Piloting an outcome-based programme evaluation tool in undergraduate medical education

Abstract

Aims: Different approaches to performance-oriented allocation of resources according to teaching quality are currently being discussed within German medical schools. The implementation of these programmes is impeded by a lack of valid criteria to measure teaching quality. An assessment of teaching quality should include structural and procedural aspects but focus on learning outcome itself. The aim of this study was to implement a novel, outcome-based evaluation tool within the clinical phase of a medical curriculum and address differences between the novel tool and traditional evaluation methods.

Methods: Student self-assessments before and after completion of a teaching module were used to compute performance gains for specific learning objectives. Mean performance gains in each module were compared to student expectations before the module and data derived from a traditional evaluation tool using overall course ratings at the end of the module.

Results: A ranking of the 21 modules according to computed performance gains yielded entirely different results than module rankings based on overall course ratings. There was no significant correlation between performance gain and overall ratings. However, the latter were significantly correlated to student expectations before entering the module as well as structural and procedural parameters (Pearson’s r 0.7-0.9). 

Conclusion: Performance gain computed from comparative self-assessments adds an important new dimension to course evaluation in medical education. In contrast to overall course ratings, the novel tool is less heavily confounded by construct-irrelevant factors. Thus, it appears to be more appropriate than overall course ratings in determining teaching quality and developing algorithms to guide performance-oriented resource allocation in medical education.

Keywords: evaluation, self-assessment, performance, learning objective, congruence, clinical studies

Introduction

Following suggestions made by the German Council of Science and Humanities [1], half of all German medical schools internally allocated resources based on teaching performance in 2009. Such performance measures are also increasingly used to distribute funds to different medical schools within German federal states. In North Rhine-Westphalia, indicators used for this purpose can be attributed to three levels [2]: structural, procedural and outcome quality of teaching. Upon introduction of their algorithm, medical school Deans in North Rhine-Westphalia acknowledged the importance of structures (resources, time-tables) and processes (instructional format, examinations) but also suggested to include global performance indicators such as student results in high-stakes examinations, study time and drop-out rates.
as measures of study outcome. Within the clinical phase of undergraduate medical education, these parameters cannot be broken down to clinical specialties and are thus of little use to guide performance-dependent resource allocation within a particular medical school. Consequently, there is need to develop quality indicators referring to structures, processes and teaching outcome that can be used to rank courses or specialties.

A survey among German medical schools carried out in 2009 showed that many schools used evaluation data (i.e., student ratings) to guide internal fund distribution [3]. Traditional evaluation forms are predominantly made up of questions regarding structural and organisational aspects of teaching as well as global ratings on six-point scales resembling those used in German schools (with 1 being the best mark). While these instruments likely allow some appraisal of structures and processes, the extent to which they actually measure outcome quality is unknown.

The definition of desired outcomes in undergraduate medical education is a matter of debate; however, it is reasonable to suggest student performance gain as a possible primary outcome variable. Ideally, an evaluation tool would provide some measure of performance gain. Such a tool has been developed at Göttingen Medical School, and its reproducibility and criterion validity have recently been established [4]. By introducing ‘learning success’ as an outcome variable, it adds to traditional evaluation methods focusing on structural and procedural aspects of teaching quality.

In order to establish discriminant validity of the novel evaluation tool this study aimed at answering the following research questions:

1. Is there a difference between course rankings derived from performance gain data and traditional evaluation parameters? We hypothesized that the two approaches cover different aspects of teaching. As a consequence, rankings based on structural/procedural criteria should be considerably different from rankings based on performance gain.

2. Is there a correlation between student expectations towards specific courses, traditional evaluation parameters obtained at the end of courses and performance gain data? We hypothesized that traditional evaluation parameters would be highly correlated with each other. Since structural and procedural aspects considerably impact on teaching quality, we expected moderate correlations between the traditional and the novel evaluation tool.

Methods

Clinical curriculum and evaluation tools at Göttingen Medical School

The three-year clinical curriculum at our institution adopts a modular structure. There are 21 modules lasting two to seven weeks. The first clinical year covers basic practical skills, infectiology and pharmacotherapy, followed by 18 months of systematic teaching on the diagnosis and treatment of specific diseases. Aspects regarding differential diagnosis are focused on in the last 6 months.

1. Traditional evaluation tool: Students complete online evaluation forms (EvaSys®, Electric Paper, Lüneburg, Germany) at the end of each module. Ratings on organisational and structural aspects of teaching are obtained on six-point scales. The following statements are used for these ratings:

   1.1. “The implementation of interdisciplinary teaching was well done in this module.” (Interdisciplinarity)

   1.2. “During this module, self-directed learning was promoted.” (Self-directed learning)

   1.3. “Regarding my future professional life, I perceive my performance gain in this module as high.” (Subjective performance gain)

   1.4. “I was very satisfied with the basic structure of this module (design, instructional formats, timetables).” (Structure)

   1.5. “This module should be continued unchanged.” (Continuation)

2. The final item of this part of the questionnaire read “Please provide an overall school mark rating of the module.”

3. In order to assess correlations between student expectations before a module and their ratings at the end of a module, an additional online survey at the beginning of each module was set up specifically for this study. In this survey, students were asked to rate the following statements on a six-point scale:

   3.1. “I believe that this module is important for my future professional life.” (Importance)

   3.2. “I am looking forward to this module.” (Anticipation)

   3.3. “The module has a good reputation with my more advanced fellow students.” (Reputation)

4. Performance gain evaluation using comparative self-assessments (CSA): The novel tool used in this study calculates performance gain for specific learning objectives on the basis of repeated student self-assessments. At the beginning of each module, students are asked in an online survey to self-rate their performance levels regarding specific learning objectives (e.g., “I can interpret an electrocardiogram.”) on a six-point scale anchored by “fully agree” (1) and “completely disagree” (6). Self-ratings were again obtained at the end of each module. The percent CSA gain for a specific learning objective was calculated by dividing the difference of mean values (pre-post) by the corrected mean of initial self-ratings across student cohorts according to the following formula (see figure 1):
Aggregated CSA gain of a module was calculated as the mean of CSA gains obtained for 15 specific learning objectives which were derived from the Göttingen Catalogue of Specific Learning Objectives (http://www.med.uni-goettingen.de/de/media/G1-2_lehre/lernzielkatalog.pdf). The major difference between the aforementioned evaluation item “Regarding my future professional life, I perceive my performance gain in this module as high.” and CSA gain was that subjective performance gain was obtained at one time-point only and was global in nature while CSA gain was calculated from ratings obtained at different time-points and related to specific learning objectives. A recently published study demonstrates a good correlation between CSA gain and an increase in objective performance measures [4].

Description of the study sample

In winter 2008/09, a total of 977 students were enrolled in clinical modules at our institution. All students were invited to participate in online evaluations via e-mail; each student automatically received three e-mails per survey containing a link to the online evaluation platform. Pre-surveys were opened three days before the beginning of a module and closed three days into the module. The same time-frame was used for post-surveys. Participation was voluntary, and all data were entered anonymously. Thus, further characterisation of the sample with regard to age and sex was not possible.

Data acquisition and analysis

Anonymous evaluation data collected during winter term 2008/09 were included in this study. During completion of the post-survey, students were asked to indicate whether they had also taken part in the pre-survey. Pre-post comparisons were only performed on data obtained from students who indicated to have completed both the pre- and the post-survey. Since data were aggregated across student cohorts, no individual labelling of students was necessary.

In order to address the first research question, the 21 modules were ranked according to mean global ratings or mean CSA gain values (aggregated from 15 specific learning objective gains). The second research question was addressed by calculating correlations between mean values of traditional evaluation items and aggregated CSA gain per module.

Analyses were run with SPSS® 14.0 (Illinois, USA). A Kolmogorov-Smirnov Test indicated that all data were normally distributed. Thus, correlations are reported as Pearson’s r. Squared r values indicate the proportion of variance explained.

Results

Response rates

Of all 977 students enrolled in clinical modules, 573 provided a total of 51,915 ratings. Response rates in specific modules ranged from 36.7% to 75.4%.

Comparison of module rankings

Table 1 compares global student ratings and CSA gain values at the level of individual modules. Modules were ranked using both indicators (see the last two columns of the table). In some instances, ranks differed substantially depending on the method used. A difference of at least 6 ranks between the two methods was found for six out of 21 modules. For example, module no. 19 received the best overall rating (1.56) despite taking position 17 in the CSA gain ranking (55.6%). Conversely, module no. 7 (“Evidence-based medicine”) received a mean overall rating of 3.67 and thus came last in this ranking while mean CSA gain was 64.3% (Rank 9). Notably, students had already provided negative ratings for this module in the pre-survey (Importance: 3.64; anticipation: 4.20; Reputation: 4.59).

Correlations between evaluation parameters

Results of correlation analyses are presented in Table 2. Student expectations before a module and ratings of organisational and structural aspects obtained after a module showed a strong positive correlation. There was also a strong correlation between overall ratings and approval of the statement “Regarding my future professional life, I perceive my performance gain in this module as high.” (r=0.94; p<0.001). In contrast, there was no significant correlation between CSA gain and student expectations in the pre-survey. Weak correlations were observed between CSA gain and two variables included in the traditional evaluation form (maximum proportion of variance explained: 22%). CSA gain was weakly correlated with the subjective perception of having learned a lot (r=-0.44; p=0.044; proportion of variance explained 19%) and non-significantly with overall module ratings (r=-0.42; p=0.061; proportion of variance explained 18%; see Figure 2).
Table 1: Overall ratings, mean CSA gain and rankings of all 21 clinical modules, listed in their order of appearance in the three-year clinical curriculum. CSA, comparative self-assessment. Each term lasts 6 months.

<table>
<thead>
<tr>
<th>Module number (term)</th>
<th>Overall rating*</th>
<th>CSA gain (%)</th>
<th>Module ranking by **</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Term 1)</td>
<td>2.09</td>
<td>67.7%</td>
<td>7</td>
</tr>
<tr>
<td>2 (Term 1)</td>
<td>2.70</td>
<td>63.0%</td>
<td>14</td>
</tr>
<tr>
<td>3 (Term 1)</td>
<td>2.82</td>
<td>56.7%</td>
<td>17</td>
</tr>
<tr>
<td>4 (Term 2)</td>
<td>1.84</td>
<td>72.3%</td>
<td>3</td>
</tr>
<tr>
<td>5 (Term 2)</td>
<td>2.35</td>
<td>55.5%</td>
<td>12</td>
</tr>
<tr>
<td>6 (Term 2)</td>
<td>2.06</td>
<td>56.6%</td>
<td>6</td>
</tr>
<tr>
<td>7 (Term 2)</td>
<td>3.67</td>
<td>64.5%</td>
<td>21</td>
</tr>
<tr>
<td>8 (Term 3)</td>
<td>1.59</td>
<td>71.2%</td>
<td>2</td>
</tr>
<tr>
<td>9 (Term 3)</td>
<td>3.34</td>
<td>56.5%</td>
<td>20</td>
</tr>
<tr>
<td>10 (Term 3)</td>
<td>2.13</td>
<td>69.5%</td>
<td>8</td>
</tr>
<tr>
<td>11 (Term 4)</td>
<td>2.73</td>
<td>64.1%</td>
<td>15</td>
</tr>
<tr>
<td>12 (Term 4)</td>
<td>2.70</td>
<td>49.2%</td>
<td>13</td>
</tr>
<tr>
<td>13 (Term 4)</td>
<td>2.02</td>
<td>70.6%</td>
<td>5</td>
</tr>
<tr>
<td>14 (Term 4)</td>
<td>2.23</td>
<td>64.9%</td>
<td>9</td>
</tr>
<tr>
<td>15 (Term 5)</td>
<td>1.99</td>
<td>70.6%</td>
<td>4</td>
</tr>
<tr>
<td>16 (Term 5)</td>
<td>2.24</td>
<td>63.2%</td>
<td>10</td>
</tr>
<tr>
<td>17 (Term 5)</td>
<td>2.32</td>
<td>70.9%</td>
<td>11</td>
</tr>
<tr>
<td>18 (Term 6)</td>
<td>2.73</td>
<td>56.8%</td>
<td>16</td>
</tr>
<tr>
<td>19 (Term 6)</td>
<td>1.56</td>
<td>55.6%</td>
<td>1</td>
</tr>
<tr>
<td>20 (Term 6)</td>
<td>2.88</td>
<td>11.5%</td>
<td>18</td>
</tr>
<tr>
<td>21 (Term 6)</td>
<td>3.19</td>
<td>44.5%</td>
<td>19</td>
</tr>
</tbody>
</table>

* six-point scale with 1 being the best option; ** Differences of >5 ranks are printed in bold letters

Table 2: Correlations between student ratings of the 21 modules and CSA gain (aggregated across 15 learning objectives per module).

<table>
<thead>
<tr>
<th>Time-point of data collection</th>
<th>before the module</th>
<th>after the module</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance</td>
<td>Importance</td>
<td>Importance</td>
</tr>
<tr>
<td>Anticipation</td>
<td>Anticipation</td>
<td>Anticipation</td>
</tr>
<tr>
<td>Reputation</td>
<td>Reputation</td>
<td>Reputation</td>
</tr>
<tr>
<td>Interdisciplinarity</td>
<td>Interdisciplinarity</td>
<td>Interdisciplinarity</td>
</tr>
<tr>
<td>Self-directed learning</td>
<td>Self-directed learning</td>
<td>Self-directed learning</td>
</tr>
<tr>
<td>Subjective performance gain</td>
<td>Subjective performance gain</td>
<td>Subjective performance gain</td>
</tr>
<tr>
<td>Structure</td>
<td>Structure</td>
<td>Structure</td>
</tr>
<tr>
<td>CSA gain</td>
<td>CSA gain</td>
<td>CSA gain</td>
</tr>
</tbody>
</table>

Legend: Cells contain Pearson’s correlation coefficients; *p < 0.05; **p < 0.001; CSA, comparative self-assessment
Discussion

Principal findings

In this study, we observed significant correlations between overall module ratings and student expectations before a module as well as with retrospective ratings of curricular structure and subjective performance gain. Conversely, there was no correlation between CSA gain and student expectations and only a weak correlation between CSA gain and perceived subjective performance gain. Accordingly, module rankings were different depending on them being based on overall ratings or CSA gain values.

Strengths and limitations

The new evaluation tool presented here facilitates a critical appraisal of performance gain by using comparative student self-assessments. It is easy to implement and addresses specific learning objectives. The data used for this study were obtained as part of our evaluation routine and comprised a large number of single ratings provided by over 500 students.

The validity of singular self-assessments has repeatedly been criticised in the past [5], [6] as their accuracy can be influenced by factors which are irrelevant to the construct examined. However, Colthart et al. [7] demonstrated that the ability to self-assess can be improved by using well-defined criteria, anchoring of scales and feedback. Accordingly, using specific learning objectives [8] for self-assessments is a prerequisite for the new tool to be functional. The potential impact of individual characteristics on CSA gain values is reduced by repeatedly collecting self-assessments from the same student group.

Relation to published research and significance of findings

Teaching quality is a multi-dimensional construct including several parameters related to process, structure, content and outcome of teaching [9]. Evaluation data provided by students contribute to the critical appraisal of teaching quality within medical schools. However, subjective ratings are influenced by various variables. As a consequence, details of the construct underlying student evaluation data may not be readily discernible. This particularly pertains to overall course ratings which are usually captured using a simple marking system. This study demonstrates that overall module ratings are strongly correlated with student expectations before taking a module as well as retrospective ratings of curricular structure and subjective performance gain (see Table 2). There are two possible explanations for this finding which may not be mutually exclusive: Student ratings may either feed into a construct of ‘good teaching’ which includes well-structured, interdisciplinary teaching that fosters self-directed learning and produces a considerable performance gain. Alternatively, it may be suggested that all seemingly different parameters used in traditional evaluation forms feed into the same, homogeneous construct within which specific aspects cannot be differentiated. In an extreme case, this would mean measuring mere student satisfaction which – as indicated by the correlations seen in this study – is strongly associated with student expectations before a module.

In fact, previous research indicates that overall ratings are not only influenced by structural and procedural aspects [10], [11], [12], [13], [14] but also by the behaviour of faculty [15], their rapport with students [16] and a lecturer’s or a course’s reputation [17]. While professional bearing of faculty might be attributed to a construct of ‘good teaching’, interpretation of student overall ratings is complicated by the extent of such contributions to the composite mark being largely unknown.

In addition to addressing structural and procedural parameters, a critical appraisal of teaching should take student learning outcome into consideration [9]. At first glance, student achievements in end-of-course assessments or high-stakes examinations might appear helpful in this regard. However, it has recently been reported that the assessments performed at many German medical schools do not live up to international quality standards [18] and thus cannot be regarded as being sufficiently valid. In addition, multiple choice questions address but one dimension of physician training (knowledge) and do not allow any conclusions to be drawn on the quality of practical training. Attributing exam results to teaching quality in specific courses or even specialties can be particularly difficult in reformed curricula. Thus, internal fund redistribution based on examination results does not seem advisable.
Practical implications

A clear definition of the construct of ‘good teaching’ and proof of validity for the tools used to assess are prerequisites for the interpretation of evaluation data. Accordingly, evaluation tools capturing specific aspects of good teaching and avoiding cross-correlations between different items are desirable. As correlations between overall ratings and CSA gain were either absent or weak in this study, the novel evaluation tool described here is likely to produce additional information that would have been missed by traditional evaluation tools. It might thus be used for a valid appraisal of teaching quality.

Research agenda

We did not assess students’ definitions of performance gain in this study; this should be done in future studies, including research into teachers’ definitions of learning outcome which may be different from students’ views. Future use of the novel evaluation tool needs to be informed by research regarding the number of learning objectives that need to be included per module in order to cover the whole range of aspects addressed in a module and, thus, produce reliable results. In addition, potential biases arising from the online data acquisition tool [19] and differing response rates across modules and their influence on CSA gain data needs to be studied in more detail.

The novel tool necessitates data collection from identical student groups at different time-points, thus potentially raising concerns regarding data protection issues as students should be able to complete evaluation forms anonymously. In order to solve this problem, the pre-/post-design could be changed to a singular data collection period at the end of each module during which students would provide retrospective ratings of their performance levels at the beginning of the module [20]. Finally, future research needs to show to what extent the novel tool can be transferred to other medical schools and curricula.

Conclusions

Overall course ratings provided by students do not address all aspects of teaching quality and are of limited use to guide fund redistribution within medical schools. In contrast, by providing an estimate of actual performance gain, the tool described here appears to cover an important dimension of teaching quality. It was found to be valid and independent of traditional global ratings. Therefore, CSA gain might become an integral part of algorithms to inform performance-guided resource allocation within medical schools.

Note

The authors Raupach and Schiekirka have equally contributed to this manuscript.

Competing interests

The authors declare that they have no competing interests.

References


Corresponding author:
PD Dr. med. Tobias Raupach, MME-D
University Medical Centre Göttingen, Dept. of Cardiology & Pneumology, 37099 Göttingen, Germany, Phone +49 (0)551/39-8922, Fax: +49 (0)551/39-6887
raupach@med.uni-goettingen.de

Please cite as

This article is freely available from http://www.egms.de/en/journals/zma/2012-29/zma000814.shtml

Received: 2011-07-08
Revised: 2011-12-12
Accepted: 2012-01-12
Published: 2012-05-15

Copyright
©2012 Raupach et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en). You are free: to Share — to copy, distribute and transmit the work, provided the original author and source are credited.