Effects of a rater training on rating accuracy in a physical examination skills assessment

Abstract

Background: The accuracy and reproducibility of medical skills assessment is generally low. Rater training has little or no effect. Our knowledge in this field, however, relies on studies involving video ratings of overall clinical performances. We hypothesised that a rater training focussing on the frame of reference could improve accuracy in grading the curricular assessment of a highly standardised physical head-to-toe examination.

Methods: Twenty-one raters assessed the performance of 242 third-year medical students. Eleven raters had been randomly assigned to undergo a brief frame-of-reference training a few days before the assessment. 218 encounters were successfully recorded on video and re-assessed independently by three additional observers. Accuracy was defined as the concordance between the raters' grade and the median of the observers' grade. After the assessment, both students and raters filled in a questionnaire about their views on the assessment.

Results: Rater training did not have a measurable influence on accuracy. However, trained raters rated significantly more stringently than untrained raters, and their overall stringency was closer to the stringency of the observers. The questionnaire indicated a higher awareness of the halo effect in the trained raters group. Although the self-assessment of the students mirrored the assessment of the raters in both groups, the students assessed by trained raters felt more discontent with their grade.

Conclusions: While training had some marginal effects, it failed to have an impact on the individual accuracy. These results in real-life encounters are consistent with previous studies on rater training using video assessments of clinical performances. The high degree of standardisation in this study was not suitable to harmonize the trained raters' grading. The data support the notion that the process of appraising medical performance is highly individual. A frame-of-reference training as applied does not effectively adjust the physicians' judgement on medical students in real-live assessments.

Keywords: rater training, rating accuracy, skills assessment, physical examination skills, randomised controlled trial

Gunther Weitz[1]
Christian Vinzentius[2]
Christoph Twesten[1]
Hendrik Lehnert[1]
Hendrik Bonnemeier[3]
Inke R. König[4]

1 Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Medizinische Klinik I, Lübeck, Deutschland

2 Institut für Qualitätsentwicklung an Schulen Schleswig-Holstein, Kronshagen, Deutschland

3 Universitätsklinikum Schlesweig-Holstein, Campus Kiel, Medizinische Klinik III, Kiel, Deutschland

4 Universität zu Lübeck, Institut für Medizinische Biometrie und Statistik, Lübeck, Deutschland

Introduction

The physical examination is a core clinical competence for every physician. A major task in medical education is to impart profound physical examination skills. However, recent literature raises concerns over declining abilities of graduates to perform a thorough physical examination [1], [2]. Factors contributing to this development include a scarcity of good teaching patients, skilled faculty, and time for bedside teaching [3], [4]. Also, increasing specialisation has led to an over-reliance on technology and a loss of the big picture [4], [5]. Hence, teaching and accurately assessing basic examination skills may more and more become a challenge in medical education.

Over the last decades several strategies have been established to secure the quality of physical examination skills training. These include the introduction of standardised patients (SPs) and patient instructors [6], [7], the application of checklists and rating forms [8], the implementation of Objective Structured Clinical Examinations (OSCEs) [9], and systematic direct observations of patient encounters [10]. In practice, however, these tools do not always take effect as intended. E.g., in a study from Taiwan, 22% of final year students reported to have never been observed in a physical examination (36% never by faculty) and 10% felt not yet confident with the procedure [11]. In our faculty, we evaluate a standardized head-to-toe examination of every third year student immediately after a tutorial over the first five weeks of the semester. How-

ever, we frequently receive complaints concerning the fairness of these assessments. Reliability and accuracy of faculty evaluation is indeed known to be low [12]. Structuring the evaluation by a rating form markedly increases the accuracy of the observations, but does not improve the agreement in the overall assessment [13]. This may be due to the fact that the raters' strategies to integrate information are rather individual and that the frame of reference differs between the raters [14], [15]. Studies from personnel psychology indicate that frame-of-reference training in groups can improve the accuracy in performance appraisal [16], [17]. The goal of such training is to teach raters to share a common conceptualisation of performance. It thereby imposes more accurate schemes [18]. We therefore planned to implement a rater training for the assessment of the physical examination skills.

Surprisingly, studies on rater training in medical education are scarce and the results are somewhat disappointing. In a small study, Newble and co-workers investigated the impact of training on the ratings of five videotaped physical examinations [19]. They gave either no training, performance feedback to the raters in one group, or feedback with additional training, including a discussion of another videotaped encounter, in a third group. There were no notable differences in the re-ratings of the videotapes after two months in either group. Holmboe and co-workers studied the effects of an intensive multi-dimensional rater training on the ratings of videotaped patient encounters eight months after training [20]. The trained faculty was more stringent and had a smaller range in some of the ratings. More recently, Cook and co-workers investigated the effects of similar but shorter training on interrater reliability and accuracy of mini-CEX ratings in a resident program [21]. The training did not improve these parameters.

To test whether a rater training would improve accuracy in our setting we undertook this study. Our setting differs from previous studies in several aspects:

Firstly, our rating focussed on a defined skill rather than assessing overall performance. Secondly, we standardized the physical examination task and all raters were familiar with the faculty standard. And thirdly, while previous studies relied on videotaped and scripted situations to determine the quality of rating, we had the chance to study real-live encounters between examinees and standardized patients (SPs). This implies that the grading was relevant and that the raters had to announce their decisions face to face to the examinees. For evaluation we videotaped all the exams and let three observers independently grade the students' performance retrospectively. We hypothesised that trained raters would rate more in line with the post-hoc observers, hence, being more accurate. We also sought to assess the effects of the rater training on stringency and the range of the grades.

Methods

Curricular embedment

The physical examination skills assessment was part of a course in physical examination to medical students at the beginning of their third year. The goal of this part of the course is to teach the students the basics of the physical examination in general internal medicine. After training the course continues with bedside teaching. The procedure is standardised to a head-to-toe screening physical examination and includes the inspection of head and mouth, the inspection and palpation of the neck, the complete examination of thorax and abdomen, an orientating examination of the vascular system (including measurement of one blood pressure), and the inspection of the limbs. A video explaining the standardised procedure is accessible to all students on the web. Other elements of the physical examination such as the pelvic, the musculoskeletal, and the neurological examination are taught in other parts of the course.

Training takes place in the first five weeks of the winter semester. It consists of five ninety-minute lectures and the same amount of training with peer examinations in groups of six students instructed by one experienced internist each. The assessment of the students' skills is scheduled in the sixth week. The students' task is to present the standardised examination with a standardized patient (SP) in a time limit of ten minutes. The raters are physicians selected from the medical departments of the University Hospital. They watch the students' performance, give feedback, and rate the performance by assigning a grade (German school grading, see table 1). They do not interfere with the students' examination nor do they ask theoretical questions. Each rater assesses six students in a time frame of fifteen minutes per student on two days each. The SPs are healthy students. They are instructed to behave passively and only to comply with coherent commands.

Table 1: German grading code (the raters were permitted to alter the grades 1 to 4 by +/-)

1 = excellent
2 = good
3 = satisfactory
4 = merely passed
5 = failed

For this study, twenty-one physicians were chosen to rate the performance of 242 students. All twenty-one raters were familiar with the learning objectives of the course, the free accessible video of the standard procedure on the web, and the feedback code. Eleven out of these twenty-one individuals were randomly chosen to undergo the rater training. For the randomisation, the raters were numbered and then assigned to the groups by numbers derived from a website creating random numbers in a given range. To determine the accuracy of the grading all the examinations were videotaped for further evaluation. Both raters and students gave written informed consent before the study. The study was approved by means of the local ethics committee. The work was carried out in accordance with the Declaration of Helsinki, and the anonymity of all participants was guaranteed.

Intervention

The eleven raters chosen for the training were split into two groups (six and five persons per group, respectively) to achieve a smaller group size. The training was scheduled in the end of the week before the skills assessment (Thursday and Friday afternoon, respectively, skills assessment on Monday and Tuesday afternoon). Training was limited to ninety minutes. In a short introduction, the moderator (author GW) stated the goals and the standards of the assessment as well as the rating dimensions (see table 2. The raters were then shown four videos showing different fourth-year students performing the standardised examination with a standardised SP at different levels of competence. The videos were presented in the same order in both training groups. After each presentation the raters were asked to assess the performance using a checklist with seven dimensions (see table 2) and to write down their grades for each item (see table 1). The raters then read out all their grades and for each item the raters with the most different grades were asked to justify their judgement. The ensuing discussions of all participants were chaired by the moderator. After all dimensions had been discussed, the moderator gave feedback featuring the embedded faults of each video.

Examination skills assessment

Because the untrained raters were not familiar with checklist forms both the trained and the untrained raters were asked to assign an overall grade for the whole performance of the students (see table 1). Hence, the scoring method of the training was abandoned for the actual assessment. After the assessment each of the tested students was asked to fill in a questionnaire about his or her views on the assessment and to grade his or her own performance. Additionally, the raters were asked to give information on their experience in assessing students, their views on the idea of rater training and on their own performance (see figure 1), and (in case of the trained raters) their satisfaction with the training on a five-point scale. The videos of the examinations were collected from the examination rooms, cut, and the allocation to trained and untrained raters was made anonymous.

Video-based re-assessment

All the videotaped examinations were re-evaluated by three observers, one faculty member and two fifth-year students who as a group underwent the same training described above (moderated by author CV). All videotapes were evaluated by global rating at first. Subsequently the observers performed the dimension-evaluation that had been applied in the training concluding with a second overall rating. The observers rated the videos independently of each other and were unaware of the randomisation. The reference rating for the analysis was defined as the median of the three observers' ratings.

Statistics

The grades are given as medians with 1st and 3rd quartiles. For reasons of graphical presentability the mean ±standard error of means (SEM) and ±standard deviation (SD), respectively, are used in the figures. The range of the ratings is given as the mean standard deviation per rater ±SD. Kendall's coefficient of concordance was calculated for every pair of observers and for all three observers together. The primary outcome measure was the difference between the raters' and the observers' ratings given as absolute value. The determining factor was the training and the studied entity were the students considering that every rater evaluated several students. This model was analysed by generalised estimating equations with exchangeable correlation structures. Parameter estimates β with standard errors are presented. Likewise, the effect of experience on accuracy and the effect of the training on grading were investigated using generalized estimating equations.
For the self-assessment of students, the effect of training on the self-assessed grades as well as on the agreement with the raters' grade was analyzed using Mann-Whitney U tests. Moreover, concordance between the raters' grading and the self-assessment was estimated by Kendall's coefficient. To control for multiple tests, we adhered to the following test hierarchy: Firstly we tested the concordance between the three observers using a significance level of 5%. Only if this was significant, we tested whether the training had an effect on the accuracy, again at a significance level of 5%. All the other tests are reported for descriptive purposes only. All analyses were performed using SPSS and R, version 2.15.0 [http://www.R-project.org].

Results

Global results

All twenty-one raters completed the study. The characteristics of trained and untrained raters are given in table

Table 2: Dimensions introduced in the rater training serving to discuss the frame of reference.

dimension	according requirement
structure of the examination	manoeuvres in reasonable order?
completeness	all required manoeuvres accomplished?
thoroughness	manoeuvres executed correctly?
guidance of the procedure	actions reasonable/formal/unnecessary?
guidance of the patient	orders reasonable? control and emphasis?
humanistic qualities	empathy?
time management	adherence to the time limit?

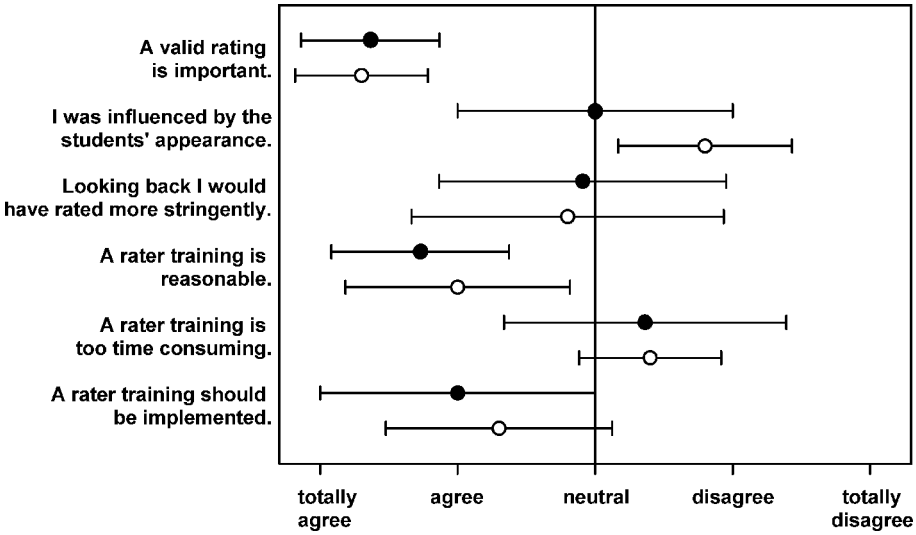

Figure 1: Raters' views on the idea of rater training and on their own performance in a five-point scale. The data are given as means±SD.

Table 3: Characteristics of the raters

	trained raters (n=11)	untrained raters (n=10)
age (mean of years ±SEM)	37,0±1,1	30,7±1,2
male / female	9 / 2	5 / 5
resident / internist / consultant	5 / 2 / 4	7 / 2 / 1
rating experience in medical exams	8	4

3. The randomly chosen training group was older and there were more males, and more senior and experienced physicians in this group. Of the 247 students scheduled for the assessment, 242 (98%) completed the assessment, and 218 assessments (90%) were successfully taped on video. 208 students of the latter group (95%) completed the questionnaire. The median of the number of rated students per rater was 11 in each group (4-12 in the untrained and 5-12 in the trained group, respectively).

Observers' ratings and their concordance

To assess the accuracy of the ratings, the median of the global ratings of the three observers was used as comparison. The difference between this median and the grade of the rater was used to estimate the (lack of) accuracy. To evaluate the adequacy of this, we estimated the coefficient of concordance between the observers, which was 0.70 (P=5.84x10^{-19}). The concordance was higher between the two student observers (0.90, P=6.58x10^{-12}) than between the faculty member and the students (0.70, P=1.26x10^{-4} and 0.73, P=1.01x10^{-5}, respectively). Sixty-one and 75% of the students' ratings equalled the median of all three raters, respectively, and 30% of the faculty's ratings. The median overall grading [1st;3rd quartile] of the observers was 2 [1-;2-] (German school grading, see table 1). Comparing the grades among observers, the faculty's median grade [1st;3rd quartile] was more stringent than the students' grades (2- [2+;3] versus 2 [1-;2-] both). The

overall ratings of the observers after assessing the seven dimensions (see table 2) were virtually the same as these ratings and did not enter further analysis.

Effect of training on grading and accuracy, effect of experience on accuracy

The median overall grading [1st;3rd quartile] of the trained raters was 2 [1-;2-] and of the untrained raters 2+ [1;2], respectively. The pairs of means (±SEM) of the raters' and median observers' gradings are given in figure 2. In the generalised estimating equations model, the trained raters were more stringent than those without the training (β=-0.94 ±0.36, *P*=0.01). No effect of the training on rating accuracy was detectable (β=-0.09 ±0.20, *P*=0.64). The factor experience of the raters did not have any influence on the accuracy of the ratings (β=-0.12 ±0.17, *P*=0.48).

Self-assessment by the students

Similar to the grades of the raters, the students in the group with trained raters assessed themselves more stringently than the students in the group with untrained raters (2 [2+;2] and 2+ [1-;2], respectively; *P*=0.01 from the Mann-Whitney U test). The concordance between the raters' grading and the self-assessment of the students was high in both groups (Kendall's coefficient 0.83 and 0.80 in the group with trained and untrained raters, respectively, $P=1.29x10^{-5}$ and $P=1.25x10^{-4}$). However, students in the group with trained raters disagreed more strongly with their assessment, finding their grade more often inadequate ($P=5.74x10^{-3}$ from the Mann-Whitney U test).

The range of grades applied by each rater did not differ between the groups. The mean standard deviations of the grades were 0.56 ±0.18 in the group of trained raters and 0.61 ±0.15 in the group of untrained raters. The corresponding standard deviation of the observers' medians were 0.67 ±0.26 and 0.66 ±0.19, and of the students' self-assessment 0.49 ±0.21 and 0.50 ±0.10.

The raters' views on the idea of a rater training and on their own performance are given in figure 1. Of the eleven trained raters, ten agreed with the notion that he or she felt more secure in their judgement after the training; one rater was neutral in this regard.

Discussion

The present study failed to show an effect of a rater training on the raters' accuracy. Trained raters were more stringent than their untrained counterparts but did not apply a wider range of grades. These results largely reflect the outcome of previous studies on rater training in a medical context. In the study by Newble and coworkers [19], raters were asked to fill in a rating form and rating quality was measured by the consistency of the raters in assessing five videotaped encounters. Similar to our study, this study focussed on physical examination skills. Despite the rather specific task, the overall consistency was only moderate to acceptable and did not change after the training. The most inconsistent ratings were given in the items "general approach to the patient" and "general observation", indicating that global rating categories (as applied in our study) were more difficult to agree on than more specific categories.

Holmboe and co-workers studied the effects of a four-day faculty development course on the rating of nine scripted videotaped clinical encounters using a mini-CEX rating form [20]. The trained faculty members felt significantly more comfortable with their evaluations of real-live encounters in a follow-up survey. After eight months the participants were re-assessed. The trained raters were found to rate significantly more stringent partially with smaller ranges of ratings. The accuracy of ratings was represented by the capability to discriminate three different levels of competence displayed in the videos. This discrimination was good in the trained and untrained raters both before and after the training. Although the approach in this study was fundamentally different to ours, the higher stringency of the trained raters and the lack of evidence for an effect on accuracy very much resemble the results of our study.

The effects of a rater training on accuracy was more specifically studied by Cook and coworkers [21]. Eighteen of the thirty-two videos used in the pre- and post-test, respectively, were the same scripted videos used by Holmboe and co-workers. The time span between training and the re-assessment in this study was one month. Accuracy was estimated by discrimination of the mean ratings between the scripted levels of competence, by the frequency with which ratings matched scripted performance, and (because of disagreements with the scripted performance ratings) by a chance-corrected agreement using intraclass correlation coefficients. The rater training had no effect at all on either of these accuracy measures. Notably, the interrater reliability for the ratings in the subcategory "physical examination" was comparably small. This might indicate that it was particularly difficult to achieve an agreement on the performance ratings in physical examination.

Our study differed to the previous studies on rater training in one decisive point. While the other studies used prepared video scenes to assess rating scores, we investigated real-live student-SP encounters and re-assessed them by video recordings. Re-assessing examinations by videotapes may have an impact on the ratings and has been formerly studied. In a study dealing with an OSCE assessing joint examination skills, the investigators found a moderate interrater-reliability between live and video raters [22]. The authors point out that the range was similar to previously published interrater-reliability scores of live raters [23]. A second study with pharmacy students specifically studied the intra-rater reliability after one month [24]. The reliability was high; however, due to a higher stringency in the video rating, more candidates would have failed in the post-hoc assessment. A higher

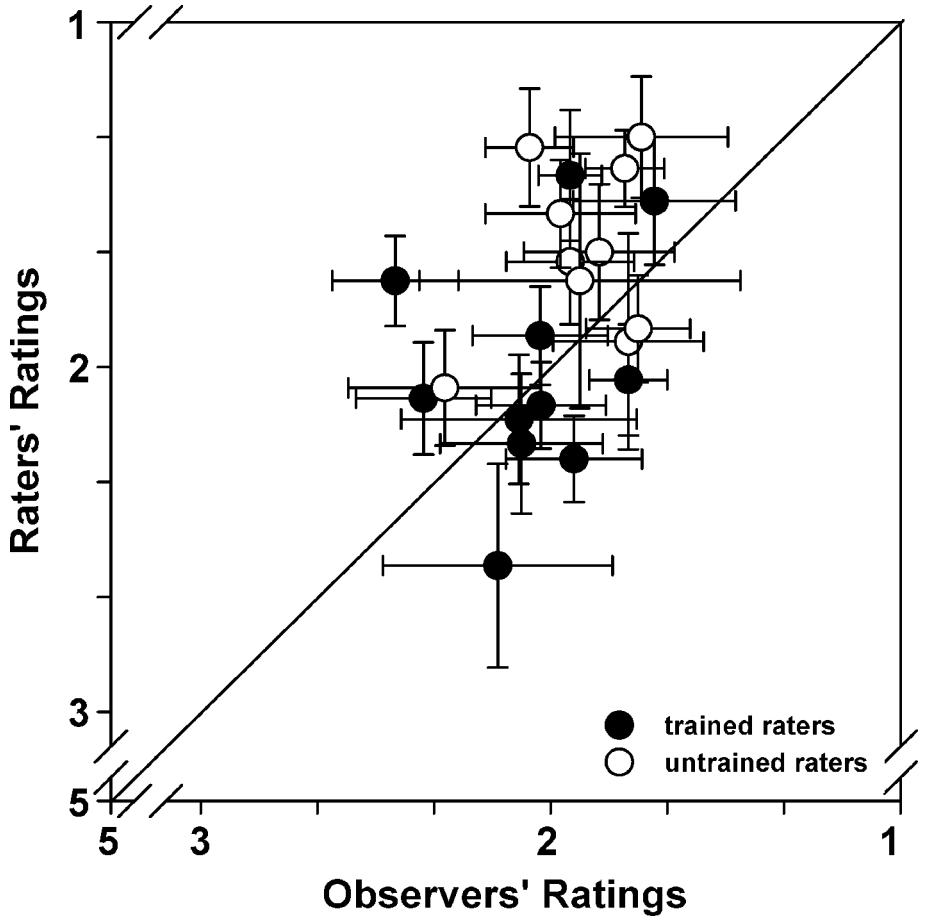

Figure 2: Means±SEM of all ratings per rater and the corresponding means of ratings by the observers (only the respective median of the three observers was taken into account). The distance from the diagonal indicates the degree of inconsistency between raters and observers, hence, the lack of accuracy. Thus, the trained raters were not more accurate but less lenient than the untrained raters. For the grading code see table 1; 1=excellent, 2=good; 3=satisfactory.

stringency in video ratings had already been observed in the first study on joint examinations and in tendency was also present in our study. This effect is most likely due to the fact that an on-scene rater has to announce his judgement face-to-face to the student, while a video observer does not have to take responsibility for his ratings. Announcing decisions face-to-face indeed influences the ratings towards greater leniency [25]. Since this affected both groups equally in our study, we do not consider it crucial for the interpretation of our data.

To overcome the problem of low interrater-reliability in rating medical encounters [26], we re-assessed the videotaped encounters by three observers each. Two of the observers were senior students; the third observer was a faculty member. Trained students have been shown to be equally reliable in rating the practical skills of their junior peers than faculty staff [27], [28]. Latter studies also show that faculty staff rate more stringently than do student assessors. This was obvious in our study. Hence, by choosing the median of the three observers as the measure for accuracy, the student observers' ratings dominated the re-ratings. This might be a concern in the interpretation of the data. Moreover, the randomisation process in our study skewed the allocation of the raters to the groups: the raters in the training group were older, more likely male, and senior and they were more frequently experienced in testing students. These factors have been shown to have no [29], [30] or marginal [31] influence on ratings. Accordingly, in our study we were also unable to find an influence of the factor "rating experience" on rating accuracy.

Other concerns might be the size of the study and the type of intervention. To reduce the effect of intra-observer variety we tried to achieve a sample size of at least ten examinees per rater. Due to the size of the students' cohort, the number of raters was therefore limited to a little over twenty. This was also the number of physicians we were able to recruit from the medical departments for the time of the exams. The time limit of the training was related to the time spent for the ratings (ninety minutes on either day). A greater number of raters or more training would not have been feasible in our setting. We also believe that the effort of a more intensive intervention with the chance of a measurable effect on accuracy would not match the benefit.

However, some other aspects of the study seem noteworthy. Firstly, the time between the training and the exams was relatively short implying that the effect of the training was still present at the time of the ratings. Secondly, the task to be presented by the students was

very clear and uniform. Hence, case specificy and contextual factors as sources of rater errors [14], [32] could largely be eliminated from the experiment. And thirdly, one can also argue that the training indeed had some kind of effect on accuracy. The stringency of overall grading of the trained raters was significantly closer to the observers' gradings and (despite the lack of individual accuracy) can be viewed as more accurate for the group. Consequently, the trained raters were rather less lenient than more stringent. The effect had already been observed in the study by Holmboe and coworkers [20] and suggests that the training in a way helped to standardise the raters' frame of reference by assigning a more appropriate range of ratings. However, the idiosyncrasy of processing the observations and converting the judgements to an ordinal scale [33] within this range obviously remained unaffected.

The untrained raters also denied the possibility of a halo effect in their ratings more consistently than the trained raters. This might well be a training effect and implies that training may be able to raise the awareness of a cognitive bias. Moreover, the students assessed by the trained raters rather felt incorrectly judged, stating more often that their grading was inadequate. This can easily be explained by the more stringent grades in this group. The observation that despite this difference there was a similarly high concordance between the self-assessment of the students and the raters' gradings in both groups could be due to the fact that the students filled in the questionnaire straight after the announcement of the grade. Hence, although the students in the trained raters' group were more likely discontent with the grading, their self-assessment was strongly influenced by the raters' judgement.

In conclusion, our study focussed on the curricular assessment of a very specific task, a brief and highly standardised physical examination. Rater training failed to have an impact on the raters' individual accuracy. However, the stringency of ratings was more in line with the observers' assessment when the raters were trained. Moreover, the trained raters were rather aware of a halo effect and their ratees were more likely discontent with their grade. The data suggest that rater training did have some kind of effect but that the idiosyncrasy of judgement in assessing complex medical skills is too strong to be influenced by a single training. The effort of implementing rater training in order to improve fairness of exams may therefore not be effective. Ratings of medical performance, however, should be interpreted with discretion.

Acknowledgements

The authors are deeply indebted to Prof. Jana Jünger and Dr. Andreas Möltner from the Center of Excellence for Assessment in Medicine Baden-Württemberg (Heidelberg, Germany) for sharing their expertise during the planning phase and in the statistical evaluation, and for providing video equipment. We would also like to thank Sebastian Sosnowki and Christopher Beck for their excellent assistance in videotaping the exams and evaluating the recordings. We also sincerely acknowledge the commitment of Jennifer Miles Davis in proof reading the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Horwitz RI, Kassirer JP, Holmboe ES, Humphrey HJ, Verghese A, Croft C, Kwok M, Loscalzo J. Internal medicine residency redesign: proposal of the Internal Medicine Working Group. Am J Med. 2011;124(9):806-812. DOI: 10.1016/j.amjmed.2011.03.007
2. Clark D, III, Ahmed MI, Dell'italia LJ, Fan P, McGiffin DC. An argument for reviving the disappearing skill of cardiac auscultation. Cleve Clin J Med. 2012;79(8):536-537, 544. DOI: 10.3949/ccjm.79a.12001
3. Smith MA, Burton WB, Mackay M. Development, impact, and measurement of enhanced physical diagnosis skills. Adv Health Sci Educ Theory Pract. 2009;14(4):547-556. DOI: 10.1007/s10459-008-9137-z
4. Ramani S, Ring BN, Lowe R, Hunter D. A pilot study assessing knowledge of clinical signs and physical examination skills in incoming medicine residents. J Grad Med Educ. 2010;2(2):232-235. DOI: 10.4300/JGME-D-09-00107.1
5. Alexander EK. Perspective: moving students beyond an organ-based approach when teaching medical interviewing and physical examination skills. Acad Med. 2008;83(10):906-909. DOI: 10.1097/ACM.0b013e318184f2e5
6. Ainsworth MA, Rogers LP, Markus JF, Dorsey NK, Blackwell TA, Petrusa ER. Standardized patient encounters. A method for teaching and evaluation. JAMA. 1991;266(10):1390-1396. DOI: 10.1001/jama.1991.03470100082037
7. Barley GE, Fisher J, Dwinnell B, White K. Teaching foundational physical examination skills: study results comparing lay teaching associates and physician instructors. Acad Med. 2006;81(10 Suppl):S95-S97. DOI: 10.1097/00001888-200610001-00024
8. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med. 2003;138(6):476-481. DOI: 10.7326/0003-4819-138-6-200303180-00012
9. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 2004;38(2):199-203. DOI: 10.1111/j.1365-2923.2004.01755.x
10. Pelgrim EA, Kramer AW, Mokkink HG, van den EL, Grol RP, van der Vleuten CP. In-training assessment using direct observation of single-patient encounters: a literature review. Adv Health Sci Educ Theory Pract. 2011;16(1):131-142. DOI: 10.1007/s10459-010-9235-6
11. Chen W, Liao SC, Tsai CH, Huang CC, Lin CC, Tsai CH. Clinical skills in final-year medical students: the relationship between self-reported confidence and direct observation by faculty or residents. Ann Acad Med Singapore. 2008;37(1):3-8.
12. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. Ann Intern Med. 1998;129(1):42-48. DOI: 10.7326/0003-4819-129-1-199807010-00011

13. Noel GL, Herbers JE Jr, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? Ann Intern Med. 1992;117(9):757-765. DOI: 10.7326/0003-4819-117-9-757

14. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. Med Educ. 2011;45(10):1048-1060. DOI: 10.1111/j.1365-2923.2011.04025.x

15. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently : Mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract. 2013;18(3):325-341. DOI: 10.1007/s10459-012-9372-1

16. Woehr DJ. Rater training for performance appraisal: a quantitative review. J Occup Organ Psychol. 1994;67:189-205. DOI: 10.1111/j.2044-8325.1994.tb00562.x

17. Lievens F. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. J Appl Psychol. 2001;86(2):255-264. DOI: 10.1037/0021-9010.86.2.255

18. Gorman CA, Rentsch JR. Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. J Appl Psychol. 2009;94(5):1336-1344. DOI: 10.1037/a0016476

19. Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. Med Educ. 1980;14(5):345-349. DOI: 10.1111/j.1365-2923.1980.tb02379.x

20. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. Ann Intern Med. 2004;140(11):874-881. DOI: 10.7326/0003-4819-140-11-200406010-00008

21. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med. 2009;24(1):74-79. DOI: 10.1007/s11606-008-0842-3

22. Vivekananda-Schmidt P, Lewis M, Coady D, Morley C, Kay L, Walker D, Hassell AB. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. Arthritis Rheum. 2007;57(5):869-876. DOI: 10.1002/art.22763

23. Newble DI, Hoare J, Elmslie RG. The validity and reliability of a new examination of the clinical competence of medical students. Med Educ. 1981;15(1):46-52. DOI: 10.1111/j.1365-2923.1981.tb02315.x

24. Sturpe DA, Huynh D, Haines ST. Scoring objective structured clinical examinations using video monitors or video recordings. Am J Pharm Educ. 2010;74(3):44. DOI: 10.5688/aj740344

25. Klimoski R, Inks L. Accountability forces in performance appraisal. Organ Behav Hum Decis Proc. 1990;45:194-208. DOI: 10.1016/0749-5978(90)90011-W

26. Martin JA, Reznick RK, Rothman A, Tamblyn RM, Regehr G. Who should rate candidates in an objective structured clinical examination? Acad Med. 1996;71(2):170-175. DOI: 10.1097/00001888-199602000-00025

27. Ogden GR, Green M, Ker JS. The use of interprofessional peer examiners in an objective structured clinical examination: can dental students act as examiners? Br Dent J. 2000;189(3):160-164.

28. Chenot JF, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B, Stanske B, Kochen MM, Himmel W. Can student tutors act as examiners in an objective structured clinical examination? Med Educ. 2007;41(11):1032-1038. DOI: 10.1111/j.1365-2923.2007.02895.x

29. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Intern Med. 1992;7(5):506-510. DOI: 10.1007/BF02599454

30. Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. Acad Med. 2010;85(10 Suppl):S25-S28. DOI: 10.1097/ACM.0b013e3181ed1aa3

31. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ. 2006;6:42. DOI: 10.1186/1472-6920-6-42

32. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003;15(4):270-292. DOI: 10.1207/S15328015TLM1504_11

33. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 2011;86(10 Suppl):S1-S7. DOI: 10.1097/ACM.0b013e31822a6cf8

Corresponding author:
PD Dr. med. Gunther Weitz, MME
Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Medizinische Klinik I, Ratzeburger Allee 160, 23538 Lübeck, Deutschland, Tel.: +49 (0)451/500-6033, Fax: +49 (0)451/500-6242
gunther.weitz@uksh.de

Please cite as
Weitz G, Vinzentius C, Twesten C, Lehnert J, Bonnemeier H, König IR. Effects of a rater training on rating accuracy in a physical examination skills assessment . GMS Z Med Ausbild. 2014;31(4):Doc41. DOI: 10.3205/zma000933, URN: urn:nbn:de:0183-zma0009338

This article is freely available from
http://www.egms.de/en/journals/zma/2014-31/zma000933.shtml

Received: *2014-01-08*
Revised: *2014-03-24*
Accepted: *2014-08-20*
Published: *2014-11-17*

Einfluss einer Prüferschulung auf die Genauigkeit der Bewertung einer Untersuchungskursprüfung

Gunther Weitz[1]
Christian Vinzentius[2]
Christoph Twesten[1]
Hendrik Lehnert[1]
Hendrik Bonnemeier[3]
Inke R. König[4]

1 Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Medizinische Klinik I, Lübeck, Deutschland

2 Institut für Qualitätsentwicklung an Schulen Schleswig-Holstein, Kronshagen, Deutschland

3 Universitätsklinikum Schlesweig-Holstein, Campus Kiel, Medizinische Klinik III, Kiel, Deutschland

4 Universität zu Lübeck, Institut für Medizinische Biometrie und Statistik, Lübeck, Deutschland

Zusammenfassung

Hintergrund: Die Genauigkeit und Reproduzierbarkeit von Prüferurteilen im Medizinstudium ist gering. Eine Schulung von Prüfern hat keinen oder allenfalls minimalen Effekt. Die dazu verfügbaren Studien beziehen sich jedoch auf die Beurteilung von Arzt-Patienten-Interaktionen in eigens dafür angefertigten Videos. Wir untersuchten, ob eine Schulung, die sich auf den Bezugsrahmen des Prüfers bezieht, die Prüfergenauigkeit bei curricularen Untersuchungskurstestaten verbessert.
Methoden: 21 Prüfer testierten 242 Studierende im dritten Studienjahr. Elf der Prüfer wurden randomisiert ausgewählt, an einer kurzen Prüferschulung teilzunehmen, die wenige Tage vor dem Testat stattfand. 218 Testate konnten auf Video festgehalten werden und wurden später unabhängig von drei Nachprüfern bewertet. Genauigkeit definierten wir als die Konkordanz zwischen der Benotung des eigentlichen Prüfers und dem Median der Benotung der Nachprüfer. Im Anschluss an das Testat füllten sowohl Prüflinge als auch Prüfer einen Fragebogen zum Testat aus.
Ergebnisse: Die Prüferschulung hatte keinen messbaren Einfluss auf die Genauigkeit der Bewertung. Die geschulten Prüfer waren aber strenger als die ungeschulten und ihr Notenspektrum lag eher in dem Bereich des Spektrums der Nachprüfer. Außerdem waren die geschulten Prüfer sich des Halo-Effektes stärker bewusst. Obwohl die Selbsteinschätzung der Studierenden in beiden Gruppen nahe bei der Prüfernote lag, waren die Studierenden, die von geschulten Prüfern testiert wurden, häufiger mit ihrer Note unzufrieden.
Diskussion: Trotz einiger marginaler Effekte hatte die Prüferschulung keinen Effekt auf die Genauigkeit der Bewertung. Diese Beobachtung bei echten Testaten stimmt mit den Ergebnissen von Studien mit Videobewertungen überein. Auch die starke Standardisierung der Aufgabe im Testat half nicht, das Prüferurteil zu harmonisieren. Unsere Studie bestätigt, dass die Bewertung ärztlicher Tätigkeiten individuell sehr unterschiedlich ist. Eine Schulung, die wie in unserem Versuch auf den Bezugsrahmen des Urteils abzielt, ist nicht in der Lage, die ärztliche Bewertung von Testatleistungen zu vereinheitlichen.

Schlüsselwörter: Prüferschulung, Prüfergenauigkeit, Testat, körperliche Untersuchung, randomisierte kontrollierte Studie

Einleitung

Die körperliche Untersuchung ist eine Kernkompetenz im klinischen Alltag. Eine wesentliche Aufgabe der ärztlichen Ausbildung muss es daher sein, die Beherrschung von körperlichen Untersuchungstechniken sicher zu vermitteln. Kürzlich publizierte Studien machen allerdings auf wachsende Defizite auf diesem Gebiet bei Absolventen des Medizinstudiums aufmerksam [1], [2]. Zu dieser Entwicklung tragen der Mangel an geeigneten Patienten, geeigneten Dozenten und Unterrichtszeit am Patientenbett bei [3], [4]. Außerdem führt die zunehmende Spezialisierung in der Medizin zu Apparategläubigkeit und Betriebsblindheit [4], [5]. Das Vermitteln und Prüfen von Fertigkeiten der körperlichen Untersuchung dürfte daher mehr und mehr zur Herausforderung im Medizinstudium werden.
In den letzten Jahrzehnten wurden zahlreiche Anstrengungen unternommen, die Qualität der Vermittlung von Fertigkeiten der körperlichen Untersuchung zu verbessern. Dazu gehören die Einführung von Schauspielerpatienten und Patienteninstruktoren [6], [7], die Anwendung von Checklisten und Bewertungsbögen [8], das Implementieren von OSCEs [9] und die systematische Beobachtung von Arzt-Patienten-Interaktionen [10]. Nicht immer hatten diese Maßnahmen den gewünschten Effekt. So gaben in

einer Studie aus Taiwan 22% der Studierenden im letzten Studienjahr an, niemals bei einer körperlichen Untersuchung supervidiert worden zu sein (36% nicht von Lehrpersonal) und 10% fühlten sich bei der Prozedur noch unsicher [11].

An unserer Fakultät muss jeder Student im dritten Studienjahr eine körperliche Untersuchung von Kopf bis Fuß im Rahmen eines Untersuchungskurstestates vorführen. Dieses Testat findet unmittelbar im Anschluss an ein fünfwöchiges Tutorium statt. In den Evaluationen wird allerdings vielfach eine mangelnde Fairness bei der Benotung beklagt. In der Tat sind Reliabilität und Genauigkeit der Bewertung durch Lehrpersonal gering [12]. Die Strukturierung der Bewertung mittels Bewertungsbögen kann zwar die Genauigkeit der Beobachtungen verbessern, hat aber keinen Einfluss auf die Übereinstimmung von Gesamtbeurteilungen [13]. Das dürfte daran liegen, dass die Strategien der Prüfer, Informationen zu integrieren, eher individuell geprägt sind und dass sich der Bezugsrahmen der Prüfer stark unterscheidet [14], [15]. Studien aus der Personalpsychologie zeigen, dass Schulungen, die sich auf diesen Bezugsrahmen beziehen, durchaus die Genauigkeit von Personalbeurteilungen verbessern können [16], [17]. Ziel eines solchen Trainings ist es, Prüfern eine gemeinsame Konzeptualisierung der zu beurteilenden Aufgabe zu vermitteln. Dabei sollen sich verlässlichere Bewertungsschemata entwickeln [18]. Diese Art der Prüferschulung wollten wir daher auf unser Setting übertragen.

Erstaunlicherweise gibt es kaum Studien zum Thema Prüferschulung in der medizinischen Ausbildung und die Ergebnisse sind eher enttäuschend. In einer kleinen Studie untersuchten Newble und Mitarbeiter den Einfluss einer Prüferschulung auf die Bewertung von fünf gefilmten Untersuchungstechniken [19]. Die Prüfer erhielten entweder keine Intervention, ein Feedback über die Prüferleistung in der zweiten Gruppe oder zusätzlich eine Prüferschulung mit Diskussion eines weiteren Videos in der dritten Gruppe. Als die fünf Filme zwei Monate später erneut bewertet wurden, gab es keine messbaren Unterschiede im Prüferurteil der drei Gruppen. Holmboe und Mitarbeiter untersuchten die Effekte einer intensiven multidimensionalen Prüferschulung auf die Bewertungen von gefilmten Arzt-Patienten-Interaktionen acht Monate nach der Schulung [20]. Die geschulten Prüfer waren strenger und nutzten bei einigen Bewertungen eine kleinere Notenskala. Des Weiteren untersuchten Cook und Mitarbeiter die Effekte einer ähnlichen, aber kürzeren Schulung auf die Interrater-Reliabilität und die Genauigkeit von Mini-CEX-Bewertungen in einem Weiterbildungsprogramm [21]. Die Schulung hatte keinen Einfluss auf diese Parameter.

Wir wollten nun untersuchen, ob eine Prüferschulung die Genauigkeit der Bewertungen in unserem Untersuchungskurstestat verbessern könnte. Unser Setting unterscheidet sich in mehrfacher Hinsicht von dem der zitierten Studien: Zum einen fokussiert unser Testat auf eine klar umrissene Fertigkeit statt auf die Bewertung einer allgemeinen Arzt-Patienten-Interaktion. Zum zweiten haben wir die Aufgabe für alle Beteiligten genau definiert und Prüflinge wie Prüfer waren mit diesem Standard vertraut. Zum dritten geht es in der vorliegenden Studie um echte Prüfungssituationen statt um die Bewertung gestellter Filmszenen. Das bedeutet, dass die Bewertungen relevant waren und der Prüfer die Note den Prüflingen auch mitteilen musste. Um die tatsächliche Leistung der Studierenden abschätzen zu können, haben wir die Videoaufnahme jeder einzelnen Prüfung drei Nachprüfern zur Bewertung vorgelegt. Die Überlegung war, dass die Bewertungen der trainierten Prüfer näher an der Einschätzung der Nachprüfer liegen würden und demnach genauer wären. Zudem wollten wir die Effekte der Prüferschulung auf Strenge der Prüfer und die Ausnutzung der Notenskala untersuchen.

Methoden

Curriculärer Zusammenhang

Das Untersuchungskurstestat ist der Abschluss eines Untersuchungskurstutoriums am Beginn des dritten Studienjahres. Ziel des Tutoriums ist es, den Studenten die Grundlagen der allgemeinen körperlichen Untersuchung zu vermitteln. Nach dem Tutorium geht der Untersuchungskurs mit Unterricht am Krankenbett weiter, wo die Studierenden die erlernten Fertigkeiten praktisch anwenden. Aufgabe im Testat ist, eine standardisierte Untersuchung von Kopf bis Fuß vorzuführen. Die Untersuchung umfasst die Inspektion von Kopf und Mundhöhle, die Inspektion und Palpation vom Hals, die komplette Untersuchung von Thorax und Abdomen, die korrekte Messung des Blutdrucks an einem Arm, die Erhebung des Pulsstatus, sowie die Inspektion der Extremitäten. Ein Anleitungsvideo ist für alle Studierenden auf unserer Homepage frei zugänglich. Weitere Untersuchungstechniken wie die genitorektale Untersuchung, die neurologische Untersuchung und die Untersuchung des Bewegungsapparates werden in anderen Teilen des Kurses vermittelt.

Das Tutorium findet in den ersten fünf Wochen des Wintersemesters statt. Es besteht aus fünf Abschnitten mit jeweils zwei Vorlesungsstunden und jeweils anderthalbstündigem Kleingruppenunterricht, in dem sechs Studierende sich unter Anleitung eines erfahrenen Internisten gegenseitig untersuchen. Das Testat findet in der sechsten Semesterwoche statt. Aufgabe ist es, die standardisierte Untersuchung in einem Zeitlimit von zehn Minuten an einem Schauspielerpatienten vorzuführen. Die Prüfer sind Ärzte der Medizinischen Kliniken. Sie beobachten den Untersuchungsgang, geben Feedback und bewerten die Leistung mit einer Schulnote (siehe Tabelle 1). Sie greifen weder in die Untersuchung ein, noch stellen sie Theoriefragen. Jeder Prüfer prüft sechs Studierende in einem Zeitrahmen von 15 Minuten pro Prüfling an jeweils zwei Tagen. Die Schauspielerpatienten sind gesunde Studierende. Sie sollen passiv agieren und nur eindeutigen Anweisungen folgen.

Tabelle 1: Deutsche Schulnoten. Die Prüfer durften die Noten 1 bis 4 mit + bzw. - auf- bzw. abwerten.

1 = sehr gut
2 = gut
3 = befriedigend
4 = ausreichend
5 = mangelhaft

Für diese Studie wählten wir 21 Ärzte aus, die 242 Testate abnehmen sollten. Alle 21 Prüfer waren mit den Lernzielen des Tutoriums, dem Anleitungsvideos und den Feedback-Regeln vertraut. Elf der 21 Personen wurden randomisiert der Prüferschulung zugeordnet. Für die Randomisierung wurden die Prüfer nummeriert und mittels Zufallszahlengenerator einer einschlägigen Internetseite den beiden Gruppen zugeteilt. Zur Bestimmung der Prüfergenauigkeit wurden alle Testate auf Video aufgenommen. Dazu gaben sowohl alle Prüfer als auch alle Studierenden vor der Studie ihr schriftliches Einverständnis. Die Studie war der Ethikkommission zur Begutachtung vorgelegt worden und es gab keine Einwände. Das Protokoll befindet sich im Einklang mit der Helsinki-Deklaration und die Anonymität aller Teilnehmer wurde gewahrt.

Intervention

Die elf Prüfer, die für die Schulung ausgewählt waren, wurden in zwei Gruppen (zu 6 und 5) aufgeteilt um eine kleinere Gruppengröße zu erreichen. Die Schulung fand am Ende der fünften Semesterwoche kurz vor den Testaten statt (Schulung Donnerstag- und Freitagnachmittag, Testate Montag- und Dienstagnachmittag). Für die Schulung wurden 90 Minuten angesetzt. In einer kurzen Begrüßung erläuterte der Moderator (Autor GW) Ziele und Standards der Testate und die Bewertungsdimensionen (siehe Tabelle 2). Im Anschluss wurden vier Beispielvideos gezeigt, in denen Studierende des vierten Studienjahres die standardisierte Untersuchung in unterschiedlicher Qualität an Schauspielerpatienten durchführten. Die Reihenfolge der Videos war in beiden Schulungsgruppen gleich. Nach jeder Präsentation wurden die Prüfer gebeten, anhand einer Checkliste mit den sieben Dimensionen (siehe Tabelle 2) die jeweilige Leistung einer Schulnote (siehe Tabelle 1) zuzuordnen. Die Prüfer lasen dann ihre Bewertung vor und die Prüfer mit der höchsten und der niedrigsten Schulnote wurden für die jeweilige Dimension gebeten, ihre Bewertung zu rechtfertigen. Die daraufhin entstehende Diskussion wurde vom Erstautor moderiert. Nachdem alle Dimensionen diskutiert waren, erläuterte der Moderator die im Video eingebauten Fehler.

Testate

Da die nicht geschulten Prüfer mit der Checkliste nicht vertraut waren, wurden alle Prüfer gebeten, im Testat lediglich eine Gesamtnote zu vergeben (siehe Tabelle 1). Die feinteilige Bewertung in der Schulung wurde fürs Testat also wieder verlassen. Nach dem Testat wurden alle Studierenden gebeten, einen Fragebogen zu ihren Ansichten über das Testat auszufüllen und sich selbst eine Note für die Testatleistung zu geben. Außerdem wurden die Prüfer gebeten, ihre bisherige Erfahrung als Prüfer offenzulegen und zur Idee der Prüferschulung und ihrer eigenen Leistung im Testat Stellung zu nehmen (siehe Abbildung 1). Die geschulten Prüfer wurden zudem gebeten, ihre Zufriedenheit mit der Schulung auf einer Skala von eins bis fünf wiederzugeben. Die Videos der Testate wurden eingesammelt, geschnitten und hinsichtlich der Allokation (geschulte/ungeschulte Prüfer) anonymisiert.

Bewertung der Videos durch Nachprüfer

Alle Videos der Testate wurden durch drei Nachprüfer (ein Dozent, zwei Studierende im fünften Studienjahr) reevaluiert, die als Gruppe ebenfalls die oben beschriebene Prüferschulung erhalten hatten (moderiert durch Autor CV). Dabei wurde zunächst eine Gesamtnote vergeben, um dann wie im Training die einzelnen Dimensionen zu bewerten, wiederum gefolgt von einer Gesamtnote. Die Bewertung der Nachprüfer erfolgte unabhängig voneinander und bezüglich der Randomisierung verblindet. Der Median der drei Nachprüfernoten wurde als endgültige Note der Nachprüfer definiert.

Statistik

Alle Noten werden als Median mit 1. und 3. Quartile angegeben. Aus Gründen der Veranschaulichung werden in den Grafiken abweichend Mittelwert und Standardfehler bzw. Standardabweichung verwendet. Die Bandbreite der verwendeten Notenskala wird als mittlere Standardabweichung pro Prüfer (ihrerseits mit Standardabweichung) angegeben. Für jedes Paar an Nachprüfern und für alle drei Nachprüfer zusammen wurde Kendalls Konkordanzkoeffizient berechnet. Zielvariable war die Differenz zwischen der Note der Prüfer und der Note der Nachprüfer als Absolutwert. Einflussvariable war das Training, Beobachtungseinheit waren die Prüflinge, wobei in dem Modell berücksichtigt wurde, dass mehrere Prüflinge vom selben Prüfer geprüft wurden. Das Modell wurde mittels verallgemeinerter Schätzgleichungen mit austauschbaren Korrelationsstrukturen ausgewertet. Angegeben werden die Schätzwerte β mit Standardabweichungen. Analog wurde der Effekt von Prüfererfahrung auf die Genauigkeit und der Effekt von Training auf die

Tabelle 2: Dimensionen für die Prüferschulung zur Diskussion des Bezugsrahmens.

Dimension	entsprechende Anforderung
Strukturiertheit der Untersuchung	Ordnung in der Vorgehensweise?
Vollständigkeit der Untersuchung	Alle erforderliche Manöver durchgeführt?
Gründlichkeit der Untersuchung	Manöver korrekt durchgeführt?
Zielführung	Manöver sinnvoll/umständlich/überflüssig?
Führen des Schauspielerpatienten	Aufforderungen sinnvoll, Durchsetzung?
Umgang mit dem Probanden	Empathische Gestaltung der Untersuchung?
Zeitmanagement	Vorgegebene Zeit eingehalten?

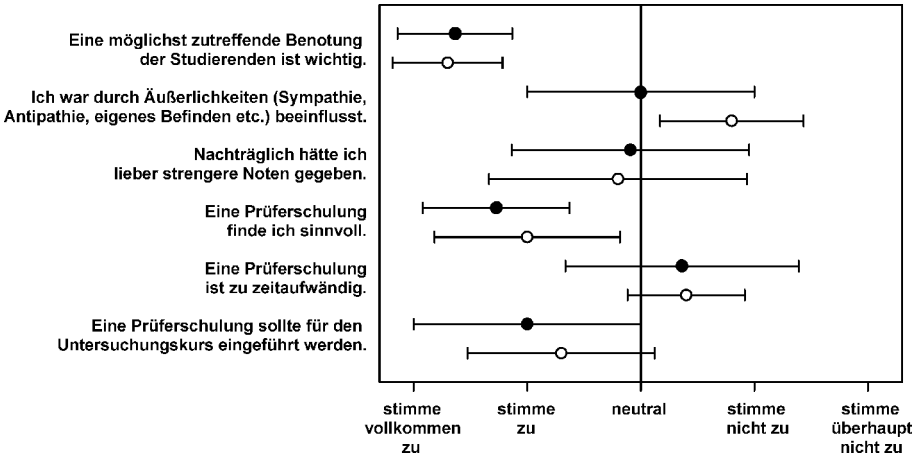

Abbildung 1: Meinungen der Prüfer zur Prüferschulung und zu den eigenen Leistungen im Testat als Fünf-Punkte-Skala (Mittelwerte±Standardabweichung).

Bewertung mittels verallgemeinerter Schätzgleichungen untersucht.

Bezüglich der Selbsteinschätzung der Prüflinge wurden zwei Parameter zwischen den beiden Gruppen von Prüflingen mit geschulten und ungeschulten Prüfern mittels Mann-Whitney U-Test verglichen: die selbst gegebene Note an sich und die absolute Abweichung zwischen selbst gegebener Note und der Note der Prüfer. Darüber hinaus wurde die Konkordanz zwischen der Prüfernote und der selbst gegebenen Note mittels Kendalls Koeffizient abgeschätzt. Zur Kontrolle des multiplen Testens legten wir folgende Testhierarchie fest: Zunächst testeten wir die Konkordanz zwischen den drei Nachprüfern mit einem Signifikanzniveau von 5%. Nur bei Signifikanz sollte getestet werden, ob die Prüferschulung einen Effekt auf die Genauigkeit hat, wiederum mit einem Signifikanzniveau von 5%. Alle anderen Tests sind rein deskriptiv. Die Auswertungen erfolgten mittels SPSS und R Version 2.15.0 [http://www.R-project.org].

Ergebnisse

Stichprobe

Alle 21 Prüfer vollendeten die Studie. Die Charakteristika der geschulten und ungeschulten Prüfer sind in Tabelle 3 wiedergegeben. Die zufällig ausgewählten Prüfer der geschulten Gruppe waren älter sowie häufiger männlich, höhergestellt und prüfungserfahren. Von den 247 Studierenden, die zum Testat angemeldet waren, absolvierten 242 (98%) das Testat und 218 Testate (90%) wurden erfolgreich auf Video dokumentiert. 208 Prüflinge der letzten Gruppe (95%) beantworteten den Fragebogen. Der Median von Prüflingen pro Prüfer war 11 in beiden Gruppen (4-12 Prüflinge in der ungeschulten bzw. 5-12 in der geschulten Gruppe).

Noten der Nachprüfer und ihre Konkordanz

Um die Genauigkeit der Bewertungen abzuschätzen, wurde der Median der Gesamtnoten der drei Nachprüfer als Vergleich herangezogen. Die Differenz zwischen diesem Median und der Prüfernote definierte die Genauigkeit (bzw. Ungenauigkeit). Um die Zulässigkeit dieses Vorgehens zu determinieren, errechneten wir den Konkordanzkoeffizienten zwischen den drei Nachprüfern. Dieser betrug 0,70 (P=5,84x10^{-19}). Die Konkordanz war höher zwischen den studentischen Nachprüfern (0,90; P=6,58x10^{-12}) als zwischen dem Dozenten und den Studenten (0,70; P=1,26x10^{-4} bzw. 0,73; P=1,01x10^{-5}). 71 bzw. 75% der studentischen Noten entsprachen dem Median, während das nur in 30% beim Dozenten der Fall war. Der Median der Gesamtnote [1.;3. Quartile] der Nachprüfer war 2 [1-;2-]. Im Vergleich der Bewertungen der Nachprüfer war der Median des Dozenten (2- [2+;3])

Tabelle 3: Charakteristika der Prüfer

	geschult (n=11)	ungeschult (n=10)
Alter (Jahre; Mittelwert±SEM)	37,0±1,1	30,7±1,2
männlich / weiblich	9 / 2	5 / 5
Assistent / Facharzt / Oberarzt	5 / 2 / 4	7 / 2 / 1
Prüfungserfahrung (Medizin)	8	4

strenger als derjenige der Studenten (beide 2 [1-;2-]). Die Gesamtnote nach Bewertung aller sieben Dimensionen (siehe Tabelle 2) war praktisch identisch mit der primär gegebenen Gesamtnote und wurde nicht weiter verfolgt.

Effekt der Prüferschulung auf Benotung und Genauigkeit, Effekt der Prüferfahrung auf Genauigkeit

Der Median der Gesamtnote [1.;3. Quartile] der geschulten Prüfer war 2 [1-;2-], die der ungeschulten Prüfer 2+ [1;2]. In Abbildung 2 sind die mittleren Gesamtnoten der Prüfer (±Standardfehler) gegen die korrespondierenden Gesamtnoten der Nachprüfer aufgetragen. Im Modell der verallgemeinerten Schätzgleichungen waren die geschulten Prüfer strenger als die ungeschulten (β=-0,94 ±0,36; P=0,01).
Es gab keinen erkennbaren Effekt der Schulung auf die Genauigkeit (β=-0,09 ±0,20; P=0,64). Der Faktor "Prüferfahrung" hatte ebenfalls keinen Einfluss auf die Genauigkeit der Benotung (β=-0,12 ±0,17; P=0,48).

Selbsteinschätzung der Prüflinge

Analog zu den Benotungen der Prüfer schätzten sich die Prüflinge, die von geschulten Prüfern testiert worden waren, strenger ein als die Prüflinge, deren Prüfer nicht geschult waren (2 [2+;2] bzw. 2+ [1-;2]; P=0,01 nach Mann-Whitney U-Test). Die Konkordanz zwischen den Prüfernoten und der Selbsteinschätzung war in beiden Gruppe hoch (Kendalls Koeffizient 0,83 bzw. 0,80 in der Gruppe mit geschulten bzw. ungeschulten Prüfern; P=1,29x10^{-5} bzw. P=1,25x10^{-4}). Allerdings waren die Prüflinge, die von geschulten Prüfern testiert worden waren, bezüglich ihrer Note eher unzufrieden und fanden die Benotung häufiger nicht adäquat (P=5,74x10^{-3} nach Mann-Whitney U-Test).
Die Bandbreite der verwendeten Notenskala unterschied sich nicht zwischen geschulten und ungeschulten Prüfern. Die mittleren Standardabweichungen der gegebenen Noten waren 0,56 ±0,18 in der Gruppe der geschulten und 0,61 ±0,15 in der Gruppe der ungeschulten Prüfer. Die entsprechenden Standardabweichungen der Mediane der Nachbeobachter waren 0,67 ±0,26 bzw. 0,66 ±0,19, die der studentischen Selbsteinschätzungen 0,49 ±0,21 bzw. 0,50 ±0,10.
Die erfragten Meinungen der Prüfer zur Prüferschulung und ihre Sicht auf die eigenen Leistungen sind in Abbildung 1 wiedergegeben. Von den elf geschulten Prüfern gaben zehn an, sich bei der Benotung sicherer gefühlt zu haben. Der elfte Prüfer äußerte sich in dieser Hinsicht neutral.

Diskussion

Die vorliegende Studie konnte keinen Effekt einer Prüferschulung auf die Genauigkeit der Bewertung belegen. Geschulte Prüfer waren strenger als ungeschulte, haben aber die Bandbreite der Notenskala nicht besser genutzt. Diese Ergebnisse spiegeln im Wesentlichen die Ergebnisse anderer Studien zu Prüferschulungen im medizinischen Kontext wider. In der Studie von Newble und Mitarbeitern [19] mussten die Prüfer Bewertungsbögen ausfüllen und die Bewertungsqualität wurde anhand der Übereinstimmung bemessen, mit der die Prüfer fünf gefilmte Situationen bewerteten. Wie in unserer Studie fokussierten die Autoren auf körperliche Untersuchungstechniken. Trotz der hohen Spezifität der Aufgabe war die Übereinstimmung allenfalls schwach bis akzeptabel und veränderte sich nach der Schulung nicht. Die geringsten Übereinstimmungen ergaben sich in den Merkmalen "allgemeiner Zugang zum Patienten" und "allgemeine Beobachtung". Dies könnte darauf hindeuten, dass allgemeine Kategorien (wie in unserer Studie) schwieriger einheitlich zu bewerten sind als konkretere.
Holmboe und Mitarbeiter untersuchten die Effekte eines viertägigen hochschuldidaktischen Kurses auf die Bewertung von neun Filmszenen einer Arzt-Patienten-Interaktion mittels Mini-CEX-Bewertungsbogen [20]. Die geschulten Prüfer fühlten sich in einer späteren Umfrage wesentlich sicherer mit ihrer Einschätzung von tatsächlichen Arzt-Patienten-Interaktionen. Nach acht Monaten wurden die Teilnehmer erneut beurteilt. Die geschulten Prüfer benoteten dabei wesentlich strenger unter geringerer Ausnutzung der Notenskala. Die Genauigkeit wurde als Fähigkeit definiert, zwischen drei verschiedenen Kompetenzstufen der gezeigten Filmszenen zu unterscheiden. Die Differenzierung war gleichermaßen gut bei geschulten wie ungeschulten Prüfern vor und nach dem Training. Obwohl sich dieser Ansatz grundsätzlich von unserem unterscheidet, spiegeln die größere Strenge der geschulten Prüfer und das Ausbleiben eines Effekts der Schulung auf die Genauigkeit in dieser Studie unsere Ergebnisse weitgehend wider.
Die Effekte einer Prüferschulung auf die Genauigkeit der Bewertung wurden noch spezifischer von Cook und Mitarbeitern untersucht [21]. 18 der 32 in Prä- und Posttest

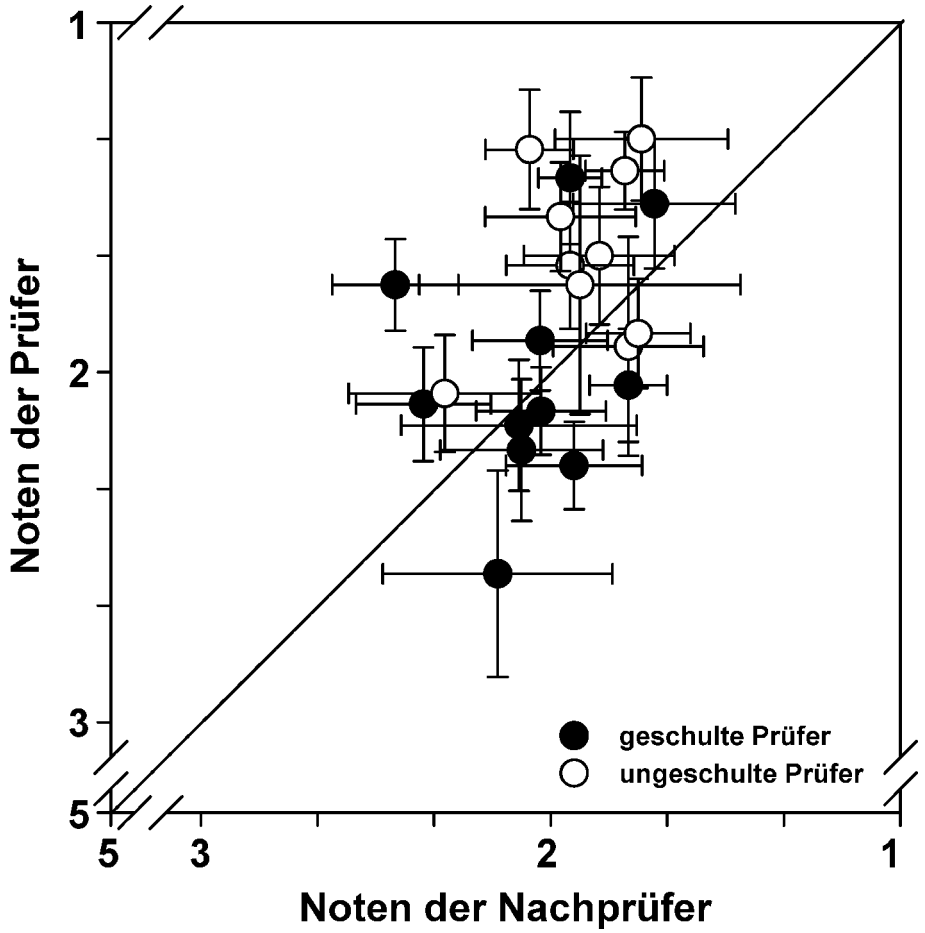

Abbildung 2: Mittelwerte±SEM aller Benotungen (Schulnoten, Tabelle 1) durch jeweils einen Prüfer mit korrespondierenden Mittelwerten der medianen Benotungen der Nachprüfer. Der Abstand von der Diagonalen deutet auf die Inkonsistenz zwischen Prüfern und Nachprüfern, also auf die Ungenauigkeit der Benotung. Demnach waren die geschulten Prüfer nicht genauer, aber weniger großzügig als die ungeschulten.

verwendeten Videos beinhalteten dieselben gestellten Szenen, die schon Holmboe und Mitarbeiter verwendet hatten. Die Zeitspanne zwischen Schulung und Nachuntersuchung betrug in dieser Studie einen Monat. Genauigkeit war definiert als Unterscheidung der Gesamtnoten zwischen den im Video dargestellten Kompetenzstufen, als Häufigkeit der Übereinstimmung der Note mit dem intendierten Ergebnis, und (wegen unterschiedlicher Meinungen über die tatsächlich dargestellten Kompetenzstufen) als zufallskorrigierte Übereinstimmung mittels Intraklassen-Korrelationskoeffizienten. Die Prüferschulung hatte auf keinen dieser Parameter einen Einfluss. Interessanterweise war die Interrater-Reliabilität für Bewertungen in der Unterkategorie "körperliche Untersuchung" auffällig niedrig. Dies könnte darauf hindeuten, dass es besonders schwierig war, in dieser Disziplin Einigkeit zu erzielen.

Unsere Studie unterschied sich von den vorigen Studien zur Prüferschulung in einem entscheidenden Punkt: Während andere Studien für die Bewertungen vorbereitete Videos verwendeten, untersuchten wir tatsächliche Prüfungssituationen, die wir mitschnitten und später nachevaluierten. Diese Nachuntersuchung könnte einen Einfluss auf die Bewertungen haben, der bereits wissenschaftlich untersucht wurde: In einer Studie über ein OSCE zu Gelenkuntersuchungen fanden die Autoren eine moderate Interrater-Reliabilität zwischen der Bewertung in der Prüfung und der Bewertung der Aufzeichnung der Prüfung [22]. Die Autoren betonen aber, dass der Unterschied ähnlich groß war wie die zuvor publizierte Interrater-Reliabilität zwischen zwei Prüfern derselben Prüfung [23]. Eine zweite Studie mit Pharmaziestudenten untersuchte die Intrarater-Reliabilität nach einem Monat [24]. Die Reliabilität war hoch, allerdings wären aufgrund einer größeren Strenge bei der Bewertung der Videoaufzeichnung nach einem Monat mehr Kandidaten durchgefallen. Eine größere Strenge bei der Bewertung von Videoaufzeichnungen war bereits in der vorgenannten Studie über die Gelenkuntersuchungen beobachtet worden und auch wir beobachten in unserer Studie eine solche Tendenz. Dieser Effekt dürfte dadurch bedingt sein, dass ein tatsächlicher Prüfer sein Urteil dem Kandidaten ins Gesicht sagen muss, während der Bewerter eines Videos für seine Bewertung keine unmittelbare Verantwortung übernehmen muss. Das Mitteilen des eigenen Urteils macht Bewertungen in der Tat großzügiger [25]. Da dieser Effekt in unserer Studie aber beide Gruppen gleichermaßen betraf, dürfte er nicht entscheidend für die Interpretation unserer Ergebnisse sein.

Um dem Problem der niedrigen Interrater-Reliabilität bei der Bewertung medizinischer Interaktionen [26] zu begegnen, haben wir alle Prüfungen von drei Nachprüfern noch einmal bewerten lassen. Zwei der Nachprüfer waren ältere Studenten, der dritte Dozent. Untersuchungen haben ergeben, das trainierte Studenten praktische Fertigkeiten ihrer jüngeren Kommilitonen ähnlich verlässlich bewerten wie Dozenten [27], [28]. Diese Studien zeigen auch, dass Dozenten dabei strenger bewerten. Dies war auch in unserer Studie der Fall. Indem wir den Median der drei Nachprüfer als Maß für Genauigkeit gewählt haben, dominierten die studentischen Bewertungen unter den Nachprüfern. Dies könnte ein Problem bei der Interpretation der Ergebnisse darstellen. Darüber hinaus verzerrte der Randomisierungsprozess die Allokation der Prüfer zu den beiden Gruppen: Die Prüfer der Schulungsgruppe waren älter, eher männlich, höhergestellt und häufiger erfahrene Prüfer. Diese Faktoren hatten in anderen Studien allerdings keinen [29], [30] oder allenfalls marginalen [31] Einfluss auf die Qualität von Prüferurteilen. Passend dazu konnten wir in unserer Studie ebenfalls keinen Einfluss des Faktors "Prüfungserfahrung" auf die Prüfergenauigkeit feststellen.

Andere Probleme könnten in der Stichprobengröße und in der Art der Intervention gesehen werden. Um den Beobachterfehler zu minimieren haben wir versucht, auf mindestens zehn Prüflinge pro Prüfer zu kommen. Angesichts der Jahrgangsgröße war die Stichprobe also auf etwas über 20 Prüfer limitiert. Das war gleichzeitig die Menge an Ärzten, die wir für die Testate aus dem laufenden Klinikbetrieb rekrutieren konnten. Die Dauer der Schulung richtete sich nach dem Zeitaufwand für die Testate (90 Minuten an beiden Tagen). Eine größere Anzahl an Prüfern und eine längere Schulung hätten wir nicht bewältigen können. Zudem sind wir der Meinung, dass der Aufwand einer noch intensiveren Schulung (mit möglicherweise messbarem Effekt auf die Genauigkeit) dem potentiellen Nutzen nicht mehr entsprochen hätte.

Einige andere Aspekte erscheinen uns noch erwähnenswert: Zunächst einmal war die Zeitspanne zwischen Schulung und Testaten recht kurz, so dass der Trainingseffekt während der Testate vermutlich noch präsent war. Zweitens war die Aufgabe, die die Prüflinge erfüllen sollten, sehr klar definiert und einheitlich. Kontextuelle Faktoren als Fehlerquellen [14], [32] waren also bereits durch den Versuchsaufbau weitgehend ausgeschlossen. Und drittens könnte man auch argumentieren, dass die Schulung doch einen gewissen Effekt auf die Prüfergenauigkeit hatte: Die Strenge der Gesamtnote der geschulten Prüfer war deutlich näher an derjenigen der Nachprüfer. Demnach war die Notengebung trotz der individuellen Ungenauigkeit im Ganzen zutreffender. Die geschulten Prüfer waren also eher "weniger nachgiebig" als "strenger". Diesen Effekt hatten bereits Holmboe und Mitarbeiter beobachtet [20]. Das würde bedeuten, dass die Schulung doch den Bezugsrahmen der Prüfer vereinheitlicht hat, indem es sie in die Lage versetzt hat, eine angemessenere Notenskala zu verwenden. Die Individualität der Informationsverarbeitung durch die Prüfer und der Konversion ihrer Beobachtung und Beurteilung in eine Schulnote [33] blieb davon jedoch unberührt.

Die ungeschulten Prüfer glaubten weniger an die Möglichkeit eines Halo-Effektes als die geschulten. Dies könnte durchaus ein Schulungseffekt sein und impliziert, dass eine Prüferschulung für kognitive Verzerrungen sensibilisieren kann. Darüber hinaus fühlten sich Prüflinge von geschulten Prüfern eher ungerecht behandelt, da sie häufiger angaben, ihre Benotung sei nicht adäquat gewesen. Dies kann leicht durch die strengere Benotung erklärt werden. Dennoch war die Konkordanz zwischen Selbsteinschätzung und gegebener Note in beiden Gruppen gleich groß. Die Prüflinge mussten sich allerdings selbst einschätzen, kurz nachdem sie ihre Note erhalten hatten. Demnach könnte trotz aller Unzufriedenheit die selbst gegebene Note noch stark von der Note des Prüfers beeinflusst worden sein.

Zusammenfassend bezog sich unsere Studie auf ein curriculäres Testat mit einer äußerst spezifischen Aufgabe, einer kurzen und stark standardisierten körperlichen Untersuchung. Die Prüferschulung hatte keinen Einfluss auf die individuelle Genauigkeit der Benotung. Allerdings lag die Strenge der Bewertungen durch die geschulten Prüfer näher an den Bewertungen der Nachbeobachter als das bei den ungeschulten Prüfern der Fall war. Zudem waren die geschulten Prüfer sich des Halo-Effekts eher bewusst und ihre Prüflinge waren mit ihrer eigenen Benotung häufiger unzufrieden. Die Ergebnisse weisen darauf hin, dass die geschilderte Prüferschulung zwar einen gewissen Effekt hatte, dass aber die außerordentliche Individualität der Urteilsbildung bei der Bewertung komplexer medizinischer Fertigkeiten zu stark ist, um von einer einzigen Schulung beeinflusst zu werden. Der Aufwand einer regulären Prüferschulung zur Verbesserung der Fairness von Testaten dürfte sich daher kaum lohnen. Vielmehr sollten die Beurteilungen von medizinischen Fertigkeiten mit Vorsicht bewertet werden.

Danksagung

Die Autoren sind insbesondere Prof. Dr. Jana Jünger und Dr. Andreas Möltner vom Kompetenzzentrum für Prüfungen in der Medizin, Baden-Württemberg, zu Dank verpflichtet für ihre Beratung in der Planungsphase und bei der statistischen Auswertung sowie für die Zurverfügungstellung der Videokameras. Außerdem möchten wir Sebastian Sosnowki und Christopher Beck für die Assistenz bei den Filmaufnahmen und die Auswertung der Filme herzlich danken.

Interessenkonflikt

Die Autoren erklären, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Artikel haben.

Literatur

1. Horwitz RI, Kassirer JP, Holmboe ES, Humphrey HJ, Verghese A, Croft C, Kwok M, Loscalzo J. Internal medicine residency redesign: proposal of the Internal Medicine Working Group. Am J Med. 2011;124(9):806-812. DOI: 10.1016/j.amjmed.2011.03.007
2. Clark D, III, Ahmed MI, Dell'italia LJ, Fan P, McGiffin DC. An argument for reviving the disappearing skill of cardiac auscultation. Cleve Clin J Med. 2012;79(8):536-537, 544. DOI: 10.3949/ccjm.79a.12001
3. Smith MA, Burton WB, Mackay M. Development, impact, and measurement of enhanced physical diagnosis skills. Adv Health Sci Educ Theory Pract. 2009;14(4):547-556. DOI: 10.1007/s10459-008-9137-z
4. Ramani S, Ring BN, Lowe R, Hunter D. A pilot study assessing knowledge of clinical signs and physical examination skills in incoming medicine residents. J Grad Med Educ. 2010;2(2):232-235. DOI: 10.4300/JGME-D-09-00107.1
5. Alexander EK. Perspective: moving students beyond an organ-based approach when teaching medical interviewing and physical examination skills. Acad Med. 2008;83(10):906-909. DOI: 10.1097/ACM.0b013e318184f2e5
6. Ainsworth MA, Rogers LP, Markus JF, Dorsey NK, Blackwell TA, Petrusa ER. Standardized patient encounters. A method for teaching and evaluation. JAMA. 1991;266(10):1390-1396. DOI: 10.1001/jama.1991.03470100082037
7. Barley GE, Fisher J, Dwinnell B, White K. Teaching foundational physical examination skills: study results comparing lay teaching associates and physician instructors. Acad Med. 2006;81(10 Suppl):S95-S97. DOI: 10.1097/00001888-200610001-00024
8. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med. 2003;138(6):476-481. DOI: 10.7326/0003-4819-138-6-200303180-00012
9. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 2004;38(2):199-203. DOI: 10.1111/j.1365-2923.2004.01755.x
10. Pelgrim EA, Kramer AW, Mokkink HG, van den EL, Grol RP, van der Vleuten CP. In-training assessment using direct observation of single-patient encounters: a literature review. Adv Health Sci Educ Theory Pract. 2011;16(1):131-142. DOI: 10.1007/s10459-010-9235-6
11. Chen W, Liao SC, Tsai CH, Huang CC, Lin CC, Tsai CH. Clinical skills in final-year medical students: the relationship between self-reported confidence and direct observation by faculty or residents. Ann Acad Med Singapore. 2008;37(1):3-8.
12. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. Ann Intern Med. 1998;129(1):42-48. DOI: 10.7326/0003-4819-129-1-199807010-00011
13. Noel GL, Herbers JE Jr, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? Ann Intern Med. 1992;117(9):757-765. DOI: 10.7326/0003-4819-117-9-757
14. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. Med Educ. 2011;45(10):1048-1060. DOI: 10.1111/j.1365-2923.2011.04025.x
15. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently : Mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract. 2013;18(3):325-341. DOI: 10.1007/s10459-012-9372-1
16. Woehr DJ. Rater training for performance appraisal: a quantitative review. J Occup Organ Psychol. 1994;67:189-205. DOI: 10.1111/j.2044-8325.1994.tb00562.x
17. Lievens F. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. J Appl Psychol. 2001;86(2):255-264. DOI: 10.1037/0021-9010.86.2.255
18. Gorman CA, Rentsch JR. Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. J Appl Psychol. 2009;94(5):1336-1344. DOI: 10.1037/a0016476
19. Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. Med Educ. 1980;14(5):345-349. DOI: 10.1111/j.1365-2923.1980.tb02379.x
20. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. Ann Intern Med. 2004;140(11):874-881. DOI: 10.7326/0003-4819-140-11-200406010-00008
21. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med. 2009;24(1):74-79. DOI: 10.1007/s11606-008-0842-3
22. Vivekananda-Schmidt P, Lewis M, Coady D, Morley C, Kay L, Walker D, Hassell AB. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. Arthritis Rheum. 2007;57(5):869-876. DOI: 10.1002/art.22763
23. Newble DI, Hoare J, Elmslie RG. The validity and reliability of a new examination of the clinical competence of medical students. Med Educ. 1981;15(1):46-52. DOI: 10.1111/j.1365-2923.1981.tb02315.x
24. Sturpe DA, Huynh D, Haines ST. Scoring objective structured clinical examinations using video monitors or video recordings. Am J Pharm Educ. 2010;74(3):44. DOI: 10.5688/aj740344
25. Klimoski R, Inks L. Accountability forces in performance appraisal. Organ Behav Hum Decis Proc. 1990;45:194-208. DOI: 10.1016/0749-5978(90)90011-W
26. Martin JA, Reznick RK, Rothman A, Tamblyn RM, Regehr G. Who should rate candidates in an objective structured clinical examination? Acad Med. 1996;71(2):170-175. DOI: 10.1097/00001888-199602000-00025
27. Ogden GR, Green M, Ker JS. The use of interprofessional peer examiners in an objective structured clinical examination: can dental students act as examiners? Br Dent J. 2000;189(3):160-164.
28. Chenot JF, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B, Stanske B, Kochen MM, Himmel W. Can student tutors act as examiners in an objective structured clinical examination? Med Educ. 2007;41(11):1032-1038. DOI: 10.1111/j.1365-2923.2007.02895.x
29. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Intern Med. 1992;7(5):506-510. DOI: 10.1007/BF02599454
30. Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. Acad Med. 2010;85(10 Suppl):S25-S28. DOI: 10.1097/ACM.0b013e3181ed1aa3
31. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ. 2006;6:42. DOI: 10.1186/1472-6920-6-42

32. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003;15(4):270-292. DOI: 10.1207/S15328015TLM1504_11

33. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 2011;86(10 Suppl):S1-S7. DOI: 10.1097/ACM.0b013e31822a6cf8

Korrespondenzadresse:
PD Dr. med. Gunther Weitz, MME
Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Medizinische Klinik I, Ratzeburger Allee 160, 23538 Lübeck, Deutschland, Tel.: +49 (0)451/500-6033, Fax: +49 (0)451/500-6242
gunther.weitz@uksh.de

Bitte zitieren als
Weitz G, Vinzentius C, Twesten C, Lehnert J, Bonnemeier H, König IR. Effects of a rater training on rating accuracy in a physical examination skills assessment . GMS Z Med Ausbild. 2014;31(4):Doc41. DOI: 10.3205/zma000933, URN: urn:nbn:de:0183-zma0009338

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2014-31/zma000933.shtml

Eingereicht: *08.01.2014*
Überarbeitet: *24.03.2014*
Angenommen: *20.08.2014*
Veröffentlicht: *17.11.2014*