Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education xxx (xxxx) xxx

www.elsevier.com/locate/hpe

In God We Trust, All Others Bring Data: A Bayesian Approach to Standard Setting

Jimmie Leppink

Hull York Medical School, University of York, Heslington, York YO10 5DD, UK

Received 9 November 2019; revised 17 January 2020; accepted 26 January 2020

Abstract

Standard setting is an inherent part of pass/fail decisions in assessment. Although various standard setting methods are available, they all have their limitations and no method provides a golden solution to all our standard setting headaches. Some methods require potentially labor-intensive standard setting panels of judges who have specific knowledge. Other methods require student cohorts of ‘sufficient’ size. However, small cohorts are quite prevalent in medical programs across the globe, and standard setting panels are not always feasible due to logistic or financial constraints or may result in inadequate judgments due to bias or a lack of specific knowledge. This manuscript presents a new standard setting method, which is based on the Bayesian principle of updating our knowledge or beliefs about a phenomenon of interest with incoming data, uses information that is not considered in methods already available and can be applied to both small and larger cohorts regardless of whether standard setting panels are available. As demonstrated in this manuscript through a worked example, the new method is easy to implement and requires only a minimum of calculations which can be done in zero-cost, user-friendly Open Source software. Options for future research comparing different standard setting methods are discussed.

Keywords: Standard setting; Small cohorts; Performance data; Bayesian; Updating

Introduction

Standard setters do not have it easy. Although we have quite a variety of standard setting methods at our disposal, there is no gold standard method that provides a solution for all standard setting problems. Whichever method we choose, we face problems. For example, the Angoff method and modifications thereof (e.g.,[1–6]) require potentially labor-intensive standard setting panels of judges who have a clear understanding of the concept of ‘borderline student’. Beuk's method[7] assumes that each judge has an opinion of what the cut-off score or pass mark should be and what pass rate should be expected. The Hofstee method[8] assumes that judges have specific knowledge of minimum and maximum acceptable cut-offs and fail rates and in practice cut-off scores do not always fall within the defined boundaries. Other methods require sufficiently large cohorts[9,10] and/or the assessment in question to be organized in a very particular way and to result in outcomes that reasonably meet specific

E-mail address: hyjl17@hyms.ac.uk.

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2020.01.003

features.[11] However, small cohorts are quite prevalent in medical programs across the globe, standard setting panels are not always feasible due to logistic or financial constraints or may result in numbers that have no real empirical basis, and assessment data do not always adhere to specific features required for other standard setting methods to be used.

This manuscript presents a new standard setting method, which is based on Bayesian statistics,[12] uses information that is not considered in methods already available and can be applied to both small and larger cohorts regardless of whether standard setting panels are available. As demonstrated in this manuscript through a worked example, the new method is easy to implement and requires only a minimum of calculations which can for example be done in zero-cost, user-friendly Open Source software called JASP.[13] Although the example focusses on setting a standard with multiple-choice or other single best answer questions (SBAQs) that can be coded as either 'correct' or 'incorrect' (i.e., a dichotomous decision for each question) for the ease of introduction, this new method can be generalized to other types of assessments as well.

Bayesian statistics in a nutshell

Just like in everyday life, Bayesian statistics is about updating our knowledge or beliefs about a phenomenon of interest, say X, as data is coming in. At any given point in time, our knowledge or beliefs about X are expressed in a probability distribution of possible outcomes of X. In the absence of any empirical data, quite a variety of outcomes of X may be likely, but with more data coming in, some outcomes become more likely while other outcomes become less likely. Very simply put, Bayesians refer to the probability distribution of X *before* seeing the data as the *prior* distribution, which is then updated with data coming in to obtain the *posterior* distribution or the probability distribution of X *after* seeing the data. However, as in everyday life, the Bayesian updating process is continuous; in the words of Lindley[14] (p. 2), "*today's posterior is tomorrow's prior*." In the simplest case, the outcome of interest is dichotomous, for example 'Correct' (success) vs. 'Incorrect' (failure) performance of a SBAQ. The probability distribution to be used for updating is then a so-called *Beta* distribution with two parameters: the number of successes (a) and the number of failures (b). We denote this as: Beta(a,b). In the absence of any knowledge or belief about the outcome of interest, the appropriate Beta distribution is Beta(1,1), which is a Uniform (i.e., rectangular) probability distribution extending from 0 to 1 (i.e., from 0% to 100% success). This distribution has a mean and median of 0.5 (i.e., 50% success) and a 95% *credible interval* (the Bayesian counterpart of the Frequentist confidence interval, which is also called posterior interval) of [0.025; 0.975]. When we observe data, outcomes in some areas of the 0−100% range become more likely while outcomes in other areas of the 0−100% range become less likely. Let us look at this with an example of coin tossing, where 'Heads' represents success and 'Tails' represents failure.

Coin tossing as an example

To assess if coin X is 'fair', we toss it several times and count the number of successes and failures. In the absence of any data on coin X, we use Beta(1,1) as prior distribution. Suppose, we throw ten times and obtain six successes and four failures. The resulting posterior distribution is Beta(1+6,1+4), hence Beta(7,5). This distribution has a median of 0.588 and a 95% credible interval of [0.308; 0.833]. Suppose, we repeat this coin tossing study nine times, meaning that we end up with a total of 10 sets of 10 tosses, or 100 observations in total. If prior to reading this article you never heard of Bayesian statistics before, and you find this process difficult to understand, do not worry, for Table 1 presents the outcomes for each of ten rounds along with the prior and posterior distribution for the different rounds.

Had we considered the 10 times 10 tosses as one study of 100 tosses, we would have obtained the same posterior distribution: Beta(1,1) updated with 54 successes and 46 failures (our data) results in a Beta(55,47) posterior distribution. This posterior

Table 1
Prior and posterior distribution for the proportion of success (Heads) for each of ten rounds in the coin tossing study.

Prior: Beta(a,b)		Data		Posterior: Beta(a,b)	
a (success)	b (failure)	Successes	Failures	a (success)	b (failure)
1	1	6	4	7	5
7	5	5	5	12	10
12	10	7	3	19	13
19	13	5	5	24	18
24	18	5	5	29	23
29	23	7	3	36	26
36	26	5	5	41	31
41	31	4	6	45	37
45	37	4	6	49	43
49	43	6	4	55	47

Please cite this article as: Leppink J, In God We Trust, All Others Bring Data: A Bayesian Approach to Standard Setting, Health Professions Education, https://doi.org/10.1016/j.hpe.2020.01.003

distribution has a median of 0.539 and a 95% credible interval of [0.442; 0.635]. The median of the posterior distribution is commonly used as a point estimate of the outcome of interest and is almost equal to the observed proportion of success (0.54); it is slightly pulled towards 0.5 and more so with the observed proportion approaching either 0 or 1 but more closely approaches the observed proportion with more observations.

The rationale behind the latter is that even if our outcomes are as extreme as 0% or 100%, we will only gain more confidence in outcomes towards either of these extremes with increasing amounts of data. If Lecturer A claims that "*all students from this university got drunk last night*", Lecturer B asks "*how did you estimate this?*", and Lecturer A responds "*well I asked a random sample of three students*", we do not have much confidence in the estimate of 100% of the students being drunk, or any extreme estimate close to that. This situation would translate in a Beta(1+3,1+0) = Beta (4,1) posterior distribution, which is a distribution with a median of 0.841 and a 95% credible interval of [0.398; 0.994]. However, if Lecturer A's random sample comprised 100 students, all of which got drunk, our confidence in a near-100% estimate would be much larger; the resulting posterior distribution would be Beta(101,1), which has a median of 0.993 and a 95% credible interval of [0.964; 1.000]. These and other numbers from different Beta distributions can be easily obtained in JASP.[13]

Applying the coin tossing example to item performance

Mathematically, N number of students responding to k number of items can − to some extent − be compared to R rounds of k number of coin tosses. In other words, the 10 rounds of 10 coin tosses could also be thought of as 10 students responding to a set of 10 items. While the different rounds of coin tossing all use the same coin, in the case of items it would be 10 students responding to the same set of 10 items. Just like the Beta(55,47) distribution constitutes the posterior distribution after the 100 tosses in the coin tossing study, we can derive a similar Beta distribution for the set of items answered by the group of students. The median and 95% credible interval of this Beta distribution can then be used to estimate the proportion of correct response for a given set of items to be put in an exam. Under the assumption that there are no large differences between cohorts of students, we can even derive a Beta posterior distribution if different items are used in different cohorts or there is only a partial overlap in items across cohorts.

Consider the following situation. For a cohort of $N = 23$ students, we are creating an exam of 100 SBAQs, drawing from an item bank of SBAQs used in exams in the previous three cohorts, which were of size $N = 20$, $N = 25$, and $N = 30$, respectively. Suppose, we use 25 items that were used only in the cohort of $N = 20$, 25 items that were used in the cohort of $N = 20$ and in the cohort of $N = 25$, 25 items that were used in the cohort of $N = 25$ and in the cohort of $N = 30$, and 25 items that were used only in the cohort of $N = 30$. To facilitate the understanding of the calculations below as in the coin tossing study, Table 2 presents the numbers of successes and failures observed for each of these in total 100 items.

Some may wonder if we cannot just calculate the proportion of success for each item and take the average of the 100 proportions calculated. For instance, for item 1, the proportion of success is 11 successes divided by 20 observations or 0.550, and for item 28 it is 27 successes divided by 45 observations or 0.600. Using this method results in a total of 59.4 across the 100 items, hence an average of 0.594. However, one problem with this method is that we

Table 2
Numbers of successes (S) and failures (F) observed for each of 100 items.

Item	*S*	*F*	Item	*S*	*F*	Item	*S*	*F*	Item	*S*	*F*
1	11	9	**26**	21	24	**51**	20	35	**76**	20	10
2	14	6	**27**	30	15	**52**	29	26	**77**	16	14
3	10	10	**28**	27	18	**53**	33	22	**78**	25	5
4	17	3	**29**	32	13	**54**	36	19	**79**	20	10
5	15	5	**30**	27	18	**55**	19	36	**80**	19	11
6	16	4	**31**	21	24	**56**	18	37	**81**	19	11
7	14	6	**32**	31	14	**57**	34	21	**82**	17	13
8	10	10	**33**	30	15	**58**	48	7	**83**	22	8
9	13	7	**34**	34	11	**59**	27	28	**84**	24	6
10	9	11	**35**	14	31	**60**	27	28	**85**	20	10
11	10	10	**36**	23	22	**61**	36	19	**86**	21	9
12	13	7	**37**	24	21	**62**	16	39	**87**	22	8
13	9	11	**38**	21	24	**63**	41	14	**88**	19	11
14	8	12	**39**	21	24	**64**	26	29	**89**	22	8
15	13	7	**40**	12	33	**65**	30	25	**90**	5	25
16	15	5	**41**	34	11	**66**	45	10	**91**	22	8
17	12	8	**42**	23	22	**67**	24	31	**92**	20	10
18	9	11	**43**	18	27	**68**	34	21	**93**	10	20
19	12	8	**44**	32	13	**69**	37	18	**94**	23	7
20	13	7	**45**	24	21	**70**	43	12	**95**	14	16
21	16	4	**46**	24	21	**71**	38	17	**96**	26	4
22	11	9	**47**	20	25	**72**	30	25	**97**	19	11
23	15	5	**48**	29	16	**73**	34	21	**98**	14	16
24	7	13	**49**	27	18	**74**	40	15	**99**	18	12
25	14	6	**50**	25	20	**75**	27	28	**100**	18	12

have more observations for some items than for other items. An easy way to account for the latter is to count the number of successes across 100 items and divide that by the total number of observations for the 100 items together: 2197 successes divided by 3750 observations results in an average of 0.586. Following the Bayesian procedure, 2197 successes and 1553 failures results in a Beta(1+2197,1+1553) = Beta(2198,1554) posterior distribution, which is a distribution with a median of 0.586 and a 95% credible interval of [0.570; 0.602]. Given the large number of observations, the median of the posterior distribution and the average obtained when accounting for unequal numbers of observations across items are about the same.

Accounting for intra-class correlation

There is one problem with the coin tossing analogy. Differences between students in knowledge of a given topic contribute to an *intra-class correlation* (ICC) that reduces the sample size from *N* times *k* to an effective sample size somewhere in between *N* and *k* and should be accounted for.[15] Given *N* students responding to *k* items, and an estimated ICC, the factor of difference between *N* times *k* and the effective sample size (*F*) can be calculated as follows:

$$F = 1 + [(k-1) * \mathrm{ICC}]$$

Thus, larger values of *k* and higher ICCs contribute to a stronger reduction in effective sample size. When dealing with larger cohorts, ICC can be estimated from the data using multilevel models,[15] but when samples are small ICC estimates often cannot be trusted and therefore ICC needs to be estimated in another way. For series of 100 or more SBAQs, ICC values in the [0.05; 0.10] range are common. Consequently, if we do not have sufficient data to obtain accurate ICC estimates from the data, we can use ICC = 0.10 as a conservative estimate. For $k = 100$ and ICC = 0.10, $F = 10.9$. We can recalculate the ICC-adjusted number of successes (a_{ICC}) and number of failures (b_{ICC}) by dividing the observed number of successes (a) and the observed number of failures (b), respectively, by *F*. The resulting posterior distribution is then Beta($1+a_{ICC}$,$1+b_{ICC}$). Doing so for the data at hand, where a = 2197 and b = 1553, we find $a_{ICC} = 201.560$ and $b_{ICC} = 142.477$, and a posterior distribution Beta(202.560,143.477), which is a distribution with a median of 0.586 and a 95% credible interval of [0.533; 0.637].

If we were to deal with a situation where ICC is likely to be larger than 0.10 (uncommon but nevertheless possible), *F* would be larger. For instance, for $k = 100$ and ICC = 0.20, $F = 20.8$. The resulting posterior distribution would then be Beta(106.625,75.663), which is a distribution with median 0.585 and a 95% credible interval of [0.513; 0.655]. Note that the median of the distribution remains almost the same, and even the 95% credible interval is not that different from the one assuming ICC = 0.10.

Setting the standard

The median of the posterior distribution provides a straightforward statistic for standard setting purposes, and the 95% credible interval expresses the degree of uncertainty around that statistic. However, using the posterior distribution median itself as a standard is problematic as it may well result in a substantial proportion of sufficiently competent students failing the exam. If we agree that the average student is sufficiently competent and we only want not sufficiently competent students to fail the exam, we need a multiplier to arrive from the posterior distribution median to the pass mark that allows for some deviation downward from the mean but simultaneously minimizes the risk of not sufficiently competent students passing the exam. If we use 0.8 (i.e., 80%) as multiplier, a posterior distribution median of 0.625 results in a pass mark of 0.500 (50%) and a posterior distribution median of 0.500 results in a pass mark of 0.400 (40%). Although these pass marks appear low, pass marks are calculated in the light of the relative difficulties of items. Larger proportions of relatively difficult items result in a lower posterior distribution median and should therefore result in a somewhat lower pass mark. Multipliers larger than 0.8 could come at the cost of a considerable proportion of sufficiently competent students failing the exam, whereas multipliers smaller than 0.8 could come at the risk of a rather substantial proportion of not sufficiently competent students passing the exam. Using 0.8 as multiplier, a posterior distribution median of 0.586 results in a pass mark of 0.469 (46.9%).

Some readers may wonder what we should do if some of the items in an exam for a new cohort have not been used before, and we therefore have no performance data available. The answer to this is that items for which we have no performance data add neither successes nor failures and thus do not contribute to the posterior distribution. This will result in a wider 95% credible interval, and especially in the case of small numbers of observations, a posterior distribution median slightly pulled towards 0.5 (e.g., see coin tossing study). This is reasonable, because we simply have less information about our set of items if we have no performance data about some items. If we had only new items, hence no item performance data at all, the posterior distribution would be Beta(1,1), which has a median of 0.500, and hence the standard would be set at 0.400 (40%).

Although the example used in this manuscript is one of dichotomous outcomes, the method introduced in this manuscript can be generalized to other types of outcome variables; we will need to use different types of prior distributions (e.g.,[15],[16]), but the basic idea of updating our prior distribution with data coming in remains the same. Another implication of this is that, while the example in this article focusses on knowledge tests, the method can also be applied to skills tests such as objective structured clinical examinations or procedural tests in the workplace.

Potential concerns and challenges with this new method

As for any standard setting method, clear concerns and challenges can be identified. A great feature of peer review is that questions and concerns can be shared by the reviewers of a manuscript. In the case of this manuscript, which proposes a new method for standard setting, the reviewers raised a series of very important questions about, concerns with and arguments against this new method. This paragraph summarizes these questions, concerns, and arguments with a concise response, which in some cases is a temporary response because future research will have to shed more light on the matter as discussed in the next paragraph.

Point 1/10: Since this method differs from all established standard setting methods, is this new method really a standard setting method?

Some scholars hold the opinion that methods like the new method proposed in this article do not qualify as standard setting methods. In their view, standard setting inherently involves having subject matter expects provide ratings to determine cut-off scores and this is generally accomplished using one of three general categories of methods: reviewing test items or groups of test items (test-centered methods), reviewing candidate work or giving ratings on how examinees would be expected to perform (examinee-centered methods) or compromise methods which usually involve providing judgments about the percentage correct and percentage of examinees expected to pass. The new method proposed in this article does not fall into any of these categories, because in the words of William Edwards Deming (1900−1993), "*Without data, you are just another person with an opinion*" and "*In God we trust; all others bring data*."Angoff, Hofstee, Beuk and variants thereof heavily rely on so-called 'expert' opinions. But who tell us they are experts? The experts!

Several scholars are against empirical methods such as Cohen[9] or modified Cohen,[10] in which the fail/pass cut-off score is determined by taking a multiplier from the 95th[9] or 90th[10] percentile of the score distribution of a student cohort at hand, because of the normative character of percentiles. However, contrary to traditional normative methods, where a fixed proportion of students can be expected to fail the exam, the proportion of students failing with Cohen or modified Cohen can vary across exams and can be zero if there is there relatively little dispersion in scores within the cohort at hand. Although Cohen and modified Cohen, like any standard setting method, have their issues as well, they do not rely on assumptions of so-called expert judgments resulting in meaningful and accurate cut-off scores. Although contrary to Cohen and modified Cohen, the new method proposed in this article uses *historical* performance data instead of the performance of a cohort at hand, these methods have in common that they use actual performance data instead of data-absent 'expert' judgments.

Point 2/10: What is the underlying philosophy of the paper when it comes to item difficulty and the appropriate standard?

Every item in an exam can be thought of as a battle between the student and the item. Using quality items, (1) given the knowledge level of a student, the more difficult the item the higher the probability of the item winning the battle and (2) given the difficulty level of an item, the more knowledgeable the student the higher the probability of the student winning the battle. If a student's knowledge level is the same as the level of difficulty of an item, both the student and the item have a probability of winning the battle of 50% (i.e., 0.50). Usually, exams are composed such that there are more items responded correctly by 50−70% of the students than there are items responded correctly by a much smaller or a much larger percentage of students in a cohort, and the average proportion of items responded correctly in a cohort of students lies around 60% (0.6).

Students performing around the average, at least in medical schools, are usually students who have sufficient knowledge (or skill, where skill is assessed) and should therefore be placed above the fail/pass cut-off score for an exam. Borderline performance is usually found a standard deviation or so below the average. Using modified Cohen, in which the fail/pass cut-off score is 65% of the 90th percentile, we usually find a cut-off score of nearly 80% of the average performance. In programs where cohort differences are small, the average performance of a new cohort will not differ much from that averaged across cohorts for which we have historical data. If a new cohort is much better than previous cohorts, the fail/pass cut-off determined using historical data will probably result in a lower proportion of students (and perhaps even zero) failing the exam. Simultaneously, if a new cohort is much worse than previous cohorts, the fail/pass cut-off determined using historical data will probably result in a higher proportion of students than usual failing the exam. This marks an important difference with Cohen and modified Cohen, where the standard set entirely depends on the performance of a cohort at hand and no historical data is used.

Point 3/10: Does the standard exist, but we do not have much information to go on in small cohorts, or are we always adjusting the appropriate standard as expectations of performance/knowledge in different areas of medicine develop over time?

Our expectations of knowledge and performance in different areas of medical do develop over time and empirical performance data serve as an important reality check; if the latter demonstrate that performance in cohorts of students is not up to standards, we may want to reflect on the way we are teaching and/or assessing in our programs, but failing larger numbers of students just because higher expectations call for a higher standard would be unfair to our students.

Point 4/10: Where does the 80% multiplier come from?

The 80% multiplier is in line with the statistical notion that even though modified Cohen does not use the average score for setting the standard it normally results in a fail/pass cut-off score of almost 80% of the average performance. For the reader who wonders how to arrive at this conclusion without giving a reference, this can be easily checked with any statistical software package by simulating score distributions for a number of cohorts in line with what we commonly encounter in exams: fairly symmetric with one fairly clear peak somewhere around the average performance. Calculate the 90th percentile, take 65% of that, and the resulting score (i.e., the fail/pass cut-off in modified Cohen) should be nearly 80% of the average of the distribution.

Some readers may wonder why if the multiplier of 80% is in line with modified Cohen, why not use modified Cohen instead of the new method proposed in this article. The reason for that is that modified Cohen may be fine if we are okay relying on the current cohort only and the cohort is sufficiently large, where we have performance data from multiple previous cohorts we have much more information about item performance than we can derive from a current cohort, we do not depend on the performance of a current cohort, and the new method can be used in a meaningful manner in smaller cohorts as well.

Point 5/10: What if the assumption about no large differences between cohorts does not hold?

If among historical cohorts there is considerable variation in performance, that probably indicates that some cohorts are somewhat stronger than other cohorts. Regardless of how much that cohort-to-cohort variation is, if we carefully sample items used in previous cohorts for the current cohort, we will use information from all these cohorts to set an informative standard. If then it turns out that the new cohort performs much better than expected based on the historical data, this probably indicates we are dealing with a relatively strong cohort and it will in that case make sense to fail relatively few students. Likewise, if the current cohort performs much more poorly than expected based on the historical data, this may well indicate that we are dealing with not such a strong cohort where in that case more students can be expected to fail. That said, the influence of cohort differences and methods of sampling items for an exam for a new cohort certainly constitute topics for further research.

Point 6/10: It is widely known that prior distributions can heavily influence results, and this is especially problematic in standard-setting contexts because they can introduce bias to cut scores. How does this new method deal with this problem?

It is also widely known that prior distributions will heavily influence estimates only if we have a very small sample (e.g., three cohorts of five students, or a statement like "*all students from this university got* *drunk last night*" being based on just a few students) or use a much more informative prior distribution than we should use. Neither is the case with the method proposed in this article. Even though our cohorts are smaller than what we usually encounter in medical programs, all items from the different cohorts taken together we have quite a bit of data (e.g., not three cohorts of five students). The coin tossing example shows that even with only 100 tosses the difference between the posterior median and the observed probability of success is very small (in the third decimal), and despite the small cohort sizes the number of observations we have on our set of items by far exceeds 100. Concerning the prior distribution, the Beta(1,1) distribution assumes no prior knowledge and is therefore widely recognized and used as a default prior distribution (e.g.,[13,17–19]), because any kind of bias introduced is very minimal at most. For that matter, a biased panel incorrectly assuming good knowledge about how students should perform (e.g., in Angoff panels, what percentage of 'borderline' students would be expected to respond item X in exam A correctly) could be expected to result in much more bias in standards set than a method that uses actual performance data and a default prior distribution that assumes no prior knowledge.

Point 7/10: Should Bayesian methods not only be used if one has a lot of prior knowledge about examinee performance and Bayesian methods are also being used to score the exam?

Researchers have been using Bayesian methods to practice statistics regardless of how students are taught or are assessed, regardless of what people eat or how they sleep, and both in the presence and absence of prior knowledge about the phenomenon of interest. Bayesian methods are intuitive because they work like the human mind: we update our knowledge or beliefs as new information comes in. The thought that we could not use Bayesian methods unless we have a lot of prior knowledge would be like saying that it is pointless to teach medical students about any kind of medical topic unless they come to class with a lot of prior knowledge about the topic already. Reference or default priors have been agreed exactly to enable Bayesian analysis where little or even no prior knowledge about the phenomenon is available, just like most teaching of new topics in a medical program have been designed to help students make the transition from little to no knowledge about the for them new topic to a state of being more knowledgeable.

Point 8/10: The approach is entirely normative, and results will change when the test is easier or harder. How is this defendable or fair?

The approach is neither normative in the sense of traditional methods in which a fixed proportion of students can be expected to fail nor in the way Cohen methods work basing standards entirely on numbers from a cohort at hand. Like with Cohen methods, everyone meeting the standard and therefore passing the exam is possible in this new method, but the standard set does *not* depend on any numbers from a cohort at hand; it uses data from previous cohorts only. Results will indeed change when a test is easier or harder, which makes sense. On the contrary, if an easier and a harder test resulted in the same standard, that would understandably raise concerns about the group receiving the harder test being disadvantaged or the group receiving the easier test being put in an unreasonable advantage. Besides, having historical performance data in place, if we carefully sample items used in previous cohorts for the current cohort, we can produce exams that in terms of difficulty do not differ that much from each other. Using judgment without empirical data to compose exams may more easily result in substantial exam-to-exam fluctuation in difficulty than careful sampling based on historical data.

Point 9/10: What to do, with software or else, if ICC = 0.10 is not a realistic assumption?

Especially in sufficiently large exams (e.g., 100 or more multiple-choice items) that cover a series of only modestly related topics, ICC = 0.10 is not an unreasonable assumption. However, as the comparison between calculations assuming ICC = 0, ICC = 0.10 and ICC = 0.20 indicates, even for samples as small as the ones in the example discussed in this article the posterior median (from which 80% is taken to set the standard) is virtually the same. ICCs of 0.20 or higher are not common in large SBAQ exams, but if assessors have solid (and preferably empirically supported) reasons to assume that in the context they are working higher ICCs are common, the formula can be used for higher ICCs as well and the resulting corrected numbers of successes and failures can be entered in JASP or any other software that allows researchers to obtain a distribution with a posterior median and 95% credible interval. Generally speaking, the more historical data we have available, either through more historical cohorts or larger historical cohorts, the less the posterior median will be affected by higher ICCs.

Please cite this article as: Leppink J, In God We Trust, All Others Bring Data: A Bayesian Approach to Standard Setting, Health Professions Education, https://doi.org/10.1016/j.hpe.2020.01.003

That said, the question of the influence of ICC assumptions on the posterior median and standard set deserves further study.

Point 10/10: What if an exam is very difficult such that the average student misses most of the questions: should this student pass?

Especially with the use of historical data, with which we can carefully sample our items for any new cohort − or for any individual re-sitting student for that matter − this scenario should never happen; instead, we could sample such that we have a paper for which historical data indicate an average performance of around 62.5% and hence a standard of 50% (i.e., 80% of 62.5%) would be defendable.

Future research

Based on the considerations in the previous section, at least four themes for future research can be identified: comparisons of the new method with existing methods, ICC assumptions, the sampling of items for a new exam especially in the case of considerable cohort differences, and the multiplier to arrive from posterior median to a standard.

To start, we would need a series of studies that would allow for direct comparison between the proposed new method and existing methods that are commonly used in our field, including Angoff, Hofstee, Beuk, and (modified) Cohen. Choices of methods at medical schools may be driven by personal preference and financial/logistic means more than by empirical studies comparing methods we can choose from. Scholars whose opinion is that setting standards by definition involves panels like in Angoff, Hofstee or Beuk tend to not be in favor of (modified) Cohen or other methods that use empirical data rather than panels and may therefore not be in favor of the new method proposed in this article either. Series of well-designed studies involving direct comparisons between panel-based and data-based methods in different types of programs, involving different kinds of students and different types of exams, may help us to identify conditions under which some methods may be preferred over other methods. Until we accumulate that body of research, any preference for one (type of) method(s) over other (types of) methods may be entirely based on personal opinion and/or financial/logistic factors to be taken into account in a given institution or program, and decisions which methods to cover in handbooks on standard setting (e.g., to cover panel-based methods but not Cohen) may be a matter of sheer preference of the authors as well. Proponents of panel-based methods may argue that setting a standard always involves 'expert' judgment, but what evidence have we got really to be confident that in the absence of empirical data we are not just drawing a line in the sand but set a standard that is appropriate? To what extent do panel-based methods really result in standards that are different from those set with data-based methods like (modified) Cohen or the new method proposed in this article? Should one of the outcomes of future research be that panel-based methods and data-based methods result in clearly correlated standards with in individual exams mostly minor if not trivial differences, that might well raise the question why invest financial and logistic resources for panels when we can achieve about the same results with much less resource-intensive data-based methods?

Comparing calculations under ICC = 0, ICC = 0.1 and ICC = 0.2 for the example discussed in this article demonstrates that, even in a relatively small-sample situation, regardless of the ICC we arrive at virtually the same posterior median and (given a fixed multiplier, here 80%) fail/pass cut-off. With increasing sample/cohort sizes, the influence of different ICC assumptions on the posterior median and cut-off will fade further; it is mainly for situations where we have much less historical data than in the example discussed in this article and/or with for large exams unrealistically high ICC values that we may see a substantial influence of ICC assumptions on the posterior median and cut-off. Future studies could result in guidelines about desired minimum amounts of historical information and ranges of ICCs for which the newly proposed method works well. ICC values above 0.20 are not common in large exams on medical knowledge but may well occur in much smaller exams that focus on very specific content. Simultaneously, smaller exams tend to use less historical data than larger exams and the less data available the more the influence of different ICC assumptions on the outcomes. Therefore, exam size may constitute another important factor to consider in studies on the influence of ICC assumptions on the outcomes of the newly proposed method.

Exam size may also constitute an important factor in studies revolving around any potential impact of cohort differences on standards set and how item sampling methods can help to reduce that impact. Until further research indicates otherwise, one may expect that carefully sampling items can help to reduce impacts of cohort differences on standards. Besides, for items that have been used in all or several previous cohorts, effects of cohort differences on the outcomes may well be smaller than for items that have been used in a single cohort only. Simultaneously, at exam level — that is: the level at which the standard is set — effects of cohort differences could be minimised by striving for a careful balance of items used in different cohorts in any new exam.

Finally, although 80% provides an intuitive multiplier to arrive from a posterior median to a standard, future research can help to examine if 80% is indeed a good multiplier or perhaps if there are conditions under which different multipliers should be considered.

To conclude

The Bayesian method introduced in this manuscript uses information for standard setting that is not considered in methods already available and can be applied to both small and larger cohorts regardless of whether standard setting panels are available. As demonstrated in the item performance example, this new method is easy to implement and requires only a minimum of calculations which can be done in zero-cost, user-friendly Open Source software. It provides a pragmatic approach to standard setting even when limited performance data is available. The posterior distribution median multiplied by 0.8 (80%) provides an intuitive pass mark that can investigated further in future studies, and the 95% credible interval provides an indication of the degree of uncertainty around our posterior distribution median. Future studies could compare this method with existing methods on past and future exams to acquire a better understanding of how the cut-off scores acquired with this new method correlate with those of existing methods, what are possible effects of different ICC assumptions and cohort differences of different magnitudes on the outcomes, and what is the best multiplier to arrive from the posterior median to a cut-off score.

References

1. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, ed. *Educational Measurement*. Washington, DC: American Council of Education; 1971:508–600.
2. Burr SA, Zahra D, Cookson J, Salih VM, Gabe-Thomas E, Robinson IM. Angoff anchor statements: setting a flawed gold standard? *MedEdPublish*. 2017. https://doi.org/10.15694/mep.2017.000167. online.
3. Clauser BE, Harik P, Margolis MJ, et al. An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Appl Meas Educ*. 2008;22:1–21. https://doi.org/10.1080/08957340802558318.
4. Hsieh M. Comparing yes/no Angoff and bookmark standard setting methods in the context of English assessment. *Lang Assess Q*. 2013;10:331–350. https://doi.org/10.1080/15434303.2013.769550.
5. Jalili M, Hejri SM, Norcini JJ. Comparison of two methods of standard setting: the performance of the three-level Angoff method. *Med Educ*. 2011;45:1199–1208. https://doi.org/10.1111/j.1365-2923.2011.04073.x.
6. Shulruf B, Wilkinson T, Weller J, Jones P, Poole P. Insights into the Angoff method: results from a simulation study. *BMC Med Educ*. 2016;16:134. https://doi.org/10.1186/s12909-016-0656-7.
7. Beuk CH. A method for reaching a compromise between absolute and relative standards in examinations. *J Educ Meas*. 1984;21:147–152. https://doi.org/10.1111/j.1745-3984.1984.tb00226.x.
8. Norcini JJ. Setting standards on educational tests. *Med Educ*. 2003;37:464–469. https://doi.org/10.1046/j.1365-2923.2003.01495.x.
9. Cohen-Schotanus J, Van der Vleuten CPM. A standard setting method with the best performing students as point of reference: practical and affordable. *Med Teach*. 2010;32:154–160. https://doi.org/10.3109/01421590903196979.
10. Taylor CA. Development of a modified Cohen method of standard setting. *Med Teach*. 2011;33:e678–e682. https://doi.org/10.3109/0142159X.2011.611192.
11. Homer M, Fuller R, Hallam J, Pell G. Setting defensible standards in small cohort OSCEs: understanding better when borderline regression can 'work'. *Med Teach*. 2019. https://doi.org/10.1080/0142159X.2019.1681388. online.
12. Wagenmakers EJ, Marsman M, Jamil T, et al. Bayesian inference for psychology, Part I: theoretical advantages and practical ramifications. *Psychon Bull Rev*. 2018. https://doi.org/10.3758/s13423-017-1343-3. online.
13. Love J, Selker R, Marsman M, et al. JASP version 0.11.1.0. Retrieved from: https://jasp-stats.org/(Accessed: 14 January 2019).
14. Lindley D. *Bayesian statistics: a review*. London: SIAM; 1972.
15. Leppink J. *Statistical methods for experimental research in education and psychology*. Cham: Springer; 2019. https://doi.org/10.1007/978-3-030-21241-4.
16. Wagenmakers EJ, Love J, Marsman M, et al. Bayesian inference for psychology, Part II: example applications with JASP. *Psychon Bull Rev*. 2018. https://doi.org/10.3758/s13423-017-1323-7. online.
17. Ly A, Verhagen AJ, Wagenmakers EJ. Harold Jeffrey's default Bayes factor hypothesis tests: explanation, extension, and application in psychology. *J Math Psychol*. 2016;72:19–31. https://doi.org/10.1016/j.jmp.2015.06.004.
18. Rouder JN, Morey RD. Default Bayes factors for model selection in regression. *Multivariate Behav Res*. 2012;47(6):877–903.
19. Wetzels R, Wagenmakers EJ. A default Bayesian hypothesis test for correlations and partial correlations. *Psychon Bull Rev*. 2012;19(6):1057–1064. https://doi.org/10.3758/s13423-012-0295-x.