The effect of two grading systems on the performance of medical students during oral examinations

Shakoor Ba-Ali[1, 2], Gregor B. E. Jemec[2, 4], Birgit Sander[1], Peter B. Toft[1, 2], Preben Homøe[2, 3, 5] & Henrik Lund-Andersen[1, 2]

ABSTRACT

INTRODUCTION: Either a pass/fail approach or a seven-point grading scale were used to evaluate students at the Danish universities. The aim of this study was to explore any effect of the assessment methods on student performances during oral exams.

METHODS: In a prospective study including 1,037 examinations in three medical subjects, we investigated the difference in the test scores between the spring- and autumn semester. In the spring semester, the students could either pass or fail the subject (pass/fail) while in the following autumn semester, the students were assessed by tiered grading (seven-point grading scale). Unknown to the students, the examiners assessed the students by the seven-point grading scale also in the spring semester. Students at the international classes who were officially assessed by the seven-point grading scale during both semesters served as control group.

RESULTS: The grading scores were significantly higher among students who were aware of being evaluated with the seven-point grading scores compared with the pass/fail group (p < 0.0001). In comparison, no significant difference between the exam results was observed from the spring- to the autumn semester for the control group (p = 0.45). Moreover, the average mark was higher among the international students (mean = 10.3, on the seven-point grading scale) than in the Danish speaking classes (mean = 9.1).

CONCLUSION: The seven-point grading scale seems to motivate students to yield a better performance; hence tiered-grading should probably be preferred to a simple pass/fail approach.

FUNDING: none.

TRIAL REGISTRATION: not relevant.

Since 1788, the Danish educational system has applied a total of ten different grading systems. Currently, either a seven-point grading scale or a simple pass/fail system is used to assess student performance in Danish university exams. The seven-point grading scale is a tiered grading system consisting of seven grades (−3, 00, 02, 4, 7, 10 and 12), while the pass/fail system is a simple dichotomous assessment (**Table 1**). The merits of the two systems are different and over the years many arguments in favour of one or the other have been expressed. Ad-vocates of the grades claim that a multi-tiered grading system is better able to identify students' professional strengths and weaknesses, while affording an optimized resource-allocation approach [1, 2]. On the other hand, the supporters of the pass/fail system draw attention to the positive contribution of this system to the psychological well-being of medical students in an already competitive environment [3]. The leading argument against the use of graded assessment has been that it may cause a grade-tethered mindset among students which does not inspire students to seek intellectual depth, while the adversaries of the dichotomous pass/fail system claim it lowers the drive for excellence in the students.

During the past 25 years, five different curricula have been implemented at the Faculty of Health and Medical Sciences, University of Copenhagen, Denmark. In three hereof, student performances in Ophthalmology, Dermatology and Oto-rhino-laryngology, have been evaluated by a grading method, whereas in the remaining two systems, the assessment was conducted by the pass/fail approach. In the 2003 study reform, the curriculum changed from a grading system to the pass/fail system without grades [4]. In 2009, the Study Committee at the Faculty of Health and Medical Sciences reverted the assessment method back to a grading method [5]. The transition from one assessment form to another gave us an exceptional opportunity to compare the pass/fail system with the seven-point grading scale.

Several studies have explored the impact of graded and pass/fail system on medical students' academic performances [1, 6]. However, to the best of our knowledge, no scientific evidence is available in favour of any of the two systems [2, 6].

In this prospective semi-blinded study, we aimed to compare the impact of the pass/fail system with the seven-point grading scale on the performances of medical students.

METHODS

Medical students, concluding the subjects of Ophthalmology, Dermatology and Oto-rhino-laryngology during the spring semester and autumn semester of 2012 at the University of Copenhagen were assessed. Officially,
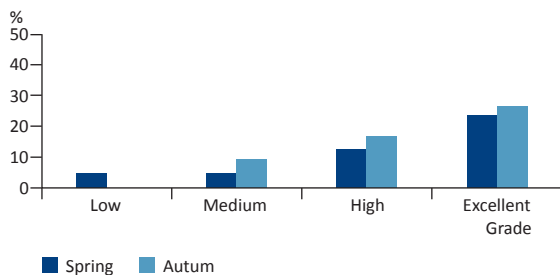
TABLE 1

The Danish seven-point grading scale consists of seven grades. The cut-off level is < 02, i.e. students with 00 and –3 do not pass the subject. For statistical purposes, we categorized the seven-point grading scale into four grade categories.

7-point grading scale	Explanation of the grade	Grade category
12	For an excellent performance	Excellent
10	For a very good performance	High
7	For a good performance	Medium
4	For a fair performance	Low
02	For an adequate performance	Low
00	For an inadequate performance	Low
–3	For an unacceptable performance	Low

FIGURE 1

Overall proportions of medical students by low, medium, high and excellent marks during spring and autumn, respectively, in Danish speaking classes.

the assessment method was pass/fail in the spring semester and graded (seven-point grading scale) in the autumn semester for the Danish speaking classes, whereas the seven-point grading scale was used to assess international students (English speaking classes) in both semesters owing to foreign university requirements. Approximately half of the students of the English speaking classes were Danish and the other half were from a host of European countries. Students from the English speaking classes therefore constituted the control group.

The study intervention consisted of adding a seven-point grading scale to the assessment, single-blinded to the examinees of the Danish speaking classes during the spring semester thereby providing data that allow for comparisons between the grades achieved using the two systems.

The exams in the three specialties were carried out in different manners. In Ophthalmology, the student undertook a clinical Ophthalmological investigation of a patient in the presence of the examiner and the assessor, which was followed by a discussion of possible diagnoses and treatments. Finally, the student was interviewed about related Ophthalmological topics. A total of 25 minutes was used. In Dermatology, the student was

asked to describe and discuss a photographic image of a Dermatological disease and a short patient history, upon completion of this, the history of a venerological case was presented and the student was asked to discuss this also. The time allotted for both questions was approximately 25 minutes. The exam in Oto-rhino-laryngology was a clinical exam mimicking a patient examination in an outpatient setting. The students were given 20 minutes to examine volunteering patients and thereafter a 20-minute exam was undertaken. In all specialties the assessors were internal lecturers from the University of Copenhagen who served as examiners and who had each been teaching a class, and also a number of external unbiased assessors, all of whom were experienced specialists. The outcome of the exam was a consensus based on to the two assessors' independent decisions. Data were anonymized before analysis, and institutional approval was not considered necessary.

Statistics

The grading was analysed as qualitative data with the chi-squared test. As only very few students scored a low grade, the grading was pooled into four classes: excellent (12 in the seven-point grading scale), high (10), me-

TABLE 2

Frequency distribution of Danish and English speaking classes by grades during the spring- and autumn semesters in Ophthalmology, Dermatology and Oto-rhino-laryngology. The values are %.

	Ophthalmology				Dermatology				Oto-rhino-laryngology			
	spring		autumn		spring		autumn		spring		autumn	
Grade	Danish	English	Danish	English	Danish	English	Danish	English	Danish	English	Danish	English
Low	5.88	0.0	6.41	0.0	28.16	16.7	10.06	0.0	15.85	6.3	6.74	0.0
Medium	29.41	12.5	10.90	8.3	30.46	16.7	15.72	18.2	24.59	0.0	15.17	18.2
High	36.36	37.5	33.97	33.3	30.46	20.8	33.34	18.2	38.25	31.3	33.15	18.2
Excellent	28.35	50.0	48.72	58.3	10.92	45.8	40.88	63.6	21.31	45.8	44.94	63.6
Total	100.00	100.0	100.00	100.0	100.00	100.0	100.00	100.0	100.00	100.0	100.00	100.0

FIGURE 2

The overall proportions of medical students according to marks (low, medium, high and excellent) during the spring- and autumn semesters in the control group, in English speaking classes.

dium (7) and low (< 7) (Table 1). Results were considered significant if the probability level was p ≤ 0.05.

Trial registration: not relevant.

RESULTS

A total of 1,037 examinations were performed; 544 in the spring semester and 493 in the following autumn semester. The total number of students who participated in the Ophthalmology exam in the spring semester was 187, and in the autumn semester the corresponding number was 156. The number of examinees in Dermatology and Oto-rhino-laryngology were approximately the same as for Ophthalmology. The mean grade during the spring semester, when students believed that they were being assessed by pass/fail only, was 8.4. The mean grade in the following autumn semester was 10.0.

During the spring semester, the proportions of students with low and excellent grades were 16.4% and 20.40%, respectively, while in the autumn semester, the percentages of students in the low and excellent categories were 7.7% and 44.8%, respectively (**Figure 1** and **Table 2**). The chi-squared test confirmed a positive association between better performances in terms of higher grades and the seven-point grading scale in the autumn semester when the students were aware of being evaluated by a graded scale (p < 0.0001).

The grading scores among the students in the English speaking classes, who constituted the control group and were assessed by the seven-point grading scale during both the spring (n = 48) and autumn semester (n = 56), did not differ between the two semesters (p = 0.45) (**Figure 2** and Table 2). In addition, the students of the English speaking classes scored higher grades (mean = 10.3) than the students of the Danish speaking classes (mean = 9.1).

Evaluating the medical subjects separately, the pro-

portion of students in Ophthalmology with an excellent grade (12 on the seven-point grading scale) increased significantly from 28.4% in the spring semester to 48.7% in the autumn semester (p < 0.0001) (Table 2). Similarly, in Dermatology and Oto-rhino-laryngology the proportions increased significantly from 10.9% to 40.9% (p < 0.0001), and from 21.3% to 44.9% (p < 0.0001), respectively.

DISCUSSION

This single-masked controlled study investigated the level of grading scores during two consecutive exams using a seven-point grading scale and a pass/fail method in Ophthalmology, Dermatology and Oto-rhino-laryngology at the University of Copenhagen. The objective was to evaluate whether assessment methods affected the presentation of medical students during oral exams. The most important finding was that test scores were significantly higher in all three specialties during the semester when the students were aware that they were assessed by the seven-point grading scale. No such differences were observed in the control groups counting international students who were officially assessed by the seven-point grading scale during both the spring and autumn semester.

The interaction between assessment and learning has been the subject of research for many years and the importance of assessments in students' education and learning has been affirmed in a wide range of studies [7-9]. In educational science, two main types of assessments, formative and summative, are used to describe the assessment of teaching and learning [10, 11]. While formative assessment aims to improve learning and teaching through consecutive evaluation and feedback, the goal of summative assessment is to measure the educational outcomes at the end of a learning activity [11]. Although summative assessment and tests without feedback are criticized for being less beneficial than formative assessments, testing is known to improve learning through a phenomenon coined the *testing effect* [12]. Both pass/fail and tiered grading are summative assessments, and since almost every topic is repeated several times in the course of medical training, the tests may bring some form of feedback.

The pass/fail as an assessment method is broadly used at Danish universities. However, its application at the universities varies significantly: at Aalborg University, for example, 15 out of 42 exams in Medicine are carried out by the pass/fail method [13, 14] whereas the University of Copenhagen applies the pass/fail approach in only four out of 35 exams [15, 16]. Knowing that tiered grading is more sensitive to elucidate students' professional strengths and weaknesses, it would be interesting to know whether the later clinical competences of doctors from the two medical universities are dif-

ferent. One study investigated the predictive validity of multi-tiered grading during the medical school on the post-graduate clinical competences as doctors [1]. Interestingly, the study showed that grades achieved during medical school are predictive of clinical performance. However, the comparison of competences at the post-graduate level is very challenging, and in the existing literature different studies have used different outcome measures. The common academic outcomes in the extant studies are the United States Medical Licensing Examination (USMLE) scores, student's ability to attend top residency programmes and evaluation by residency program directors. These studies, however, do not clearly show evidence in favour of either a pass/fail or a multi-tiered grading scale [6, 17-20].

The effect of the assessment methods on the academic performances during preclinical years has been investigated in several studies, and while there is strong evidence that the pass/fail method contributes to the psychological well-being of students, no clear evidence has been established concerning the academic impact of the two methods [1, 3, 6, 17].

In the present study, we found that tiered grading produced better performances than pass/fail assessment. It has been suggested that greater student-awareness during tiered grading plays a distinctive and explanatory role with respect to performance. Alternatively, systematic differences in the students' intellectual prowess may also be hypothesized, although such a systematic difference is unlikely due to the sample size and the stringent uptake criteria required for admission to medical studies.

The main qualifications of the current study are the large number of observations, the use of a control group and the single-blinded design of the study. The clear instructions to the examiners to follow the official requirements in the assessments of student-performances also contributed to reduce bias. Correspondingly, the most important weaknesses were the observer-expectancy effect, as both the internal examiners and the external assessors were aware of the study when grading students in the spring semester. Nevertheless, the intervening six-month break between the spring and autumn exams and the random allocation of external examiners limits the chance of conscious based bias i.e. remembering the grades given in the past spring during the autumn exam. However, observer bias was difficult to eliminate in the current situation. One approach to reducing the observer-expectancy effect was video recording of the exams and subsequent assessment by external assessors blinded to the goal of study. Another factor that would be interesting to explore is the perceived stress of the students in the two cohorts. It would, however, be more appropriate to investigate this factor in a repeated measures design where the students could serve as their own controls.

CONCLUSION

Students who were aware that their oral examinations resulted in a graded score achieved higher scores than students who thought they were assessed by a simple dichotomous pass/fail system only. It may be speculated that the assessment method resulting in specific test scores created an incentive for students to make a stronger effort to gain higher scores [12].

LITERATURE

1. Gonnella JS, Erdmann JB, Hojat M. An empirical study of the predictive validity of number grades in medical school using 3 decades of longitudinal data: implications for a grading system. Med Educ 2004;38:425-34.
2. Kreiter CD, Ferguson KJ. An investigation of the generalizability of medical school grades. Teach Learn Med 2016;28:279-85.
3. Reed DA, Shanafelt TD, Satele DW et al. Relationship of pass/fail grading and curriculum structure with well-being among preclinical medical students: a multi-institutional study. Acad Med 2011;86:1367-73.
4. 2000 studieordning for bacheloruddannelsen og 2003-kandidatstudieordning for kandidatuddannelsen i medicin, Det Sundhedsvidenskabelige Fakultet, Københavns Universitet, Version pr. 01.09.2011. http://sund.ku.dk/uddannelse/vejledning-information/ studieordninger/medicin/tidligere-kandidatstudieordninger/ medicin-ka-2000-2003-11.09.01.pdf (28 Sep 2016).
5. 2009-kandidatstudieordning, kandidatuddannelsen i medicin, Det Sundhedsvidenskabelige Fakultat, Københavns Universitet, Version efterårssemestret 2011. http://sund.ku.dk/uddannelse/vejledning-information/studieordninger/medicin/tidligere-kandidatstudieordninger/ medicin-ka-2009-11.09.01.pdf (30 Sep 2016).
6. Bloodgood RA, Short JG, Jackson JM et al. A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. Acad Med 2009;84:655-62.
7. Brown S, Knight P. Assessing learners in higher education. London: Kogan Page, 1994.
8. Ramsden P. Learning to teach in higher education, 2nd ed. London: RoutledgeFalmer, 2003.
9. Heywood J. Assessment in higher education. London: Wiley, 1977.
10. Huhta A. Diagnostic and formative assessment. United Kingdom: Wiley-Blackwell, 2010.
11. Scriven M. The methodology of evaluation. Monograph series on evaluation, no. 1. American Educational Research Association, 1967.
12. Roediger HL, 3rd, Karpicke JD. The power of testing memory: basic research and implications for educational practice. Perspect Psychol Sci 2006;1:181-210.
13. Studieordning for bacheloruddannelsen i medicin, 1.-6. semester, 2010, Aalborg Universitet, opdateret november 2014, version 4. www.smh.aau. dk/digitalAssets/96/96467_bsc-medicin-02.12.2014.pdf (30 Sep 2016).
14. Studieordning for kandidatuddannelsen i medicin, 1.-6. semester, Aalborg Universitet, 2013. www.smh.aau.dk/digitalAssets/96/96469_ studieordning-for-kandidatuddannelsen-i-medicin_-02.12.2014.pdf (30 Sep 2016).
15. 2012-studieordning for bacheloruddannelsen i medicin ved Det Sundhedsvidenskabelige Fakultet ved Københavns Universitet, Skolen for Human Sundhed og Medicin, version 2016. 2012.
16. 2015-studieordning for kandidatuddannelsen i medicin ved Det Sundhedsvidenskabelige Fakultet på Københavns Universitet, School of Medical Sciences, version marts 2015. http://sund.ku.dk/uddannelse/vejledning-information/studieordninger/medicin/Medicin-ka-2015_pr._01-09-2016_ pr._09.09_indsat_nye_aktivitetskrav.pdf (30 Nov 2016).
17. White CB, Fantone JC. Pass-fail grading: laying the foundation for self-regulated learning. Adv Health Sci Educ Theory Pract 2010;15:469-77.
18. Dietrick JA, Weaver MT, Merrick HW. Pass/fail grading: a disadvantage for students applying for residency. Am J Surg 1991;162:63-6.

19. Provan JL, Cuttress L. Preferences of program directors for evaluation of candidates for postgraduate training. CMAJ 1995;153:919-23.
20. Hughes RL, Golmon ME, Patterson R. The grading system as a factor in the selection of residents. J Med Educ 1983;58:479-81.