Flinders
UNIVERSITY

ASME medical education www.mededuc.com

Very short answer questions: reliability, discrimination and acceptability

Journal:	*Medical Education*
Manuscript ID	MED-2017-0755.R1
Manuscript Type:	Research Papers
Keywords:	Testing/Assessment

Very short answer questions: reliability, discrimination and acceptability

Abstract

Background:

Single best answer questions (SBAs) have been widely used to test knowledge because they are easy to mark and demonstrate high reliability. However, SBAs are open to the criticism that they are subject to cueing. We used a novel assessment tool that facilitates efficient marking of open-ended very short answer questions (VSAs). We compared VSAs with SBAs with regards to reliability, discrimination and student performance, as well as and evaluating the acceptability of VSAs.

Method:

Medical students were randomised to sit a 60-question assessment in either VSA then SBA format (group 1, n=155) or the reverse (group 2, n=144). The VSAs were delivered on a tablet, which was computer-marked and subsequently reviewed by two examiners. Standard error of measurement (SEM) across the ability spectrum was estimated using item response theory.

Results:

Review of the machine-marked questions took on average one minute 36 seconds per question for all students. The VSAs had a high reliability (alpha: 0.91), significantly lower SEM than the SBAs (p<0.001) and higher mean item-total point biserial correlations (p<0.001). The VSA scores were significantly lower than the SBA scores (p< 0.001). The difference in scores between VSAs and SBAs was attenuated in group 2. Although 80.4% of students found the VSAs more difficult, 69.2% found them more authentic.

Conclusions:

VSAs demonstrated high reliability and discrimination and were perceived as more authentic. SBAs were associated with significant cueing. Our results suggest a higher degree of validity for VSAs.

Introduction

Multiple choice single best answer (SBA) questions are widely used in undergraduate and postgraduate medical assessment programmes worldwide, as they tend to have high levels of reliability and can be machine-marked efficiently and accurately. However, there have long been concerns that multiple choice questions may not provide a true reflection of knowledge as they rely on answer recognition rather than recall.[1, 2] Students tend to perform better in multiple choice questions than open-ended, free response short answer questions.[2-7] This may be due to the effect of cueing when candidates are presented with a list of options. A core principle of the validity of an assessment is the extent to which the test measures the competency it is supposed to measure.[8] Thus, if the aim of the assessment is to examine the student's ability to synthesise or generate rather than recognise a correct answer, short answer questions may provide a greater validity.[9]

It is well recognised that assessment drives learning,[10, 11] and that students prepare differently for exams in different formats.[10, 12-14] Indeed recognition and recall require different learning operations and their distinction has long been recognized in cognitive psychology.[15] Learning exam technique for multiple choice questions is a recognised phenomenon, which may lead to exam success at the expense of a deeper understanding of the subject being tested.[16-18] Furthermore, students demonstrate greater long-term information retention after studying for or completing short answer question rather than multiple choice questions questions.[19-22]

A meta-analysis investigating construct equivalence of multiple-choice and constructed response (e.g. short answer) items showed a higher mean correlation between the two formats when stem-equivalent items were used.[23] However, studies of stem-equivalent items include assessment of topics that lend themselves to multiple-choice questions. In our experience, question-writers can find it challenging to think of sufficient plausible distractors for some topics. The content of the assessment may also be skewed as core knowledge becomes too easy, causing question writers to resort to the testing of obscure material.[24] Short answer questions can offer greater flexibility for question-writers by allowing them to focus on common and relevant themes rather than academic minutiae.[24, 25]

Despite the potential advantages of short answer questions, their use in large-scale assessments has been limited by feasibility as, historically, they have not been amenable to being machine-marked[4, 26] and thus have been unable to sample the curriculum efficiently due to resource limitation. We have previously used Microsoft Excel to mark very short answer (VSA) questions[9]. In the current study, we used an online assessment management system, which allows questions to be posed in a VSA format on an electronic platform

requiring one-to-four word answers. Assessment item types can be categorised according to the degee of constraint on the respondent's options for answering ranging from fully constrained/selected responses (multiple choice questions) to fully constructed responses (essays).[27] Very short answer questions fall in the 'intermediate constraint' items which can be marked efficiently and accurately by computer using new information technologies.

The utility of an assessment method can be evaluated by its reliability, validity, acceptability, educational impact and costs.[28] The purpose of this study was to compare VSA and SBA questions to assess the utility of VSA questions as an assessment method. Although it is difficult to link assessment formats with learning behaviour directly,[13] we assessed student opinions and potential educational impact of VSA questions with a post-test student survey.

Methods

Participants and assessments

This study was approved by the Medical Education Ethics Committee at Imperial College London. Ethical approval was granted to invite all medical students in year 3 at Imperial College School of Medicine by an email from the faculty to sit a formative exam. All students had been on one surgical and two medical attachments in year 3. There were no other inclusion or exclusion criteria.

Medical students sat a formative examination consisting of 60 questions under exam conditions. The clinical vignettes were constructed to allow them to be posed in both SBA and VSA formats without changing the content (Figure 1). During the construction of the VSA questions we were able to generate a list of acceptable answers for each item within five minutes. Content validity was ensured by blueprinting against the year 3 curriculum at Imperial College School of Medicine to ensure broad sampling of relevant topics and close alignment with the syllabus. Items were written with short case descriptions and tested a range of cognitive processes, including clinical reasoning, decision-making and knowledge recall. They were independently reviewed to minimise construction errors.

Examinees were randomly assigned to two groups. In group 1, the examinees were presented with 60 questions posed in VSA format (a 90 minute exam), immediately followed by the same 60 questions posed in SBA format with five options (a 60 minute exam). In group 2, the first test consisted of SBA questions and the second of VSA questions. The students were given more time for the VSA questions because of the time taken to type the answers. Students had to complete the first test before commencing the second one, and could not return to the previous test. VSA questions were posed using a new online

examination management software (*Practique, Fry*) where the students provided answers on an iPad. The SBA format was delivered using a traditional paper-based system with a machine-marked scoring card (*Multiquest*, *Speedwell*). Following completion of both test formats, the students were invited to complete a feedback form to evaluate student opinions on the VSA questions.

Marking

Answers to the SBA questions were machine-marked with the individual student and question performance exported for statistical analysis. The students' answers to the VSA questions captured by the iPad App (*Practique, Fry*) were sent to a server over an encrypted connection. At the end of the exam the server applied an automated matching algorithm using the Levenshtein distance to match each answer against pre-approved acceptable answers for each question. Subsequently two markers reviewed all the non-exact matches and match failures to see if any of the non-exact matches should be disallowed, or any of the match failures should be allowed. Grouping similar answers in blocks facilitated this verification process. The system applied the examiner marking judgments to all the identical answers, ensuring consistency and saving the marker time. The system also learns the new marking judgments for each question and adds this to the pre-approved answer list for each question to improve the automatic marking for the next time that particular question is used.

Figure 2 shows an example of the marking system. The green answers have been automatically marked as correct based on the pre-approved answers. The yellow answers have been marked as correct based on their similarity to pre-approved answers. Answers marked as incorrect are shown in red, however during the review process this can be over-ridden, with all identical answers automatically given the same mark. The time taken for the examiners to review each item was recorded to give a measure of acceptability.

Analysis

Statistical analyses were performed with *IBM SPSS Statistics* version 24.0 and *Prism* version 5.0c software (GraphPad Software). The Kolmogorov-Smirnov test showed a normal distribution of all variables. Pearson's correlation coefficient was used to assess the correlation between the raw scores of the two formats. The difference in sex distribution between groups was tested using the chi-square test. Differences in raw scores, item-total correlations and standard error of measurement (SEM) within and between groups were analyzed using mixed design analysis of variance (ANOVA) with effect sizes expressed as partial eta squared (η_p^2). Differential item functioning (DIF) for sex was assessed with *Xcalibre* 4.2 (Assessment Systems) for all items using the significance of the z-test based on the Mantel-Haenszel coefficient.

Students' responses to the same questions in the two formats were compared, with 'positive cueing' defined as percentage of questions answered correctly in the SBA format and incorrectly in the VSA format. 'Negative cueing', occurred where distractors caused students who knew the correct answer in the VSA question to answer the SBA question incorrectly.[6]

Three-parameter logistic model analysis was carried out to include a specific parameter to estimate the probability of correct answers due to guessing. Analysis was carried out using the R package *mirt*.

In the post-test survey, students were asked to rate the following four statements on a 5-point Likert scale (strongly disagree, disagree, neutral, agree, strongly agree).

1. Questions in the single best answer format are easier than the very short answer format.
2. Very short answer questions are a better representation of how I would be expected to answer questions in clinical practice.
3. Having examinations in very short answer format would change my learning and revision strategy.
4. Using very short answer questions in assessments would help improve my preparation for clinical practice.

The questionnaire also provided a space for students to write comments about the use of VSA questions in medical school assessments. The *NVivo* software was used to identify themes in the students' free text responses.

Results

Of 340 students in year 3 at Imperial College School of Medicine, 302 students (89%) sat the formative examination, with 299 students (99%) completing both parts. The sex distribution was similar in both groups (female/male ratio in group 1: 72:83, and group 2: 63:81, p=0.64). Among the tests in both groups, only one item (a VSA question group 1) had significant differential item functioning (DIF) for sex, with bias against males (p = 0.04).

There was a significant positive correlation between the two formats (p<0.001, r=0.83). Due to the high reliability and the identical content of the tests, correction for attenuation was not performed, as this was likely to overestimate the correlation.

Acceptability

The VSA questions were reviewed by two examiners. Based on the preloaded acceptable answers, the system was able to identify 80.2% of correct answers prior to review, which was instrumental in allowing efficient marking. The remainder of correct answers were marked during the review process. Of answers marked 'correct' by the system, 0.2% were

deemed to be 'incorrect' on review (due to a spelling error significantly changing the meaning of the answer). The total time taken to review the machine-marked answers to all 60 VSA questions for all 299 students was 95 minutes 51 seconds. The average time spent by the examiners to review the answers to each question for all 299 students was 1 minute and 36 seconds (standard deviation=1 minute 2 seconds). The marking system allowed for multiple correct answers aside from trivial spelling or terminology differences, and in 8% (5/60) of questions at least one student offered an alternative answer to the question that was judged to be correct following review by examiners.

Effect of cueing

The raw scores for each test are shown in Table 1. The raw scores for VSA and SBA questions in group 1 were 52.4% and 68.2% respectively. In group 2, the raw scores for VSA and SBA questions were 65.7% and 69.7% respectively. There was a significant difference in raw scores between item type with a large effect size [$F(1,297)=384,339$, $p<0.001$, $\eta_p^2=0.56$]. There were also significant differences in the interaction between item type and group [$F(1,297)=136,343$, $p<0.001$, $\eta_p^2=0.32$, $p<0.001$) and between the groups [$F(1,297)=19,854$, $p<0.001$, $\eta_p^2=0.06$). The difference in VSA and SBA scores between the two groups is likely due to students seeing the answer options in the SBA questions before answering the VSA questions and the associated cueing effect. Positive cueing, with students answering the SBA question correctly and the equivalent VSA question incorrectly, was seen in 19.2% of items in group 1 and 7.5% of items in group 2. Negative cueing, with students answering the VSA question correctly and the equivalent SBA question incorrectly, was in seen in 3.5% of items in group 1 and 3.5% of items in group 2. On an item level, positive cueing occurred for every item (100%) and negative cueing occurred in 50 out of 60 (83.3%) items. There were four items where the students performed better in the VSA than the SBA questions, possibly because the answer options distracted the students. Notably in 20% of items, over 30% of students in group 1 could only get the answer correct in its SBA question format. For example, whilst 13% of students in group 1 were able to generate a diagnosis of erythema multiforme in the VSA question, 68% were able to select the correct answer in the SBA question.

Reliability and Discrimination

Table 1 shows reliability (Cronbach's alpha) and standard error of measurement (SEM) values for the VSA and SBA tests in both groups. SBA tests had Cronbach's alpha values of 0.84 and 0.85 in groups 1 and 2 respectively. VSA tests had a Cronbach's alpha of 0.91 in both groups 1 and 2.

We also used the three-parameter logistic model to estimate the standard error of measurement (SEM) for the two tests in both groups across the ability spectrum (Figure 3).

VSA questions had significantly lower standard error of measurement [$F(1,297)=213,782$, $p<0.001$, $\eta_p^2=0.42$]. The effects of the group and the interaction between item format and group were not significant ($p=0.23$ and 0.97 respectively).

Figure 4 shows individual estimates for information according to theta ability estimates and item type and group combination. VSA questions had significantly higher information with a large effect size [$F(1,297)=311,998$, $p<0.001$, $\eta_p^2=0.51$]. There was no significant interaction between item format and group [$F(1,297)=3,584$, $p=0.06$, $\eta_p^2=0.01$] and the group effect size was small [$F(1,297)=4,742$, $p=0.03$, $\eta_p^2=0.02$].

Mean item-total score point-biserial correlations for VSA questions were 0.36 and 0.35 in groups 1 and 2 respectively. Mean item-total score point-biserial correlations for SBA questions were 0.26 and 0.27 in groups 1 and 2 respectively. VSA questions had significantly higher item-total correlations [$F(1,118)=89,235$, $p<0.001$, $\eta_p^2=0.43$]. The effects of the group and the interaction between format and group were not significant (p=0.81 and 0.30 respectively).

Potential impact on learning behaviour

Figure 5 shows the percentage of students selecting each point on the Likert scale for each statement. 80.4% of students agreed/strongly agreed that SBA questions were easier than VSA questions. With regards to authenticity, 69.2% agreed/strongly agreed that VSA questions were more representative of how they would be expected to answer questions in clinical practice. Almost half the cohort (49.3%) agreed/strongly agreed that having VSA questions in summative examinations would change their revision and learning strategy and would help improve their preparation for clinical practice.

Thirty one percent of students who thought SBA questions were easier, commented this was due to the presence of options. Fifty six percent of free text comments regarding authenticity were related to how in practice they will be expected to recall information without options. Seventy percent of comments on a change in revision strategy indicated more emphasis on thoroughness and 9% would spend more time learning spelling.

Discussion

Our results indicate that VSA questions have advantages over SBA questions in terms of reliability and discrimination. The cueing effect associated with SBA questions was apparent in the analysis of the scores for the VSA questions and SBA questions in the two groups. The difference in scores between the two formats was attenuated when the students saw the

SBA questions first, likely due to the cueing effect of the options. Therefore, VSA questions have higher validity when testing ability to arrive at a correct answer without cueing or guessing.

Another potential limitation of SBA questions is the implication that there is one best answer for any question, therefore discouraging question writing in areas of medicine where there may be multiple defensible answers.[1, 3] Very short answer questions allow students to offer alternative answers that may be as good or second-option alternatives. Allowing students to demonstrate the scope of their knowledge improves the validity of the assessment.

Almost 70% of students thought VSA questions better represent how they would be expected to answer questions in real life clinical practice, and this suggests they have greater authenticity. More authentic examinations promote both deeper learning methods and increase student motivation.[29] Furthermore, a major component of learning from assessment is the quality of the feedback provided.[8, 11] Very short answer questions can offer an opportunity to provide more specific and detailed feedback based on the diverse range of incorrect answers proposed by the students. These can therefore be addressed or targeted by curriculum developers.

A major limitation to the widespread use of VSA questions would have been their acceptability in terms of resources. Introduction of any novel assessment method should be weighed against the extra time and resources incurred. Machine-marked VSA questions should be reviewed by subject matter experts. With the assessment software used in this study, the total time taken to review the machine-marked answers to all 60 VSA questions for 299 students was under two hours. Questions that took longer to mark had a higher number of permutations of the correct answers making it hard to create a comprehensive answer key (for example: ultrasound left leg, Doppler USS LL, Left leg US). However, every time an unforeseen answer is marked as 'correct' by the examiner, the system will learn the variation. Therefore, in future uses of the question, the answers will already be present in the system making marking more efficient. With advances in computational linguistics and machine-learning across educational fields,[30, 31] it is likely that speed and reliability of marking short answer questions will continue to improve.

Limitations of our study include the sample size and inclusion of students from one centre. Whilst the students were randomly assigned to the two groups and there were no differences in sex distribution, the possibility of differences in other characteristics cannot be excluded. For example we did not have access to data on the cultural background of the students for inclusion in the differential item functioning. Only one VSA question showed differential item functioning with a bias against males. Interestingly this item was based on the presentation

of urinary tract infection in a young female. Another limitation of our study is the lack of data on clinical experience and anticipated specialty of those who agreed VSA questions were more representative of real-life practice. Furthermore, there was no external validation measure to assess and compare students' competence. Future studies should investigate the utility of VSA questions in multi-centre studies involving a larger number of students. It would also be interesting to examine the effects of cueing in candidates with varying levels of expertise.

Given their high correlation with short answer questions, SBA questions are widely used as an efficient assessment tool across medical education as a proxy for assessing applied knowledge. However, high correlations between assessments do not necessarily imply that the same cognitive facility is being tested. Indeed answer generation rather than recognition is tested in short answer questions.[2, 4, 32] Very short answer questions have efficiency and acceptability that is approaching that of SBA questions. Furthermore, compared with the SBA format, VSA questions demonstrate higher reliability, discrimination and authenticity. The results of this study demonstrate the utility of VSA questions as a an assessment instrument, which has the potential to improve existing assessment programs.

Conflict of interests

The authors and their institutions do not have a financial interest in the software (*Practique, www.fry-it.com*) used in the study.

Acknowledgement

We thank Mr Keean Schupke for his help with the provision of the *Practique* software.

References

[1] Elstein AS. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. *Acad Med*. 1993;**68**:244-249.

[2] Veloski JJ, Rabinowitz HK, Robeson MR, Young PR. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. *Acad Med*. 1999;**74**:539-546.

[3] Shaibah HS, van der Vleuten CP. The validity of multiple choice practical examinations as an alternative to traditional free response examination formats in gross anatomy. *Anat Sci Educ*. 2013;**6**:149-156.

[4] Newble DI, Baxter A, Elmslie RG. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Med Educ*. 1979;**13**(4):263-268.

[5] Desjardins I, Touchie C, Pugh D, Wood TJ, Humphrey-Murto S. The impact of cueing on written examinations of clinical decision making: a case study. *Med Educ*. 2014;**48**(3): 255-261.

[6] Schuwirth LWT, van der Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. *Med Educ*. 1996;**30**(1):44-49.

[7] Schuwirth LWT, van der Vleuten CPM, Stoffers HEJH, Peperkamp AGW. Computerized long-menu questions as an alternative to open-ended questions in computerized assessment. *Med Educ*. 1996;**30**:50-55.
[8] van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;**39** (3):309-317.
[9] Sam AH, Hameed S, Harris J, Meeran K. Validity of very short answer versus single best answer questions for undergraduate assessment. *BMC Med Educ*. 2016;**16**:266.
[10] Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;**356**:387-396.
[11] Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001;**357**:945-949.
[12] Cilliers FJ, Schuwirth LW, van der Vleuten CP. A model of the pre-assessment learning effects of assessment is operational in an undergraduate clinical context. *BMC Med Educ*. 2012;**12**:9.
[13] Al-Kadri HM, Al-Moamary MS, Roberts C, Van der Vleuten CP. Exploring assessment factors contributing to students' study strategies: literature review. *Med Teach*. 2012;**34 Suppl 1**:S42-50.
[14] Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ*. 1983;**17**(3):165-171.
[15] Eagle M, Leiter E. Recall and Recognition in Intentional and Incidental Learning. *J Exp Psychol*. 1964;**68**:58-63.
[16] McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach*. 2004;**26**:709-712.
[17] Newble DI, Entwistle NJ. Learning styles and approaches: implications for medical education. *Med Educ*. 1986;**20**(3):162-175.
[18] Willing S, Ostapczuk M, Musch J. Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Adv Health Sci Educ Theory Pract*. 2015;**20**:247-263.
[19] McConnell MM, St-Onge C, Young ME. The benefits of testing for learning on later performance. *Adv Health Sci Educ Theory Pract*. 2015;**20**:305-320.
[20] Larsen DP, Butler AC, Roediger HL, III. Test-enhanced learning in medical education. *Med Educ*. 2008;**42**(10):959-966.
[21] Wood T. Assessment not only drives learning, it may also help learning. *Med Educ*. 2009;**43**(1):5-6.
[22] McDaniel MA, Roediger HL, 3rd, McDermott KB. Generalizing test-enhanced learning from the laboratory to the classroom. *Psychon Bull Rev*. 2007;**14**:200-206.
[23] Rodriguez MC. Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement*. 2003;**40**:163-184.
[24] Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Hum Pathol*. 1997;**28**:526-532.
[25] Damjanov I, Fenderson BA, Veloski JJ, Rubin E. Testing of medical students with open-ended, uncued questions. *Hum Pathol*. 1995;**26**:362-365.
[26] Case SM, Swanson DB. Extended-matching items: A practical alternative to free-response questions. *Teaching and Learning in Medicine*. 1993;**5**:107-115.
[27] Scalise K, Gifford B. Computer-Based Assessment in E-Learning: A Framework for Constructing “Intermediate Constraint” Questions and Tasks for Technology Platforms. *The Journal of Technology, Learning, and Assessment*. 2006;**4**:1-44.
[28] Van Der Vleuten CP. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996;**1**:41-67.
[29] Gulikers JTM, Bastiaens TJ, Kirschner PA. A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*. 2004;**52**:67-86.
[30] Burrows S, Gurevych I, Stein B. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*. 2015;**25**:60-117.

[31] Pulman SG, Sukkarieh JZ. Automatic short answer marking. *Proceedings of the second workshop on Building Educational Applications, Ann Arbour, Michigan*. 2005:9-16.
[32] Ozuru Y, Briner S, Kurby CA, McNamara DS. Comparing comprehension measured by multiple-choice and open-ended questions. *Can J Exp Psychol*. 2013;**67**:215-227.

Single best answer (SBA) question

A 24 year old woman has two months of lethargy, dizziness, weight loss and nausea. She has type 1 diabetes and reports erratic blood sugars and one episode of loss of consciousness. She has hyperpigmentation in her palmar creases and her oral mucosa. Her temperature is 36.8˚C, pulse rate 101 bpm, BP 78/61 mmHg (standing), respiratory rate 16 breaths per minute and oxygen saturation 99% breathing air. Her capillary blood glucose is 3.2 mmol/litre.

Investigations:

Sodium	129 mmol/L	(135–146)
Potassium	5.4 mmol/L	(3.4–5.0)
Urea	7.7 mmol/L	(2.5–7.8)
Creatinine	67 µmol/L	(50–95)

What is the most likely diagnosis?

A. Addison's disease
B. Congenital adrenal hyperplasia
C. Cushing's disease
D. Hypothyroidism
E. SIADH

Very short answer (VSA) question

A 24 year old woman has two months of lethargy, dizziness, weight loss and nausea. She has type 1 diabetes and reports erratic blood sugars and one episode of loss of consciousness. She has hyperpigmentation in her palmar creases and her oral mucosa. Her temperature is 36.8˚C, pulse rate 101 bpm, BP 78/61 mmHg (standing), respiratory rate 16 breaths per minute and oxygen saturation 99% breathing air. Her capillary blood glucose is 3.2 mmol/litre.

Investigations:

Sodium	129 mmol/L	(135–146)
Potassium	5.4 mmol/L	(3.4–5.0)
Urea	7.7 mmol/L	(2.5–7.8)
Creatinine	67 µmol/L	(50–95)

What is the most likely diagnosis?

Correct answers:
• Addison's disease
• Addison's
• Adrenal insufficiency
• Primary adrenal insufficiency
• Hypoadrenalism

Figure 1. Example of a question written in both single best answer (SBA) format with five options (left) and very short answer (VSA) format with the acceptable variations of the correct answer that would automatically receive a mark (right).

Figure 2. Example of marking of a very short answer question by computer. The green answers have been automatically assigned a mark (1.00) as they were included on the list of acceptable answers. The yellow answers have been marked as correct based on their similarity to the acceptable answers. Answers marked by computer as incorrect are shown in red, however during the review process this can be over-ridden (e.g. 'primary hypoadrenalism'), with all identical answers automatically receiving a mark.

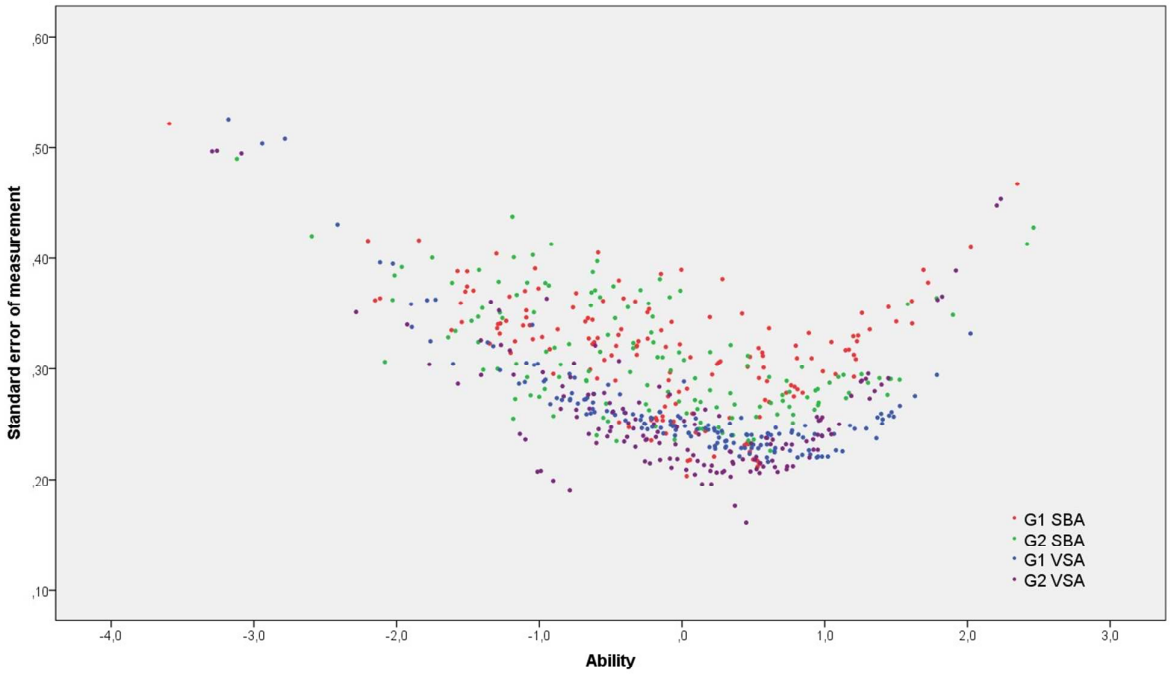

Figure 3. Individual estimates for the standard error of measurement according to theta ability estimates, and item type (single best answer: SBA, very short answer: VSA) and group (G) combination.

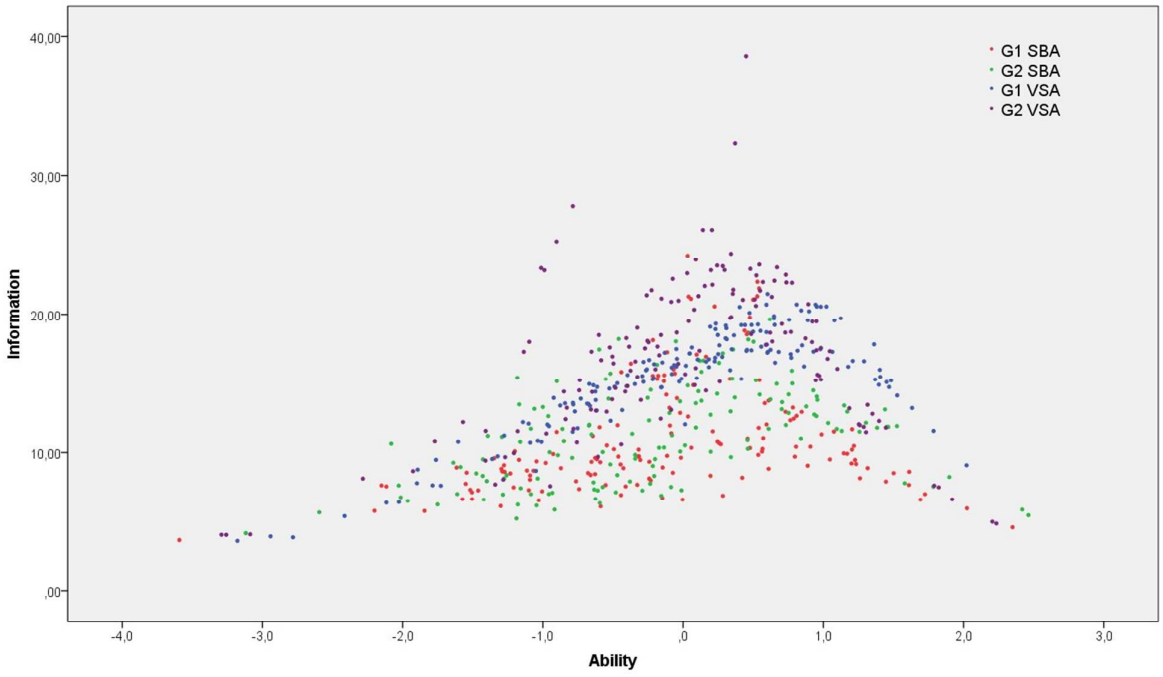

Figure 4. Individual estimates for information according to theta ability estimates, and item type (single best answer: SBA, very short answer: VSA) and group (G) combination.

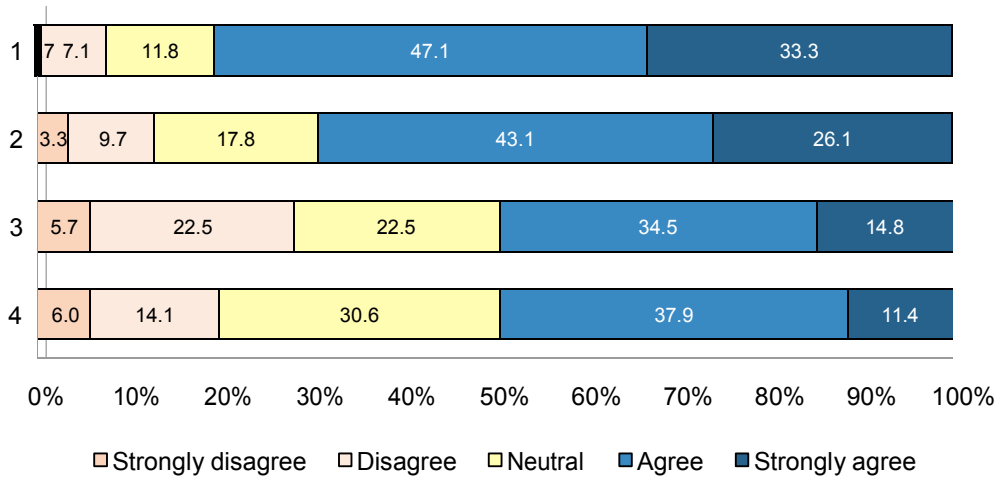

Figure 5. Students were asked to rate their agreement with each statement on a 5-point Likert scale (strongly disagree, disagree, neutral, agree, strongly agree). For each of the following four statements, the percentage of students selecting each point on the Likert scale is shown.

1. Questions in the single best answer format are easier than the very short answer format.
2. Very short answer questions are a better representation of how I would be expected to answer questions in clinical practice.
3. Having examinations in very short answer format would change my learning and revision strategy.
4. Using very short answer questions in assessments would help improve my preparation for clinical practice.

	Group 1 (n=155)		Group 2 (n=144)	
	VSA	SBA	SBA	VSA
Mean (%)	52.4	68.2	69.7	65.7
SD %	17.4	12.5	12.9	16.5
Cronbach's alpha	0.91	0.84	0.85	0.91
SEM	5.235	5.03	4.970	5.09
Mean item-total score point-biserial correlation	0.36	0.26	0.27	0.35

Table 1. The mean raw scores, standard deviation (SD), Cronbach's alpha, standard error of measurement (SEM) and mean item-total score point-biserial correlations for the very short answer (VSA) and single best answer (SBA) tests in groups 1 and 2.