for some individuals. Depending on how scores are used by medical schools, even an increase of a few marks may be sufficient to secure a place in the next phase of the selection process. This is an important caveat in the context of our discussion of extended currency, since in some circumstances, extended currency might be counterproductive if it means that a candidate chooses to forgo the opportunity of a (theoretically small but potentially sufficient) increase, which might turn an unsuccessful score into a successful one (Andrich et al., 2017; Griffin et al., 2019). It remains, of course, up to the candidate to decide whether re-attempting the test might be in their own interests, but it would clearly assist candidates to make informed decisions if testing agencies drew attention to data which suggested that (modest) increases in scores were not uncommon in the performance of resitting candidates.

Accounting for this increase in scores in both our own and wider data is complex and debatable. As noted by several authors, the potential reasons for such improvement are multifactorial, including formal learning in the interim, further development of relevant abilities, increased familiarity with the test, reduced anxiety, recall of previous questions and/or responses, and regression to the mean (Andrich et al., 2017; Hausknecht et al., 2007; Scharfen et al., 2018). As a test of reasoning ability rather than curricular knowledge, we, along with other scholars (e.g. Kuncel & Hezlett, 2007; Huber et al., 2015), would argue that the impact of subsequent learning or ability development is likely to be minimal, as is the memorisation factor, given common practice (including on GAMSAT) to create equated forms with minimal overlap. Furthermore, as discussed above, the impact of coaching has been specifically researched by several authors in the medical selection context, with the accepted

view now falling on the side of minimal impact. Certainly a reduction in anxiety is a very plausible benefit of re-sitting an exam, while regression to the mean cannot be discounted as having a potential dampening effect on the magnitude of the score increase, although this has been estimated as less than 10% of the practice effect size (Hausknecht et al., 2007; Puddey et al., 2014).

In our view, this leaves increased basic *familiarity* with the test content, format and process as the most likely explanation for score improvement on high-stakes cognitive ability tests such as GAMSAT. Such familiarity can be gained through a variety of sources, including use of official practice material, consultation with others who have previously sat the exam, and simply through more time spent in preparation (Kulkarni et al., 2022). Commercial coaching, of course, might also incidentally lead to greater familiarity with the test content (depending on the fidelity of the coaching material compared with the actual test) as well as increased preparation time, although this would appear to be a rather uneconomical strategy for candidates, given the abovementioned findings about limited effectiveness of coaching in the medical education literature beyond basic familiarity (Griffin et al., 2008; Wilkinson and Wilkinson, 2013).

One further source which has so far received minimal discussion in the literature is the dramatically increased familiarity which follows a so-called 'practice run' on cognitive ability test, that is, a registered formal initial sitting of the exam.[5] This phenomenon

[5] This is likely what is intended by the category 'Sat a practice GAMSAT exam' in the research by Kumar et al. (2018), although this is not elaborated in the text. Kulkarni et al. (2022) are more explicit, albeit in the context of a different medical selection exam, noting that: 'It is possible that some candidates choose to have a 'practice run' at the UCAT, sitting the test in the year prior to making their application for medical school, so as to familiarize themselves with the process' (p.3). Given the similar high stakes across medical selection contexts, preparation for GAMSAT is unlikely to be very different.

does not appear to have been formally researched, and therefore its existence and potential impact remains anecdotal, although it is amply represented in online candidate forums and coaching websites.[6] A significant element in this strategy, it seems, is that the test is commonly attempted 'cold', that is, without significant preparation, as a way of ascertaining one's 'base' ability while attempting to benefit from the exposure and experience of an actual sitting.[7] Whatever the practical benefit of this approach in terms of optimising one's score, there can be little doubt that its impact will be to under-estimate the candidates' ability on their first sitting, and therefore, potentially over-estimate the gain in scores from subsequent sittings.

It is not our intention to comment here on the wisdom or even legitimacy of this practice, except to point out that, knowing such a practice amongst first-time candidates occurs, an increased score in a subsequent sitting is perhaps less of a surprise. This provides further context to the magnitude of score improvement we observed through re-sits, in the sense that the magnitude of the increase in scores (especially the higher

[6] By way of example, one medical student/tutor from an online coaching agency gives the following account of their original approach on the GAMSAT: 'I sat the GAMSAT 3 times before I was happy with my scores. The first time was a classic 'practice run'. I had studied a little, but I really didn't know what I was doing or how I should've spent my time. I wasn't overly happy with my results, but I guess I was glad to have understood how the exam and day worked.' (https://medium.com/halad-to-health/gamsat/home)

[7] Concerns with this strategy appear to be behind retesting rules by some testing agencies and/or educational authorities which limit the number of attempts or, more controversially, invalidate scores which differ significantly from a previous sitting (Hausknecht et al., 2007).

outliers in Figures 1 and 2) is likely to be inflated by this deliberate low preparedness on the first attempt. (Certainly this would be a valuable area for research.) Of particular interest for the purposes of this paper are the potential implications of extended currency on candidates adopting this strategy, especially since such an option is likely to be less available to lower SES candidates (for whom the cost of additional sittings may be prohibitive). It may well be that the current limited currency of two years inadvertently encourages such a deliberately under-prepared approach, since any score on an early first attempt (i.e. during the first or second year of the pre-medical degree) will no longer be valid by the time the candidate is eligible to apply for medicine. It may therefore turn out to be a side benefit of extended currency of GAMSAT to allow such 'practice runs' to count should candidates perform better than they expect. The data presented in this paper would suggest that, from a construct validity perspective at least, the results of such 'early attempts', if sufficiently high, would provide a reasonable basis for selection by institutions several years on.

Of course, as mentioned above, extending the currency of selection test scores may not necessarily increase access of lower SES applicants to medical schools, at least not on its own. As Griffin and colleagues (2019) have shown for a similar cognitive ability test for entry into undergraduate medicine, other social and educational factors affect who decides to re-sit a selection test. This is consistent with findings from the widening participation literature, which indicates that many complex factors influence the decision-making of under-represented and/or lower SES applicants considering medicine as a career, including career advice and support, language issues, personal and cultural networks, economic factors and academic assumptions (Cleland et al., 2018; Martin et al., 2018; Tsouroufli & Malcolm, 2015). Appropriate outreach

activities are increasingly recognized as at least as important as any intrinsic changes to the selection process (Martin et al., 2018; McLachlan, 2005).

Finally, testing agencies will need to keep in mind that extending currency of a score may impact on the opportunity and timing of any future changes to the test constructs. While selection test constructs tend to remain consistent over long periods, occasionally significant changes are made (MCAT in 2015 is a case in point – see Kirch et al., 2013) that might make comparability of subsequent scores difficult. Clearly this may need to be factored into not only the timing of changes to score currency, but also the length of currency as well.

Conclusion

To sum up, GAMSAT scores show sufficient stability at the cohort level for institutions to be reasonably confident that a test score will continue to provide a valid representation of cognitive ability for at least up to a five-year period. While there appear to be sound construct, equity and psychometric bases for extending the currency of GAMSAT scores accordingly, there are practical implications that also need consideration. Firstly, the opportunity to use a score with extended currency may actually be relatively limited, depending on the medical school. For schools which rely heavily on selection tests in their selection process and/or set relatively high score cutoffs, an unsuccessful score is unlikely to be successful on re-application. In such cases, it is clearly in the candidate's interest to re-sit and attempt to improve their score. However, for applicants to medical schools with lower selection test cutoffs and/or broader selection criteria, the extended currency would allow time to reconsider their preferences, obtain any further required information, and submit another application within the timeframe of extended score currency. For such candidates, even the minor

improvements that we have observed from re-sits might be sufficient to progress to the next stage of the selection process at some medical schools. Further, our data suggests that it would be appropriate and defensible to allow any score within the currency period to also remain valid, therefore allowing candidates to use their best score in their applications. In other words, a 'no disadvantage' policy re-sits seems warranted.

There are several limitations to our study that should be considered. We have focussed on improvement at the level of mean scores of entire cohorts. An area of further interest would be to explore the impact of specific demographics such as academic background, socioeconomic factors, and ethnicity on score improvement in GAMSAT. This would be particularly useful for helping us to better understand the nature of the resit data and the factors contributing to score improvement. We have also not modelled how extended currency might actually impact on the achievement of typical entry score for candidates, mainly due to the highly varied way in which GAMSAT scores are used and combined with other selection methods by the various medical schools. Finally, we have had to rely on anecdotal evidence about under-prepared first attempts or practice runs which we have suggested may impact the magnitude of score improvements with subsequent sittings. Clearly our conclusions would be better supported if we had stronger evidence for this practice, but it should nevertheless be borne in mind as another potential factor. This would be a highly desirable area for future research.

Determining the optimal currency of a cognitive ability test for selection purposes is therefore a balance of several competing purposes, considerations and complex data. It may also present somewhat of a paradox if the primary intention is to set a currency period which will facilitate access to medicine for lower SES students. While 'pulling the levers' on selection policies by extending score validity and making more practice

material freely available may assist with some of the barriers currently faced by lower SES applicants, these are only some of the elements which can impact on widening participation into medicine. As noted above, widening participation requires a significant shift in policy to include suitable recruitment and outreach strategies. In the meantime, however, we argue that removing any unnecessarily restrictive currency of scores, regardless of how small the potential impact, is a step in the right direction.

References

Aldous, C.J., Leeder, S.R., Price, J., Sefton, A.E., & Teubner, J. K. (1997). A selection test for Australian graduate-entry medical schools. *Medical journal of Australia, 166*(5), 247-250.

Andrich, D., Styles, I., Mercer, A., & Puddey, I.B. (2017). On the validity of repeated assessments in the UMAT, a high-stakes admissions test. *Advances in Health Sciences Education, 22,* 1245–1262.

Australian Council for Educational Research (ACER). (2022). *GAMSAT Information Booklet*. https://gamsat.acer.org

Camara, W. J., & Mattern, K. (2022). Inflection Point: the role of testing in admissions decisions in a postpandemic environment. *Educational Measurement: Issues and Practice, 41*(1), 10-15. doi:https://doi.org/10.1111/emip.12493

Castro, M. R. H., Calthorpe, L. M., Fogh, S. E., McAllister, S., Johnson, C. L., Isaacs, E. D., Ishizaki, A., Kozas, A., Lo, D., Rennke, S., Davis, J, & Chang, A. (2021). Lessons from learners: adapting medical student education during and post COVID-19. *Academic Medicine, 96*(12), 1671-1679. doi:10.1097/acm.0000000000004148

Cleland, J. A., Patterson, F., & Hanson, M. D. (2018). Thinking of selection and widening access as complex and wicked problems. *Medical Education*, 52(12), 1228-1239. doi:10.1111/medu.13670

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (*2nd ed*.). Hillside, NJ: Lawrence Erlbaum Associates.

Corridon, P., R. (2021). Admissions to medical school during the COVID-19 era without the MCAT. *Journal of Medical Education and Curricular Development, 8.*

Elliott, S. L., & Epstein, J. (2005). Selecting the future doctors: the role of graduate medical programmes. *Internal Medicine Journal, 35*(3), 174-177. doi:10.1111/j.1445-5994.2004.00796.x

Fielding, S., Tiffin, P. A., Greatrix, R., Lee, A. J., Patterson, F., Nicholson, S., & Cleland, J. (2018). Do changing medical admissions practices in the UK impact on who is admitted? An interrupted time series analysis. *BMJ Open, 8:*e023274.

Fyfe, M., Horsburgh, J., Blitz, J., Chiavaroli, N., Kumar, S. & Cleland, J. (2022). The do's, don'ts and don't knows of redressing differential attainment related to race/ethnicity in medical schools. *Perspectives on Medical Education, 11*(1):1-14. doi: https://doi.org/10.1007/S40037-021-00696-3

Girotti, J. A., Park, Y. S., & Tekian, A. (2015). Ensuring a fair and equitable selection of students to serve society's health care needs. *Medical Education, 49*(1):84-92.

Griffin, B., Harding, D.W., Wilson, I.G., & Yeomans, N.D. (2008). Does practice make perfect? The effect of coaching and retesting on selection tests used for admission to an Australian medical school. *Medical Journal of Australia, 189*.

Griffin, B., Auton, J., Duvivier, R., Shulruf, B., & Hu, W. (2019). Applicants to medical school: if at first they don't succeed, who tries again and are they successful? *Advances in Health Sciences Education, 24*, 33-43. doi:10.1007/s10459-018-9847-9

Griffin, B. (2018). Coaching Issues. In: F. Patterson & L. Zibarras (Eds). *Selection and recruitment in the healthcare professions: research, theory and practic*e (pp. 223-248). Palgrave Macmillan.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, 92(2), 373–385

Ho, C., Hu, W., & Griffin, B. (2022). Cultures of Success: How elite students develop and realise aspirations to study Medicine. *The Australian Educational Researcher,* 1-21. doi:10.1007/s13384-022-00548-x.

Huber, C. R., Kuncel, N. R., Sackett, P. R., & Beatty, A. S. (2015). Validity stability across entering college cohorts: exploring the temporal generalizability of local validity estimates. *International Journal of Selection and Assessment, 23*(3), 237-246. doi:10.1111/ijsa.12111

Kelly, M., Tiffin, P., & Mwandigha, L. (2018). Aptitude testing in healthcare selection. In: F. Patterson & L. Zibarras (Eds). *Selection and recruitment in the healthcare professions: research, theory and practic*e (pp. 27-50). Palgrave Macmillan.

Kirch, D. G., Mitchell, K., & Ast, C. (2013). The new 2015 MCAT: testing competencies. *JAMA, 310*(21),2243–2244. doi:10.1001/jama.2013.282093.

Kreiter, C. D., & Kreiter, Y. (2007). A validity generalization perspective on the ability of undergraduate GPA and the Medical College Admission Test to predict important outcomes. *Teaching and Learning in Medicine, 19*(2), 95-100.

Kulkarni, S., Parry, J., & Sitch, A. (2022). An assessment of the impact of formal preparation activities on performance in the University Clinical Aptitude Test (UCAT): a national study. *BMC Medical Education*, 22(1), 747. doi:10.1186/s12909-022-03811-y

Kumar, K., Roberts, C., Bartle, E., & Eley, D. S. (2018). Testing for medical school selection: What are prospective doctors' experiences and perceptions of the GAMSAT and what are the consequences of testing? *Adv Health Sci Educ Theory Pract, 23*(3), 533-546. doi:10.1007/s10459-018-9811-8.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315*, 1080-1081.

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19*(6), 339-345. doi:10.1177/0963721410389459

Lam, J. T. H., Hanson, M. D., & Martimianakis, M. A. (2020). Exploring the socialization experiences of medical students from social science and humanities backgrounds. *Academic Medicine, 95*, 401-410

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, 92, 1672–1682.

Lohman, D. (2004). *Aptitude for college: the importance of reasoning tests for minority admissions*. In R. Zwick (Ed.), *Rethinking the SAT: the future of standardized testing in university admissions*. Routledge.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Martin, A., J., Beska, B., J., Wood, G., Wyatt, N., Codd, A., Vance, G., & Burford, B. (2018). Widening interest, widening participation: factors influencing school students' aspirations to study medicine. *BMC Medical Education, 18*(1),117

McDonald, A., Newton, P., Whetton, C., & Benefield, P. (2000) *Aptitude testing for university entrance: a literature review*. National Foundation for Educational Research.

McGaghie, W. C., Downing, S. M., & Kubilius, R. (2004). What is the impact of commercial test preparation courses on medical examination performance? *Teaching and Learning in Medicine, 16*(2), 202–211.

McLachlan, J. C. (2005). Outreach is better than selection for increasing diversity. *Medical Education, 39*(9), 872-5.

McManus, I. C. (1992). Does performance improve when candidates resit a postgraduate examination? *Medical Education, 26*(2), 157-62.

Medical Deans of Australia and New Zealand (MDANZ). (2021). *Medical Schools Outcomes Database: National Data Report 2021*.

Mercer, A., Crotty, B., Alldridge, L., Le, L., & Vele, V. (2015). GAMSAT: A 10-year retrospective overview, with detailed analysis of candidates' performance in 2014. *BMC Medical Education, 15*(31), 1-9.

Nicholson, S. (2005). Commentary: The benefits of aptitude testing for selecting medical students. *BMJ: British Medical Journal, 331*, 559-560.

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers. College. Record, 112*, 1137–1162.

Puddey, I. B., Mercer, A., Andrich, D., & Styles, I. (2014). Practice effects in medical school entrance testing with the undergraduate medicine and health sciences admissions test (UMAT). *BMC Medical Education, 14*.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rees, E., & Woolf, K. (2020). Selection in context: the importance of clarity, transparency and evidence in achieving widening participation goals. *Medical Education, 54*(1), 8-10. doi:10.1111/medu.14023

Ross, D., Loeffler, K., Schipper, S., Vandermeer, B., & Allan, G. M. (2013). Do scores on three commonly used measures of critical thinking correlate with academic success of health professions trainees? A systematic review and meta-analysis. *Academic Medicine, 88*(5), 724-734.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: a meta-analysis. *Intelligence, 67*, 44-66.

Schleicher, D. J., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2010). If At First You Don't Succeed, Try, Try Again: Understanding Race, Age, and Gender Differences in Retesting Score Improvement. *Journal of Applied Psychology, 95*(4), 603-617. doi:10.1037/a0018920.

Stemler, S. (2012). What should university admissions tests predict? *Educational Psychologist, 47*(1), 5-17.

Stemler, S., & Sternberg, R. (2013). *The assessment of aptitude. APA handbook on testing and assessment.* American Psychological Association.

Stringer, N. (2008). Aptitude tests versus school exams as selection tools for higher education and the case for assessing educational achievement in context. *Research Papers in Education, 23*,53-68.

Thomson, S. (2018). Achievement at school and socioeconomic background—an educational perspective. *npj Science Learn* 3(5). https://doi.org/10.1038/s41539-018-0022-0.

Tsouroufli, M., & Malcolm, I. (2015). Equality, diversity and fairness in medical education: international perspectives. *Medical Education*. *49*(1),4-6.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement,* 26(1), 77–87.

Wilkinson, T. M., & Wilkinson, T. J. (2013). Preparation courses for a medical admissions test: effectiveness contrasts with opinion. *Medical Education, 47*(4), 417-24.

Younger, K., Gascoine, L., Menzies, V., & Torgerson, C. (2019). A systematic review of evidence on the effectiveness of interventions and strategies for widening participation in higher education. *Journal of Further and Higher Education,* 43(6), 742-773. doi:10.1080/0309877X.2017.1404558