

Accuracy and confidence of objective structured clinical examination pass-fail decisions



Wei-Ming Hay
Year 5 Medical Student
University of Otago Wellington

This project was carried out as a summer studentship 2008-9.

Acknowledgements:

I would like to thank the supervisors Mike Tweed, Tim Wilkinson and Mark Fawcett-Thompson for their help, support and contribution to the planning and development of this project and the writing up of the report.

This project could not have taken place without the financial support received from Ako Aotearoa.



This work is published under the Creative Commons 3.0 New Zealand Attribution Non-commercial Share Alike Licence (BY-NC-SA). Under this licence you are free to copy, distribute, display and perform the work as well as to remix, tweak, and build upon this work noncommercially, as long as you credit the author/s and license your new creations under the identical terms.

Table of Contents

Summary.....	3
Introduction	5
Methods	6
Analysis.....	8
Results	9
Discussion.....	13
Tables	15
References.....	22

Summary

Introduction

Assessment of the consultation skills of undergraduates is important. Information may be acquired from different consultations, such as an OSCE, and decisions made on aggregated information. The effort to improve OSCEs, including blueprinting, scoresheet development and examiner training, is wasted if the decisions from these data are inappropriate. The aim of this study was to investigate the accuracy and confidence for decisions made by staff assessors given increasing information on students' performances.

Methods

Medical students at University of Otago sit a 10 station OSCE at the end of the 5th year of six. Staff assessors were shown authentic anonymised student scores for an increasing number of OSCE stations and asked to make pass-fail decisions and give a degree of confidence in this decision. The scores chosen were used to demonstrate good performance and also variable performance with different degrees of under-performance. The student results given included several parameters including examiner scores, passmark scores and running totals. Subsequently the staff assessors were given a fictional anecdote from a single observation made previously. This information was made to be deliberately discordant with the staff assessors' views to that point and again they gave decisions and confidence.

Following completion of forms, the staff assessors were interviewed regarding the rationale for their decisions.

Confidence was given on a 0-100% Likert scale. Accuracy was defined as the comparison with a gold standard, which was determined in two ways: the actual decision of the Board of Examiners (compensatory method on total scores) and also from consensus of decisions made by staff assessors during this study.

Results

35 staff assessors made 11 pass-fail and confidence decisions for a mean of 5.9 candidates each.

Across the 10 stations for the candidate who was above pass threshold for all stations the mean level of confidence in a pass increased from 80 to 90%. For the students that failed the most stations the level of confidence in fail varied between 70 and 80%. The confidence-accuracy difference, a measure of overconfidence, was greatest for students whose performance was closest to the pass-fail threshold.

Despite provision of progressively poor performances the staff assessors were not as confident in assigning fail. Internal and external factors affecting decision making process tend to contribute to their doubts. The anecdotal information changed 12% of the decisions to the extent that pass was altered to fail or vice versa.

Discussion

Making decisions on students whose performance is close to decision thresholds is difficult and associated with overconfidence. As a majority of students perform above threshold, staff assessors are more comfortable assigning a pass rather than fail. In the presence of uncertainty the staff assessors will tend to pass the student. This has implications for standard setting and decision-making processes.

To some assessors the plausible but unreliable anecdote is given an equal or greater importance than the evidence from the OSCE, and so leads to a decision change.

Introduction

Proficiency of consultation skills is essential for healthcare professionals [McKinley et al., 2001]. In order to make sure that undergraduate medical students have attained the required skills at the appropriate time, are progressing satisfactorily and to aid learning, there need to be assessment. This includes taking a history from a patient, performing appropriate clinical examination, and informing patients and carers/family about health and illness issues and discussing these issues. Therefore the assessment of the consultation skills of medical undergraduates is a necessity.

Given that both consultation skills and healthcare working environment are complex, consultation skills should be assessed by different methods in different context and at different times. Just one source of information cannot be relied upon. This requires aggregation from multiple sources of information to make a decision whether students have reached the required level of practice.

When making a decision, stakeholders, which include the faculty, staff, student, patients and the public, need to have confidence in the decisions. The decision from an assessment is based on format, content, scoring, and aggregation with decision processes. Faculty spends time assuring format, content and scoring with authentic contexts, blueprints, mark sheets and examiner training. Confidence is required not only in these but also the aggregation.

Direct observation of medical undergraduates undertaking a broad range of consultations with structured scoring schemes is frequently used in the assessment process, such as objective structured clinical examinations (OSCEs) [Harden & Gleason 1979]. A limited number of OSCE stations may reduce the statistical reliability of the scores [Newble 2004] and also the confidence of staff assessors in their own accuracy [Tweed & Ingham 2009]. Staff assessors recognise that judging the student's performance based on one consultation might not be representative of other consultations [Tweed & Ingham 2009], and so it will be difficult for assessors to be fully confident in their decision. Thus increasing the numbers of stations would improve statistical reliability and may increase staff accuracy or at least their confidence. Empirically, more information could lead to a better decision. However, in addition to good information, other information that is inconsequential to the final decision will be obtained [Bastardi & Shafir 1998]. Increasing information may only improve confidence of a decision [Schwartz 2004] but not its accuracy. Furthermore, confidence and accuracy may increase but may change at different rates and plateau at different points [Hall et al 2007].

Confidence in assessment of consultation skills of medical undergraduates that requires aggregation of multiple sources of information is important and has not been investigated.

In this study, the effects of increasing information available to faculty staff assessors from OSCE results on a number of students were investigated. The aim of the study is to address the following questions:

1. What is the effect of increasing the assessment results (number of OSCE stations) on the faculty staff assessor's confidence and accuracy?
2. What information is important to faculty staff assessors in making decisions?
3. Would anecdotal information affect the decision and/or confidence?

Methods

At the end of Year 5, of 6, medical students in University of Otago are required to sit a high stakes examination limiting progress to 6th year (Trainee Intern year). This comprises a written examination and an OSCE. In 2008 this OSCE consisted of 10 stations with two examiners marking independently using scores sheets of content and process. Scores were aggregated by compensation and passmark threshold derived by borderline regression [Kramer et al. 2003; Wood et al. 2006].

A database of students' results was developed for this study from the actual Year 5 OSCE results in 2008. They were chosen from a sample of student of various degrees of abilities across the cohort with a focus on more difficult decisions. From all the results obtained, the numerical information from each station was presented incrementally in a table form. The order of increments in station information was fixed, as per station order set out in the actual examination by the Faculty Board of Examinations (BoE) for Year 5.

The scoring patterns included features of:

1. Different number of stations failed
2. Difference in position of stations failed across all ten stations
3. Different extent of failure on individual stations
4. Inter-examiner concordance and variability
5. Passing all stations

On the results table, the following results were presented by station and running totals:

- Scores for content, process and total from each examiner;
- Passmark scale (to be used in borderline regression calculation) from each examiner;
- Inter-examiners' difference in the scores;
- Passmark threshold calculated from all students;
- The score below the passmark threshold.

In addition, the discipline, task and examinable content for each station were included. All data were anonymised.

The subjects for the study were staff who teach and/or assess consultation skills. During recruitment, staff across all campuses of University of Otago were invited to participate by email. After that, the time and place of interview were arranged to accommodate the convenience of the staff assessors. The interview was digitally voice-recorded and later transcribed. Before the interview started, an explanation and instruction of the interview procedure was given, and written consent gained.

The staff assessor was shown the table of increasing number of stations and after each station result, they would give a pass or fail decision and rate their confidence for each decision made on a scale of 0 to 100% [Adams 1957].

At the end of viewing a student's ten stations, the staff assessor was asked on the following:

1. What other information would increase your confidence in making a decision?
2. What other information would be needed to change your current decision?
3. Given that the Professor from the Department of (participant's department) had recently observed this candidate on an end-of-run OSCE station for (participant's department) and commented that this candidate had performed exceptionally good/bad, which is at registrar level/totally inept and inefficient. With this additional information, do you consider the student a pass or a fail? What level of confidence do you have in this decision?

This last additional description was discordant with the pass/fail decision given by the staff assessor after 10 stations.

Additionally after viewing their last student results in the time allocated, the staff assessors were asked about their overall opinions on the whole decision making process for OSCE.

Analysis

Confidence

Pass-fail decisions and confidence were converted to a pass-fail confidence continuum (Table 1).

Accuracy

The gold standard against which the individual participant's responses were judged was defined in 3 ways:

- the actual decision made by the BoE;
- the median judgement of all staff assessors participating having viewed 10 station results;
- the likelihood of the participating staff assessors awarding a pass grade having viewed 10 station results.

Analysis of confidence, accuracy and the confidence-accuracy relationship

Analysis of the confidence accuracy relationship was made in several ways. The confidence in the decisions was converted to a fail-pass confidence continuum and this compared with accuracy, defined as listed, using different criteria.

- Calibration curves correlations, point bi-serial correlations and gamma statistic [Krug 2007];
- Confidence minus accuracy [Benner 1996];
- Confidence-judgement accuracy quotient [Shaughnessy 1979].

Analysis of increasing information

The effect of increasing numbers of station results on pass-fail decisions and confidence was calculated.

Analysis of anecdote

The proportion of decisions changing, and the change in confidence in pass to fail and vice versa were calculated.

Analysis of interview

Analysis of transcripts of interviews was carried out after the numerical analysis. The main themes for study of transcripts were predetermined as comments on pass/fail decision, its confidence, and accuracy. Themes from within the interviews were used to verify or contradict the numerical analysis (Grbich 1999, Cohen 2000). Initially independently, the investigators reviewed the transcript for themes, then themes and exemplars were agreed by consensus.

This study was been approved according to University of Otago policy on ethical practices in research and teaching involving human participants.

Results

The results of 12 students were chosen for the database. Eight were awarded a pass by the BoE using compensatory method on total scores [Schuwirth 2008].

Of the 106 faculty staff assessors invited 35 agreed to participate in the project, a response rate of 33%. Each reviewed a mean of 5.9 candidate scores over 10 stations and the additional information.

Two of the staff assessors felt that they could only make decision based on all 10 stations, and so data was not available for stations 1-9 for these.

Confidence

Confidence scale response was calculated from the fail-pass continuum.

Accuracy

The 3 standards for measuring accuracy, the BoE result, the median of the pass-fail decisions by participants after 10 stations, and mean level of confidence for all participants given results of all 10 stations, were calculated

Confidence-accuracy relationships

The participating staff assessor's confidence is less than accuracy when analyses by the median judgment of pass-fail decisions (Figure 1) and BoE decision (Figure 2).

The correlation between accuracy (the likelihood of being deemed a pass after 10 stations) and confidence (mean level of confidence in the last decision as pass) was $r=0.96$ (Table 3).

There was overconfidence when the likelihood of awarding a pass was 0.5 (totally uncertain) after 10 stations and underconfidence for other points (Figure 4). The degree of underconfidence in decisions was greater for those failing rather than passing.

There was a quadratic correlation between confidence as likelihood of absolute decision and confidence-accuracy. For the uncertain decision-making confidence was greater than accuracy (Figure 5). As likelihood increases underconfidence developed and then levelled.

The confidence accuracy quotient was 4.5 across all decisions. This is a measure of the increasing confidence in correct as opposed to incorrect decisions. A CAQ of 0 means that a judge cannot discriminate when they are correct or not. The CAQ was the same whether the BoE decision was fail or pass and for both significantly greater than 0 (Figure 6).

When analyzed by likelihood of being awarded a pass, the CAQ was highest for the very poor performer and some of the better performers (figure 7). The variability for those judged likely to pass may be due to the variability within this group of results with occasional poor performance.

Gamma is a number of correct decisions made as a proportion of total. The gamma for all decisions was 0.69. This means that 69% of the decisions after each station were the same as the BoE decision. When analysed by likelihood of being awarded a pass, there was less variability in gamma for better performance (figure 8). The staff assessors agree on passing candidate more than for those of lesser performance.

What was the effect of increasing number of station results?

Despite every staff member, after every station, deeming student 10 a pass, the mean level of confidence in these decisions only rose from 80 to 90% across 10 station results (Figure 9). This implies that the staff assessors felt they are could be incorrect 1 in 10 with this pass decision even after 10 passing stations agreed by all assessors.

Only 48 of 207 evaluations for student 1 were deemed a pass with any degree of confidence. All the evaluations after station 10 were fail. Despite this the mean confidence after 10 stations was only 80% (Figure 10). The assessor expected this to be a false failing result 1 time in 5.

For those students deemed fail by the BoE the staff assessors did not reach any degree of confidence in this until after 8 stations (Figure 11).

What was the effect of anecdotal information?

There were 178 events with no change and 24 with a change. By defining one group as changers and comparing with non-changers, the alteration in confidence in those who did not change their decision was minimal. This implies that people either changed their decision or made no changes.

Interviews

The themes of the quantitative analysis were backed up by the comments of the staff.

There was underconfidence despite near certainty in fail and pass decisions. The main factors that contributed to this were inter-examiner variability; perceptions of difficulty of content and marking, familiarity with marking process and perceptions of subjectivity of marking.

“The mark given in the content must be consistent between both examiners. There should be no discrepancy in the mark given by both examiners. Discrepancy of 1 mark is fine but discrepancy of 2 or more marks is questionable.”

Assessor 18

“... to know the relative strengths of each examiner are in terms of whether their specific area of interest. ... in the area he specialises then he may have high expectation on that part.”

A21, Student 3

“...to know how this examiner examines station 2 (Med-Ed-Post MI). I would like to know how everyone examines station 7 (Surg-H-Dysphagia)... there were broad problems for this station that was too hard or imprecisely marked. That might increase my confidence to pass him and discount how badly he did. If these examiners were very tough on station 2 (Med-Ed-Post MI)

for everyone, I would like to discount the terrible mark for station 2.”

A24, Student 4

“...to know what the content issues were for some of the stations and what the discrimination between the examiners was about, particularly on station like number 10 where there was a difference in the opinion on the process...”

A25, S10

“...to see verbal communication between examiners. For example, they are at the ability to

“... I don't understand the format and not used to the format. The questions on how would previous things change your confidence in pass or fail need to know the previous exam on how many stations the previous OSCE had, also what portion the final mark contributed in this OSCE.”

A32

“There is a large subjective element. I think it is quite hard for the examiner, especially if there are many students. Lots of the marks are from process are really subjective and the passmark standard is quite subjective. Therefore, it could vary quite a lot between the examiners to a degree but overall it is pretty fair.”

A35

There was more underconfidence for failing than passing students. Factors included aggregating total scores with number of stations passed; ideas of passing scores that may be different to that calculated by borderline regression; perceived rigidity of marking, and the inter-examiner variability when the participating staff assessor wanted to see both examiners awarding failing scores.

“I don't think doing well on some stations makes up doing completely hopeless in others. If I am the one designing the OSCE, I'd say he could fail a minimum of three stations. I would want to know if there is any medical reason that he may have done poorly.”

A2, S11

“The way I observe this score operating is that if the total is under 50% then those students are failing. If these total score instead of being 14, 17, 19 and if they are down in the 10, 12, 11 then they are not fit for TI and you are going to fail them.”

A19, S10

“...one particular station where there is quite a discrepancy on what the examiners have scored. That might explain on the area of expertise of the examiners and that might defend for someone. If it is close fail, they have got to be very sure about that. If it is clear, I think that looking across the spectrum of number of different areas, they haven't achieved, not just the one bad result, but it is probably three stations he has failed.”

A21, S3

“If I found out that these examiners were marking station 10 (Psych-Ex-Cognition) in a very rigid way... the person is failing is because of the way the station was designed, not the student's performance and if station 7 (Surg-H-Dysphagia) is a difficult station and so everyone was doing really badly, that would make me thinking about bumping the person up to a pass because he is so close to the passmark running total.”

A24, S9

Comments were different between those changing decisions on basis of anecdotal evidence. There was different opinion in the reliability of the OSCE and prior results but may be related to concerns on compensation across disciplines.

“It will only influence me slightly, 5% the most. In the OSCE result, they are all very consistent and so I won't put in a lot of weighting having done, even if one OSCE station is brilliant and all the others are marginal. I won't decide based on this OSCE.”

A1, S12

“There is always a concern that people might not perform on the day and that might take that into consideration when people are close. I think it is looking on a number of different specialties and it is a bit difficult to translate one performance in OSCE. But if it is a general perception throughout the year that he has done well clinically then it should be taken into account as well. I think that people have to correlate to which one specialty he has done poorly in this occasion, either it is a medical one or O&G then it might alter my perception then.”

A21, S3

“Because I am only 50% happy with the fail mark, so I am not absolute confident and might give some discretion of that, particularly if it was the one he did for the one he failed in the particular exam, I might push him through. More inclined to give him a pass, if particularly last run he had was one of the one he did not do well.”

A29, S8

“I think they need to be separated. Because end of 5th year OSCE is different and I think it is not fair to bring through things from the year, unless there is something really particular, like being quite ill on the day and try to do the exam and did not perform very well on that day. Then you can take into account the performance during the year.”

A31, S11

“That might suggest this person got specific problem which affected his performance in the OSCE and also last end of run OSCE. If that is a consistent thing that he performed badly for several times, my confidence will be low that this person has reached the minimum standard of a TI. If it is that one end of run, then I start to wonder if he has some problems which also affect his end of run and OSCE.”

A34, S8

Discussion

The outcome of the BoE and median of the participating staff was similar in terms of pass-fail decisions. However the level of confidence of staff assessors in some of these decisions was not high. Confidence and accuracy were correlated. Overconfidence was noted for the most difficult decision, while underconfidence was apparent for easier decisions. Confidence only reached up to 90% even if the student candidate had passed all the stations. Confidence in the decision for the candidate recognised by all as a fail only reached 80%. Confidence and accuracy did increase with number of station results although the levels of correlation were not high. When anecdotal information was given, 12% of decisions were change.

The strength of this research was that authentic results were collected and a gold standard outcome, that of the BoE, was available. Faculty staff, who are involved in teaching and/or assessing consultation skills, were recruited, meaning a more realistic decision making process can be reflected.

There were some limitations to this study. Firstly, the response rate was relatively low, at 33%. Those that chose to respond might not be representative of the staff assessors as a group. The responders may be more confident, and overconfident in their decisions, or perhaps more reflective and underconfident. Secondly, some assessors found it hard to comprehend the whole decision making process and confidence process. Although they may be experienced in individual assessments or OSCE stations they may not have considered the aggregation of information and the decision-making processes. Thirdly, most assessors were not from the BoE and their decisions might not reflect the real-life decision. The BoE is the final arbiter of the whole examination and is made up of University staff with considerable experience in University policy and practice and assessment.

Building on the earlier study [Tweed & Ingham 2009], there were several strengths. As for the earlier study, with assessors observing recordings of a single consultation, the assessors were overconfident around the decision point and underconfident at the extremes. Similarly the assessors were more comfortable in giving a pass to students. Unlike the previous study the additional information derived from the 10 stations meant that accuracy and confidence were correlated. Also the range of scores for which overconfidence occurred was significantly reduced. Additional information did reduce but not exclude overconfidence.

The current study only focussed on the specific of group of candidates. It might not be representative to the overall cohort of the students. The pass rate for the OSCE is usually in excess of 95%. The assessors are not as used to seeing candidates with poor performance and may not be as familiar or comfortable. Despite this, ensuring that those who progress are fit to do so is an important role of the assessment process. There is no place for a 'sympathy pass'.

Assessors were over-confident when it comes to making difficult decisions. As a majority of students perform above threshold, staff assessors are more familiar with assigning pass rather than fail. In the presence of uncertainty, they will tend to pass the students. However, some assessors had expressed their concerns on the variation on the marks given between both examiners. This shows that assessors were concerned that the marks given were not a true reflection of the student's performance.

Additional information such as comments or dialogues of both examiners may help. Further analysis of the data such as multiple regression analysis may allow for evaluation of factors such as inter-examiner differences, station numbers passed, marks below pass thresholds and running totals. To some assessors, the plausible but unreliable anecdote given is an equal or greater importance than the evidence from the OSCE, and so leads to a decision change.

The number of OSCE stations is decided on a combination of practicality, predicted score reliability, and a blueprint of content. Although aiming for high levels of assessors' confidence might be desirable, it seems unlikely that levels above 90% would be achieved. There were concerns on various external factors that may contribute to apparent failure, such as cultural background, nervous status, and fairness of station and internal factors, derivation of passmark threshold, fairness of examiners and the flow of the stations.

Further questions raised relates to what factors assessors would take into account either consciously or subliminally and why staff are under-confident in awarding a pass and even more so for a fail decision. A reluctance to award a fail may influence the borderline group and borderline regression and similar standard setting processes and hence final outcomes.

Tables

Table 1: The fail-pass continuum of confidence

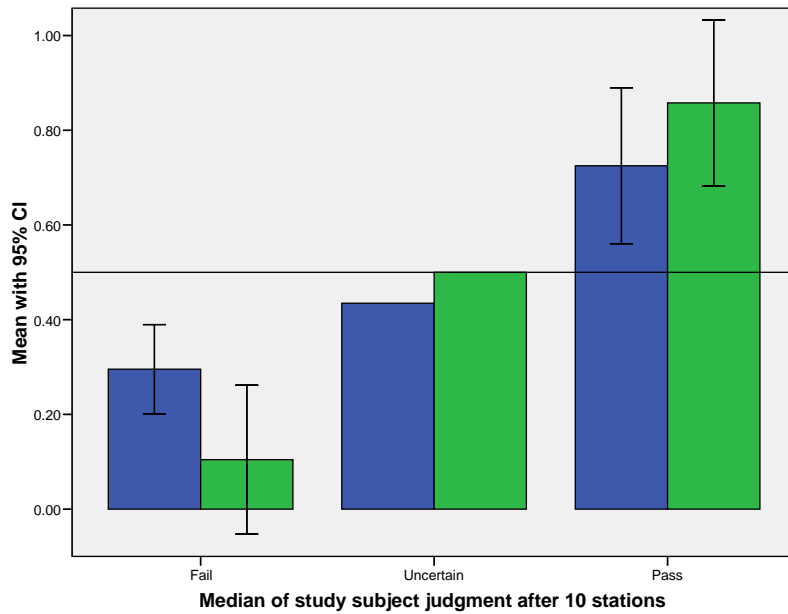
	Fail ↓					Uncertain ↓					Pass ↓
	100	90	80	70	60	50	60	70	80	90	100
Confidence in pass	0	10	20	30	40	50	60	70	80	90	100
Confidence in fail	100	90	80	70	60	50	40	30	20	10	0

Table 2: Different methods used to assess accuracy

Candidate	Likelihood of being deemed a pass after 10 stations	Median pass-fail decision after 10 stations	BoE
1	0.16	0	0
2	1.00	1	1
3	0.05	0	0
4	0.80	1	1
5	0.00	0	0
6	1.00	1	1
7	0.61	1	1
8	0.50	0.5	1
9	0.21	0	0
10	1.00	1	1
11	0.59	1	1
12	1.00	1	1

The median pass-fail decision by the staff assessors was the same as that for the BoE for 11 of the 12 students. For the twelfth, a pass was awarded by the BoE, the staff assessors' median was uncertain.

Figure 1: Mean likelihood and confidence in pass after 10 station by median participant decision



Blue = confidence (converted from % to 0-1 scale)
 Green = likelihood

Figure 2: Mean likelihood and confidence in pass after 10 station by BoE decision.

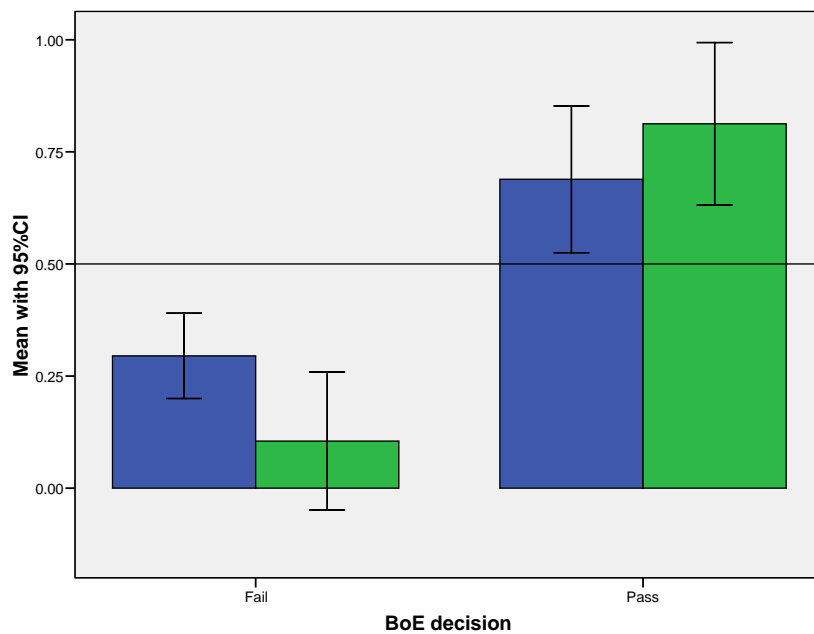


Table 3: The relationship between accuracy and confidence

Candidate	Likelihood of being deemed a pass after 10 stations	Mean level of confidence in the last decision	Mean level of confidence in the last decision as pass
1	0.16	72.50	27.50
2	1.00	75.00	75.00
3	0.05	75.68	24.32
4	0.80	65.00	65.00
5	0.00	71.67	28.33
6	1.00	91.82	91.82
7	0.61	52.64	52.64
8	0.50	56.53	43.47
9	0.21	61.96	38.04
10	1.00	92.08	92.08
11	0.59	51.67	48.33
12	1.00	82.68	82.68

Figure 4: The relationship between confidence and likelihood of passing after 10 stations

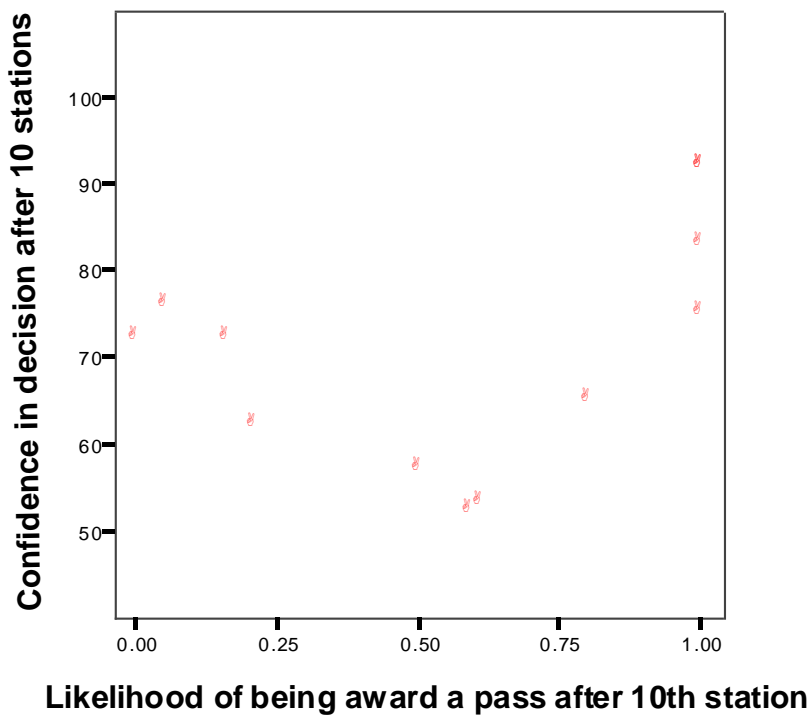


Figure 5: The Confidence-accuracy relationship using likelihood of an absolute decision being made

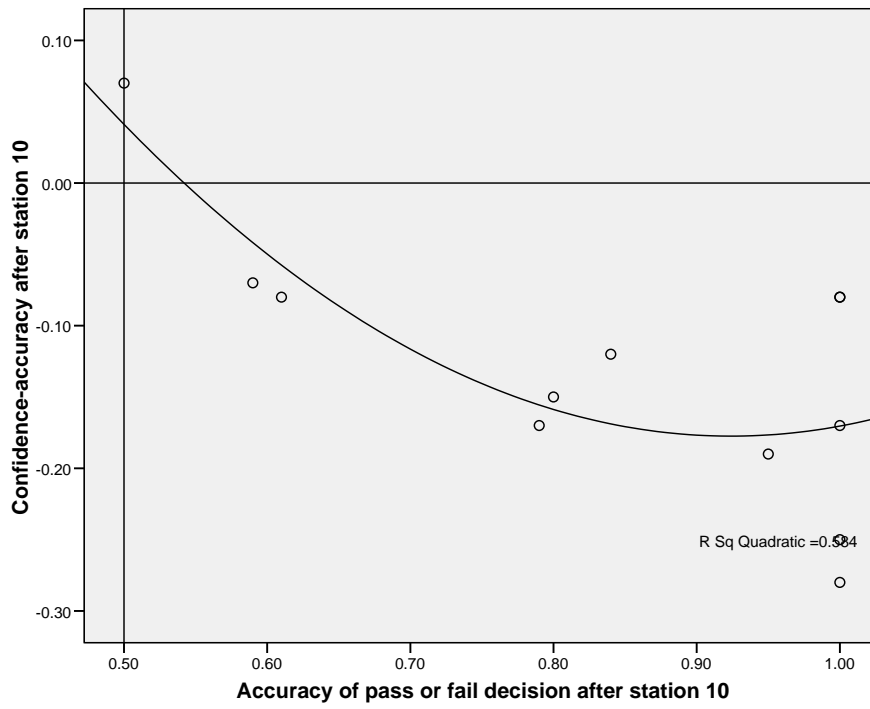


Figure 6: Mean CAQ by BoE decision

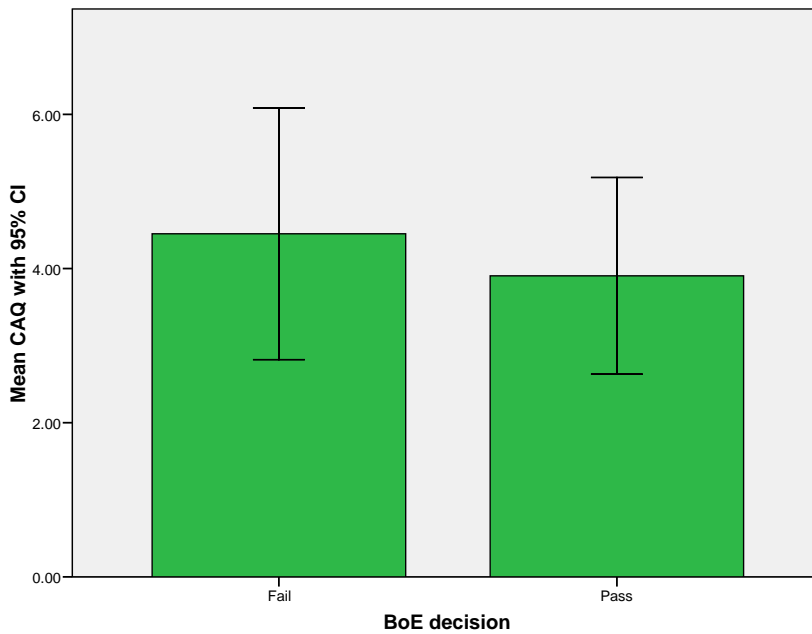


Figure 7: Mean CAQ by likelihood of student being awarded a pass

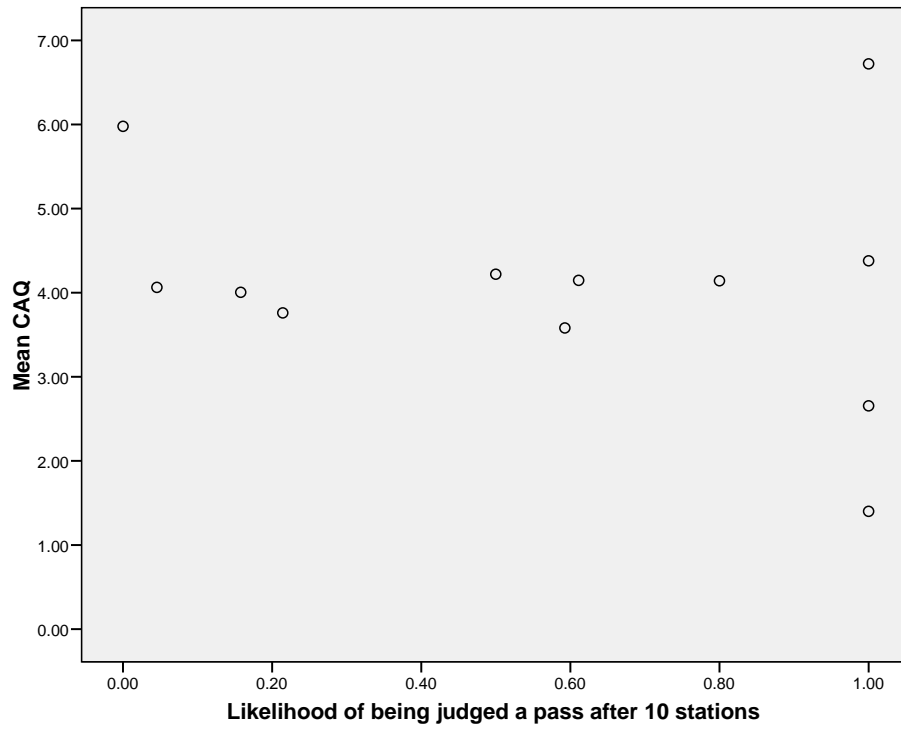


Figure 8: Gamma statistic by likelihood of student being awarded a pass

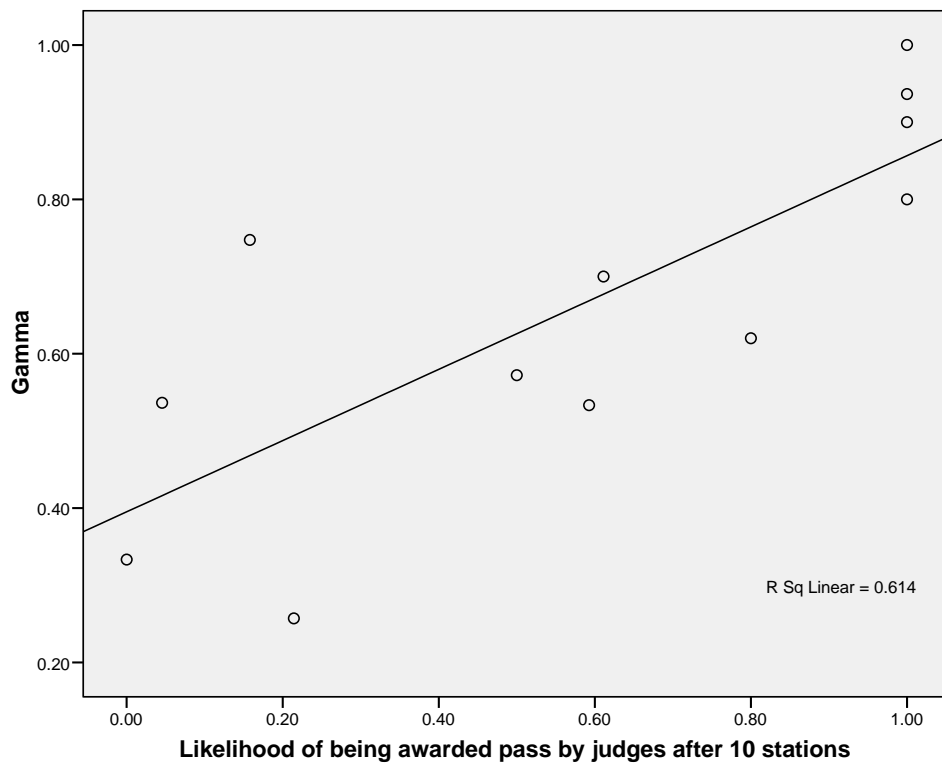


Figure 9: Change in mean confidence for student 10 (pass every station clearly)

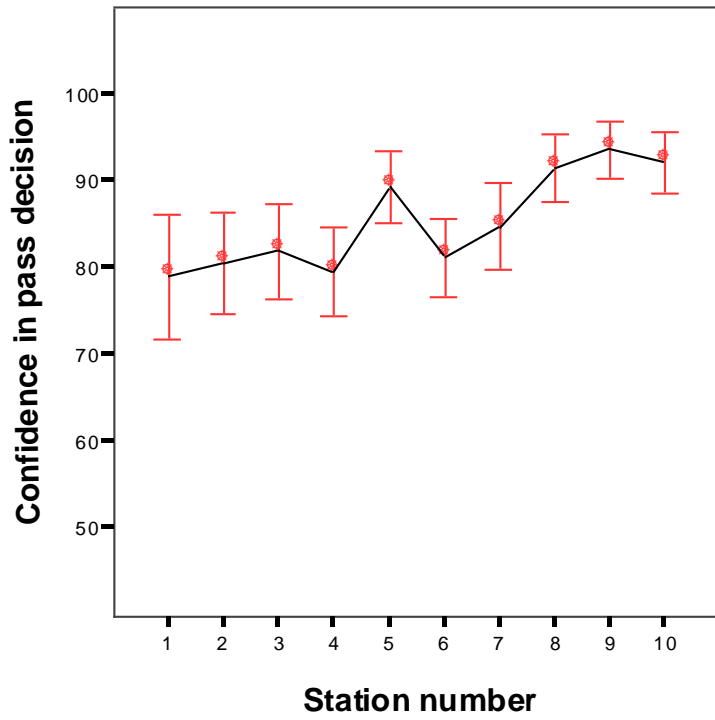


Figure 10: Change in mean confidence for student 1

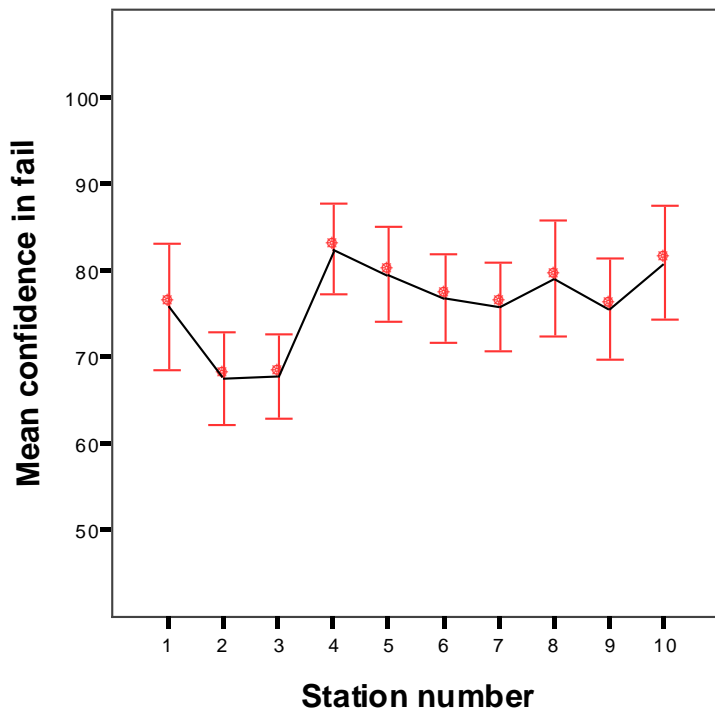


Figure 11: Mean confidence in a pass decision with increasing station results by BoE decision

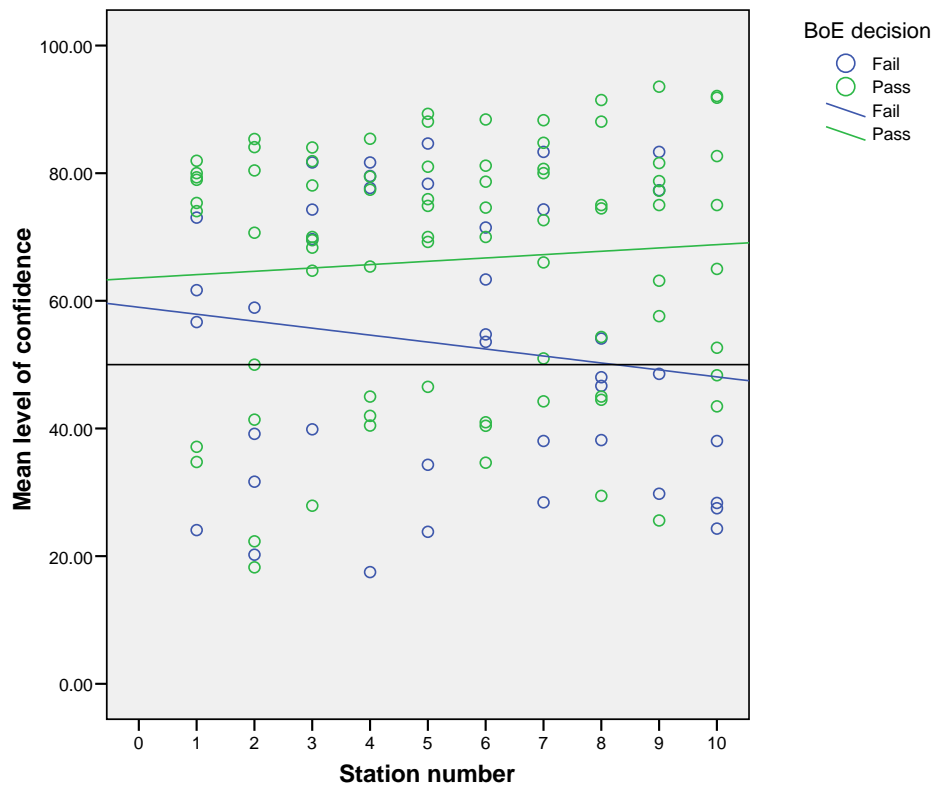


Table 4: The effect of anecdotal information of decision after 10 stations

		After information	
		Fail	Pass
Before information	Fail	70	12
	Pas s	12	108

Table 5: The degree of decisions changed after anecdotal information

	Distance from 50% before information	change	Distance from 50% after information
Change	17%	35%	18%
No change	32%	2%	29%
	p=0.2	p<0.001	p=0.4

References

- Adams, J. K. (1957). A Confidence Scale Defined in Terms of Expected Percentages. *American Journal of Psychology*, 70(3), 432-436
- Brenner LA, Koehlet DJ, Liberman V, Tversky A (1996). Overconfidence in probability and frequency judgements: A critical examination. *Organizational Behavioural and Human Decision Processes* 65(3):212-9
- Cohen, L., Manion, L., & Morrison, K. (2000). *Research Methods in Education* (London: Routledge)
- Grbich, C. (1999). *Qualitative research in health: an introduction* (St Leonards, N.S.W.: Allen & Unwin)
- Hall CC, Ariss L, Todorov A (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes* 103:277-90
- Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;13 (1):41–54.
- Kramer, A., Muijtjens, A., Jansen, K., Dusman, H., Tan, L., & Van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education*, 37(2), 132-139
- Krug, K. (2007). The relationship between confidence and accuracy: Current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice*, 3(1), 7-41
- McKinley, R. K., Fraser, R. C. & Baker, R. (2001). Model for directly assessing and improving clinical competence and performance in revalidation of clinicians. *British Medical Journal*, 322, 712-715
- Newble D (2004). Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education* 38: 199-203
- Schuwirth L WT. Assessing medical competence: finding the right answer The *Clinical Teacher* June 2004; 1(1): 14-18.
- Shaughnessy, J. J. (1979). Confidence-Judgment Accuracy as a Predictor of Test-Performance. *Journal of Research in Personality*, 13(4), 505-514
- Tweed MJ & Ingham C (2009). How do assessors make decisions on marking and standard setting for observed consultations? *Focus on Health Professional Education* (in press)
- Wood, T. M., Humphrey-Murto, S. & Norman, G. (2006). Standard setting in a small scale OSCE: A comparison of the modified borderline-group method and the borderline regression method. *Advances in Health Sciences Education*, 11(2), 115-122