User manual for the CSA gain evaluation tool

Author: Prof. Dr. Tobias Raupach
Co-Author: Dipl.-Psych. Sarah Schiekirka

Content

1. Introduction

This user manual provides an introduction to a novel Class Climate® feature facilitating estimation of student learning outcome from comparative self-assessments (CSA). This new feature can be used as an addition to existing questionnaires as it generates information that is generally not picked up by traditional evaluation tools. CSA evaluation data are unrelated to other dimensions of evaluation, e.g. student satisfaction. Instead, they help to identify specifics strengths and weaknesses of a curriculum on the level of specific learning objectives, thus closing the feedback loop of curricular development. The new feature was designed according to current advances in evaluation research in the context of medical education. A number of studies have assessed the reliability and validity of this approach towards outcome-based evaluation.

Following a brief outline of the scientific background of outcome-based evaluation, the CSA gain feature is being described using examples from undergraduate medical education and referring to recent findings on the reliability and validity of the tool. At the end of this manual, you will find step-by-step instructions on how to set up your own CSA gain evaluation. After reading this manual, you will be able to tailor the CSA feature to your evaluation needs, and you will know how to interpret the output file.

2. Scientific Background

One purpose of evaluations in higher education is to critically appraise teaching quality in modules, courses or even entire programs. In addition to providing formative or summative feedback to individual teachers and/or teaching co-ordinators, the ultimate goal of this process is to improve teaching quality. This begs the question of how ‘teaching quality’ be best defined. Reaching a consensus on a specific construct of high-quality teaching is essential as any evaluation tool aimed at measuring teaching quality needs to be aligned to this very construct. For example, if evaluators agree that teaching quality is mainly reflected by lecturing skills, these should be the main focus of the evaluation tool used. In turn, if designed along these lines, the resulting evaluation tool will only identify skilled lecturers as delivering high-quality teaching while it will not pick up excellent bed-side teaching resulting in favorable learning outcome despite the teacher not being an outstanding lecturer. Thus, if the construct of ‘high-quality teaching’ underlying an evaluation tool is unknown, evaluation results need to be interpreted with great caution.

According to Gibson and colleagues [1], there are four dimensions of teaching quality, each of which can (and probably should) be specifically addressed. On a curricular level, structural as well as procedural aspects of teaching impact on the quality of the learning experience. The third dimension relates to the expertise and skills of individual teachers while the fourth dimension is reflected in the outcome of teaching activities.

Structural aspects include teaching facilities and resources as well as the structure of the curriculum itself. The quality of the teaching process is reflected in the way students interact with each other. Thus, this dimension reflects the learning environment in a module, course or entire program. On the level of individual teachers, teaching expertise and thorough session preparation both contribute to excellent teaching. At the same time, many questionnaires assessing teacher performance also include questions regarding an individual's enthusiasm for teaching. Finally, teaching outcome may be apparent in successful student learning. Other definitions of teaching outcome include a culture change within higher education institutions towards valuing teaching just as much as research output.

The first three dimensions of teaching quality may be captured using student ratings, and a number of questionnaires have been developed for this purpose. While most tools were shown to produce reliable results, the validity of student ratings is limited owing to confounding from various sources: For example, enthusiastic teachers who are well known for their presentation skills [2] tend to receive more favorable ratings than others – even if their presentations are flawed [3]. Another important confounder of student ratings is student motivation: Students who are interested in the subject matter will provide more favorable ratings for a course than students who are less enthusiastic – all of this is independent of actual teaching quality [4].

Although student learning might be perceived as the most important effect of high-quality teaching, other outcome measures have been proposed as well. Blumberg [5] considers the development of study skills (including expertise in self-directed learning) and of a teaching culture within a higher education institution as important outcomes that may arise from high-quality teaching. However, in this user manual, outcome will be defined as student learning outcome in terms of professional competencies (i.e., 'performance gain'). In medicine (and potentially in most other higher education contexts as well), each learning objective relates to one out of three domains: factual knowledge (cognitive domain), skills (practical / psychomotor domain), and attitudes / professionalism (affective domain).

At first glance, assessing student learning outcome in the cognitive and practical domains might seem straightforward as student performance is usually captured in written, oral or practical summative exams. Unfortunately, despite being objective in most cases, exam results do not necessarily represent valid indicators of student learning. End-of-course exam results merely represent performance levels at one particular point in time. Due to the lack of information on initial student performance levels, they do not provide any information on the **gain** in knowledge, skills and attitudes that would have occurred as a result of teaching and learning activities **during** the course. A potential causal relationship between favorable exam results and high-quality teaching might be inferred, but other explanations for high performance need to be considered as well. For example, students might have mastered some of the learning objectives even before entering the course – in this scenario, high scores in a final exam merely represent good retention but not acquisition of new competencies. In order for exam data to be valid

indicators of learning outcome during a course, students would have to be tested at least twice (i.e., at the beginning and at the end of the course/module). This approach is likely to prove logistically challenging. However, even if course co-ordinators managed to implement repetitive objective testing, they might be faced with another problem in that some exams fail to adequately cover the full range of learning objectives addressed in a course. In addition, assessment formats need to be aligned to the type of learning objectives taught in a course [6]. This is not always the case. For example, medical students need to learn how to perform an arterial blood gas analysis. This would usually be taught in a lecture; the subsequent exam is most likely going to be a multiple choice test, the results of which do not allow any conclusions to be drawn on student skills in taking a blood sample and interpreting lab results. Practical learning objectives are not confined to medical education, so these considerations are relevant to non-medical subjects as well.

As stated above, exam questions need to cover the entire range of learning objectives relevant to a specific subject in order to provide a valid representation of the underlying construct. In the context of medical education, one example of such a construct is ‘a medical school graduate who is adequately prepared to take a full history, carry out a physical examination, derive a differential diagnosis, order and correctly interpret specific diagnostic tests as well as suggest an appropriate treatment’. If high-quality teaching is defined as helping students to achieve this goal, the evaluation tool used needs to address all of these aspects – and so should the exam if it is meant to be used as an evaluation tool. However, most exams fail to truly represent a complex construct as they do not cover the full range of learning objectives contributing to the construct, i.e., there is ‘construct under-representation’. This source of bias threatens the validity of exam data as exams highlighting only some but not all relevant aspects place students at an advantage who happen to be well-prepared for those aspects covered in the exam, regardless of their knowledge on aspects not covered in the exam. Another source of confounding is ‘construct-irrelevant variance’ which tends to occur if the wording of exam questions is confusing, thus disadvantaging students who might have excellent knowledge, skills and attitudes but are struggling with the language. In this case, variance in exam results reflects variance in language proficiency – clearly, this variance is irrelevant to the construct underlying the exam (unless the exam was specifically designed to address language fluency and intelligence).

Finally, student performance in exams is highly dependent on exam consequences. While formative exams provide feedback helping students to adjust their learning to individual strengths and weaknesses, they usually do not generate a huge incentive to learn. In contrast, summative (i.e., graded) exams are perceived as threatening as they can be failed so students spend considerably more time revising for them. A recent study demonstrated the substantial effect of exam consequences on student performance irrespective of the teaching formats used before the exam [7]. Thus, exam results do not necessarily reflect the quality of teaching but rather the effect of an upcoming exam on study behavior. While this might be a desired effect,

evaluators need to be clear about what they are evaluating (i.e., their construct of high-quality teaching).

In summary, one important dimension of teaching quality – student learning outcome – is not necessarily fully covered by exam results. An evaluation instrument assessing learning outcome in the cognitive, psychomotor and affective domains that is less prone to bias arising from post-hoc measurements, construct under-representation and construct-irrelevant variance would be desirable. Such an instrument has been developed and will be presented in the following section.

3. Description of an outcome-based evaluation tool developed and piloted at Göttingen University Medical Centre

In 2007, Göttingen University Medical Centre introduced a novel evaluation tool enabling educators to gauge student learning outcome with regard to all three domains of medical education (factual knowledge, skills, and attitudes). The raw data used for this approach are derived from student self-assessments. The following paragraph summarizes some research on the validity of self-ratings.

The validity of isolated self-assessments, i.e. self-ratings obtained from individual students at one single point in time, has been shown to be low to moderate [8, 9]. This is due to considerable inter-individual variance in the ability to provide realistic self-assessments. For instance, individuals with favorable self-esteem tend to make more positive attributions to themselves. As students, these individuals would be expected to provide more favorable self-ratings than their peers. Another confounding factor relevant to the area of self-ratings is the so-called 'above average effect': Studies performed in various contexts have shown that considerably more than 50% of the surveyed population believe that their abilities are above average – by definition, this is impossible. There is evidence of the above average effect being more pronounced in specific groups: In one study, 94% of higher education lecturers believed that their presentation skills exceeded those of 50% of their colleagues [10].

In contrast to isolated self-ratings, repeated (comparative) self-assessments yield higher validity as within any given individual the ability to self-assess is relatively stable over time [11]. Thus, repeated self-assessments can be used to determine any change that has occurred between data collection points. Accordingly, learning outcome may be measured by calculating the difference between student self-assessments obtained at the beginning and at the end of a course. Any tendency to over- or underestimate oneself should only have limited effects on results as that tendency can be expected to remain unchanged over time.

The novel evaluation tool is based on repeated student self-assessments for specific learning objectives. Obviously, the quality of the statements used for self-ratings is crucial to the validity

of this approach. Only specific learning objectives addressed in a particular module should be used to evaluate learning outcome of that module. General statements such as 'I am a good doctor.' are unlikely to generate useful information on the level of specific learning objectives. Instead, educators should frame these statements in a way that will allow them to identify specific weaknesses of a curriculum so that outcome-based evaluation feeds directly into curriculum development activities. For example, if comparative student self-ratings of the statement 'I can list all risk factors for the development of coronary artery disease.' suggest that learning outcome on this specific subject was suboptimal, teachers will know exactly which part of the curriculum needs to be improved. Writing up self-rating statements according to the SMART criteria is the most time-consuming step in the development of outcome-based evaluation tools, and a detailed description of how to best approach this task is provided at the end of this user manual.

Once relevant learning objectives have been identified and related first-person statements have been drafted, students will be asked to self-rate their knowledge, skills and attitudes on a six-point scale. As this tool was developed in Germany, the scale resembles the marking system in German schools. Thus, it is anchored at 1 ('fully agree') and 6 ('completely disagree'). Anchors are not displayed in Class Climate® so this is only relevant for data interpretation. Different anchors may be used, and the formulas given below might need to be adjusted accordingly. However, in order to be able to compare results to our previous reports, the most positive response (whether this translates into a 1 or a 6) should be kept on the left side of the scale. This detail was taken care of in the design of the new Class Climate® feature. As described, students need to provide self-ratings for the same set of statements at the beginning and at the end of a module/course.

Data analysis is performed on the level of the entire student cohort (rather than on the level of individual students). A study by Lam demonstrated how the unit of analysis impacts on evaluation results [12]. Similarly, we found correlations between subjective and objective performance gain to be low to moderate when analyzed on an individual level while the correlation was much stronger when data were aggregated across the student cohort [13]. (As yet, the minimum sample size needed for this approach still needs to be determined. As a rule of thumb, any student group of at least 20 students should be large enough to produce meaningful results).

Learning outcome within a student cohort is reflected by the change in mean values between the two data collection points. However, as more advanced students most probably know more than their less advanced peers, their initial self-ratings will be more favorable, thus reducing the scope for absolute numerical change on a six-point scale. If left unadjusted, this is likely to result in learning outcome to be underestimated for advanced students compared to novices in whom learning outcome will be overestimated. Adjustment for initial performance levels can be achieved by dividing the difference in mean self-ratings by mean ratings obtained before

exposure to the module/course ('pre-ratings'). As students progress, these pre-ratings will improve (i.e. numerical values will decrease owing to the fact that 1 is the most favorable option), and any absolute pre-post change will be divided by a smaller value. At the same time, novice students are more likely to provide less favorable pre-ratings so that any absolute pre-post change will be attenuated by dividing it by the pre-rating (which is numerically higher). Student learning outcome is calculated using the following formula:

$$\text{CSA gain [\%]} = \frac{\mu_{pre} - \mu_{post}}{\mu_{pre} - 1} \times 100$$

Formula 1: CSA = comparative self-assessment; µpre = mean student self-ratings before exposure to teaching; µpost = mean student self-ratings after exposure to teaching; the correction term '-1' in the denominator reflects the fact that the six-point scale is anchored by 1 and not 0.

This formula produces values between -100% and +100%. A CSA gain of 100% requires all students to tick the most favorable scale option at the end of the module/course. The following table illustrates how adjusting for initial performance levels impacts on CSA gain values.

Pre-rating	Post-rating	Absolute difference	Adjusted pre-rating	CSA gain
5	3	2	4	2/4 = 50%
4	2.5	1.5	3	1.5/3 = 50%
2	1.5	0.5	1	0.5/1 = 50%
3	1	2	2	2/2 = 100%

If students provide less favorable self-ratings after a module than at the beginning of a module, learning outcome will be negative. This is likely to occur if students were confused by the way content was presented or if self-rating statements were unrelated to what has actually been taught in a module (both of which would be significant findings for teaching co-ordinators). If the numerator of the above formula is <0, dividing this difference by the adjusted pre-rating produces disproportionally low CSA gain values. For this reason, Class Climate® automatically uses a separate formula in the case of negative pre-post differences:

$$\text{CSA gain [\%]} = \frac{\mu_{pre} - \mu_{post}}{6 - \mu_{pre}} \times 100$$

Scantron Corporation – 1313 Lone Oak Road – Eagan, MN 55121 1-800-722-6876

A number of studies addressed the feasibility, reliability and validity of this novel outcome-based evaluation tool [13-16]. As this approach was developed in the context of medical education, additional research is needed to confirm that it works just as well in other higher education contexts. For the time being, we can only draw on the results of studies run at Göttingen University Medical Centre.

a) Feasibility

As a first step, the novel tool was implemented across the entire clinical curriculum at Göttingen Medical School. This three-year study phase is preceded by a two-year pre-clinical phase and is followed by one year of elective periods. The three-year clinical phase is made up of 21 consecutive modules of varying duration (between two and seven weeks). For each module, 15 self-rating statements referring to the three domains of medical education (knowledge, skills and attitudes) were derived from the institution's Catalogue of Specific Learning Objectives. CSA gain data for selected learning objectives taught in different modules are presented in Figure 1. Results demonstrate that CSA gain can be used to identify learning objectives with high, moderate or low performance gain in each of the three educational domains.

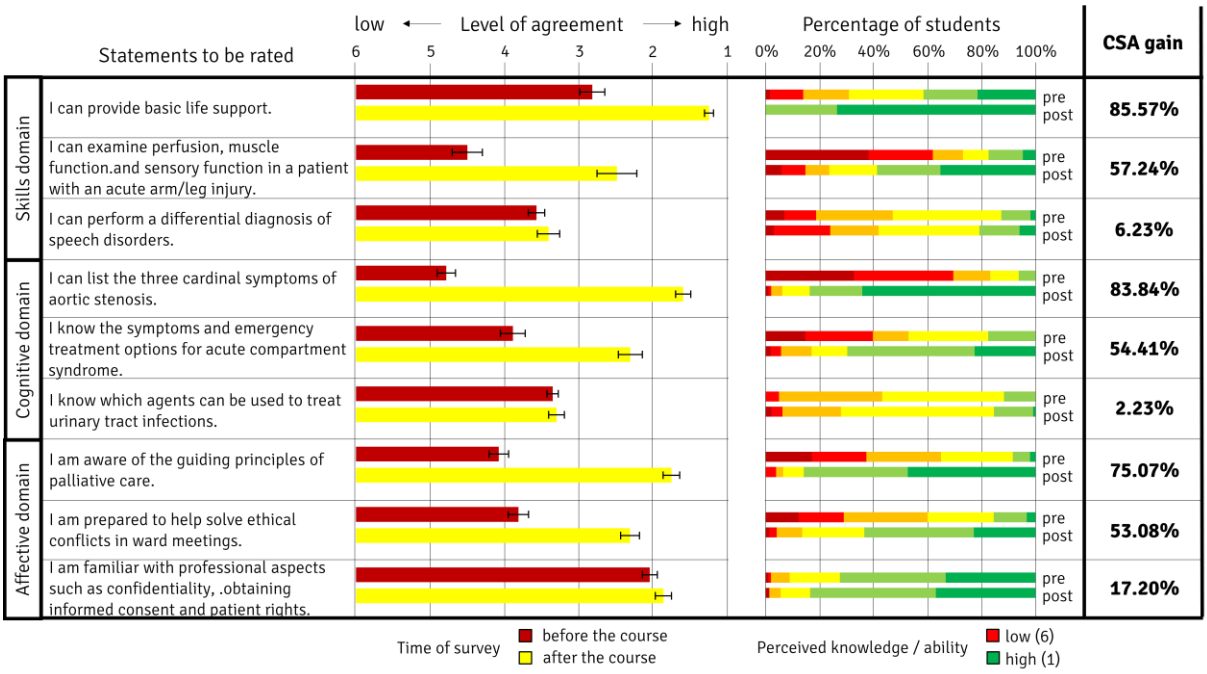

Figure 1: CSA gain report for module co-ordinators. Columns on the left-hand side represent mean student self-ratings before (red) and after (yellow) exposure to teaching. On the right-hand side, student self-assessments are broken down by scale options. CSA gain is given in the final column. For each domain of medical education, there were learning objectives with high, moderate and low performance gain. Modified from [14].

b) Reliability

Figure 2 presents CSA gain for the same set of learning objectives addressed in a six-week module on cardiovascular and respiratory disease in two consecutive student cohorts. Teaching was identical for both cohorts so learning outcome was expected to be similar. In fact, CSA gain data for 14 learning objectives were highly correlated (Pearson's r = 0.98), indicating favorable reliability of the evaluation tool. However, there was one 'outlier' (see arrow): In winter 2008/09, performance gain for this objective was considerably higher than in the preceding term. A potential reason for this finding will be discussed in the next paragraph.

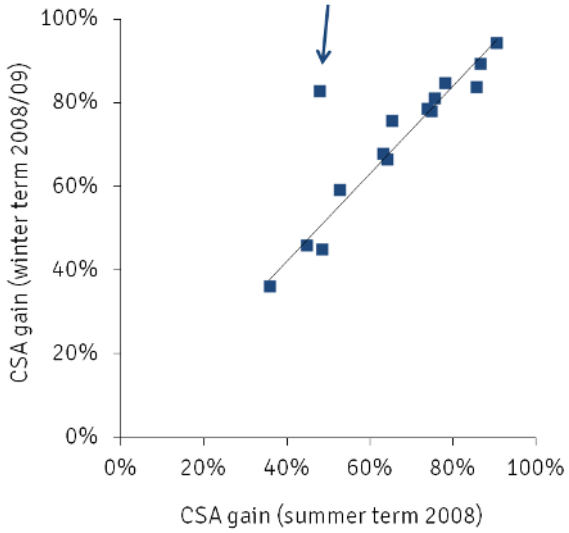

Figure 2: CSA gain for 15 selected learning objectives addressed in a six-week module on cardiovascular and respiratory disease in two consecutive student cohorts. Each dot represents one learning objective. A high correlation was found for 14 learning objectives (r = 0.98). CSA gain for the remaining objective (arrow) was higher in winter 2008/09 than in summer 2008. The self-rating statement for this learning objective referred to the ability to read an electrocardiogram (ECG). Teaching on ECG interpretation skills was improved in winter 2008/09. See text for details. Modified from [17].

c) Validity

Content validity of the novel evaluation tool is self-evident as the statements used for student self-ratings are directly matched to the intended learning objectives. Obviously, the wording of these statements is crucial to their content validity. Drafting self-rating statements can be challenging, and some time and effort will be required to create statements that will produce meaningful results. Readers with limited expertise in this area are referred to the end of this

manual where they will find some hints and clues as well as a few examples of suboptimal and improved statements.

Assessing construct validity of outcome-based evaluation is less straightforward. Any association between changes in teaching and changes in CSA gain might be taken as evidence that the novel measure is a valid proxy of a construct of high-quality teaching resulting in improved student learning outcome. The first piece of such evidence emerged when results of the implementation phase of the novel tool at Göttingen Medical School were discussed. Figure 1 reveals that CSA gain for the statement 'I know which agents can be used to treat urinary tract infections.' was close to zero. When this became apparent at the end of the module during which this should have been taught, the co-ordinator of that module was contacted. It transpired that the lecture addressing this learning objective had been cancelled without substitution. Accordingly, the absence of effective teaching was associated with the absence of a tangible CSA gain. In contrast to this, Figure 2 provides evidence for a teaching intervention resulting in a considerable increase in performance gain. In summer 2008, the practical skill of reading an electrocardiogram (ECG) was mainly taught in lectures. Thus, there was a mismatch between the learning objective domain (skill) and instructional format (didactic lecture). Accordingly, CSA gain was moderate (approximately 48%). In winter 2008/09, teaching was enhanced in that students practiced ECG interpretation in small groups. In addition, performance levels were assessed in a summative written exam at the end of the module. As a result, CSA gain increased to approximately 83%. While this association between an improvement of teaching and an increase in CSA gain is in line with favorable construct validity of the new approach, a direct comparison between CSA gain and objective learning outcome is required in order to establish criterion validity of the approach.

A prerequisite for the assessment of criterion validity is the presence of an adequate external criterion. In the case of learning outcome, one potential external criterion representing true performance gain could be derived from repeated exams matched to the same learning objectives as the evaluation tool. In 2011, all students enrolled in the six-week cardio-respiratory module took formative entry and exit exams targeting 33 cognitive learning objectives. For each of these, one question containing five true/false items was created, thus producing a score between 0 and 5 for each learning objective. Following re-coding of the data, the above formulas were used to calculate subjective and objective performance gain as exam data and self-assessments were both represented on six-point scales. Details of the methods used in this study have been published elsewhere [13]. Main findings of the trial are shown in Figure 3.

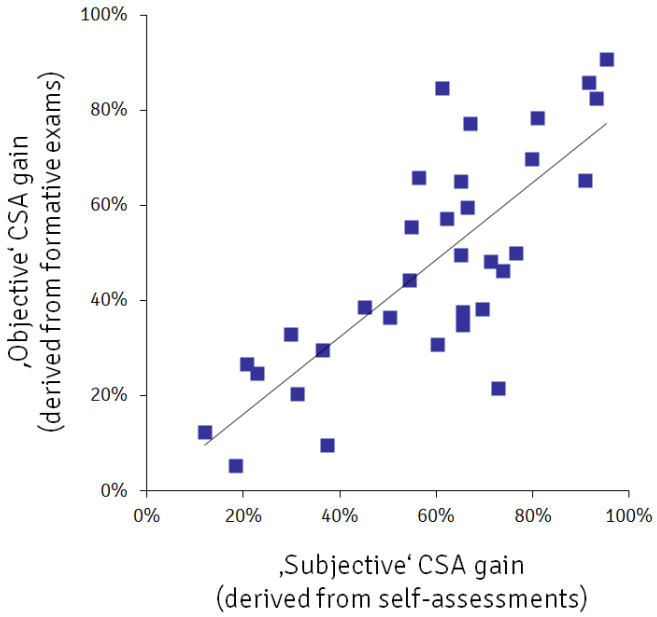

Figure 3: Student learning outcome for 33 learning objectives in the cognitive domain. CSA gain is shown on the x axis while objective learning outcome as measured in two formative exams is shown on the y axis. Pearson's r between the two learning outcome measures was 0.78. Modified from [13].

There was a good correlation between learning outcome calculated from comparative self-assessments and learning outcome calculated from repeated formative exams (Pearson's r = 0.78). A ROC (receiver operator characteristic) analysis revealed that, when using a CSA gain cut-off of 54.7%, the novel tool correctly identified 59% of learning objectives for which objective testing actually indicated a favorable learning outcome (positive predictive value = 0.59). More importantly, the novel tool picked up 100% of cases with suboptimal learning outcome in objective testing (negative predictive value = 1). In summary, CSA gain reliably identifies learning objectives with suboptimal learning outcome. Thus, the novel evaluation tool provides information on potential weaknesses in a module/course which paves the way for improvements in teaching on the level of specific learning objectives.

4. Additional research

Overloading students with evaluation activities can have detrimental effects on response rates [18]. When it was first conceived, the CSA gain tool described above required students to complete two surveys: one before and another one after exposure to a module/course. In order to ensure that the student samples completing each survey were identical, matching of pre- and post-ratings requiring individual labelling of evaluation questionnaires was necessary. This

approach is ethically challenging. As an alternative, students could be asked to provide current self-ratings (post-test) and retrospective self-ratings (then-test) during the same survey at the end of a module. In addition to reducing the burden of evaluations by 50%, this approach is likely to produce more realistic pre-ratings: Students self-assessing their performance levels at the beginning of a module might not fully grasp the complexity of the content taught. As a consequence, they might provide inflated ratings. Exposure to teaching and elaboration of the content taught might result in altered internal standards. Thus, retrospective pre-ratings are likely to be less favorable than prospective pre-ratings ('response shift bias' [19]).

According to these considerations, using only one data collection point (including real post-ratings and retrospective then-ratings) is likely to result in inflated CSA gain values compared to the traditional data collection method outlined above. This hypothesis was tested in another trial involving students enrolled in the cardio-respiratory module at Göttingen Medical School [16]. Self-rating statements were related to the same 33 cognitive learning objectives as in the previous study, but CSA gain was calculated in two different ways: (a) prospective data collection (true post- and pre-test) and (b) retrospective data collection (true post-test and retrospective then-test). The level of agreement between the two methods (Pearson's r = 0.98) is illustrated in Figure 4.

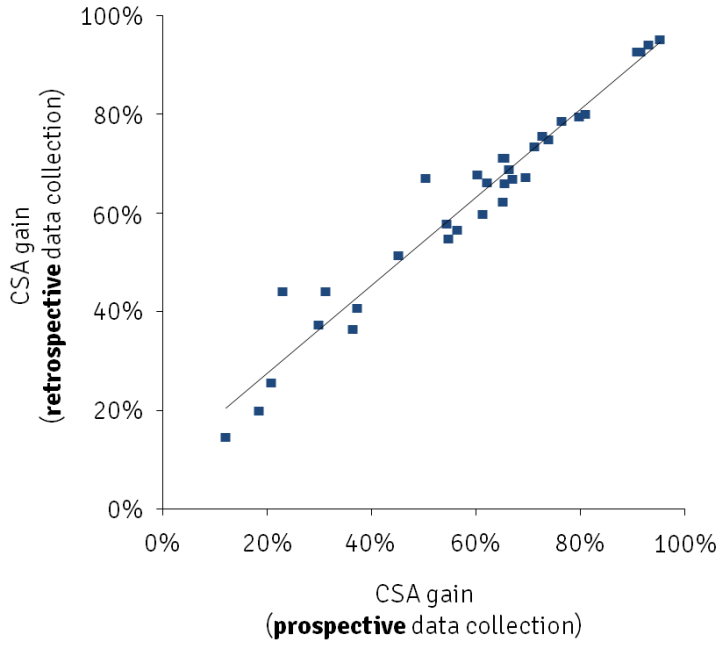

Figure 4: CSA gain for 33 cognitive learning objectives. Data collection was either prospective (x axis) or retrospective (y axis). Pearson's r = 0.98. Modified from [16].

Scantron Corporation – 1313 Lone Oak Road – Eagan, MN 55121 1-800-722-6876

Contrary to the above expectations, the impact of response shift bias on CSA gain was negligible. The mean difference between retrospective and prospective CSA gain was 3.3%, i.e. CSA gain was inflated by 3.3% when using retrospective self-ratings. Given the advantages of using only one data collection point with regard to response rate and the redundancy of student identification, it is advisable to calculate CSA gain retrospectively. However, as all data presented in this user manual refer to undergraduate medical education, the extent to which these findings can be transferred to other higher education contexts needs to be established in future trials. There is also scope for further studies on the minimum/maximum number of learning objectives to be included in a CSA gain questionnaire. More fine-grained analyses require more specific learning objectives to be assessed. This can be challenging in comprehensive modules/courses addressing a large number of learning objectives. While reducing the number of self-rating statements carries the risk of construct under-representation, educators may need to focus on those areas of education where CSA gain data are most likely to have a beneficial impact on curriculum development. The results of outcome-based education should always be interpreted on the level of specific learning objectives with the potential use of a cut-off value as suggested by the ROC analysis discussed above. Calculating 'mean CSA gain' for modules and courses is not recommended as this would unduly reduce the richness of the data.

5. Customizing the outcome-based evaluation tool to your module or course

If you want to use the Class Climate® feature to run your own outcome-based evaluation, this section provides some instructions on how to phrase self-rating statements. Details on the transfer of these statements to the Class Climate® feature, collecting data, retrieving as well as interpreting the PDF report are given in a separate document.

As discussed above, careful preparation of self-rating statements is crucial for the added value (i.e., suggestions for curriculum development) derived from outcome-based evaluation. Adhering to the SMART criteria helps to operationalize learning objectives in such a way that they can be used for outcome-based evaluation purposes. 'SMART' is an acronym representing five features of high-quality learning objectives [20]:

Specific: Descriptions of what students are supposed to learn need to be as specific as possible. (good example: 'I can take a full smoking history.' – bad example: 'I am a good doctor.')

Measurable: A good learning objective lends itself to objective measurements of student performance. Learning outcome may not be quantifiable for all learning objectives (the extent to which affective learning objectives have been met is particularly hard to express in numbers), but there is usually some objective external criterion that can be used to quantify the outcome of

an educational intervention. Since learning objectives must also be specific, blending different objectives in one self-rating statement is not recommended. (good example: 'I can list all risk factors for the development of coronary artery disease.' – bad example: 'I can treat a myocardial infarction *and* provide expert counselling to smokers interested in quitting.')

Attainable: Statements must be pitched to the student level. In other words, the learning objective must lie within the learners' zone of proximal development (good example for first-year students: 'I can name all types of epithelial tissue prevalent in the human body.' – bad example for the same student group: 'I can fully describe the patho-physiological mechanisms underlying impaired endothelium-dependent vaso-relaxation.')

Relevant: It goes without saying that learning objectives used for CSA gain calculations must be relevant for the overall goal of the module/course in question. (good example for future physicians: 'I can detect an acute myocardial infarction on an ECG tracing.' – bad example: 'I can describe all chemical reactions required in the manufacture of aspirin.')

Time-bound: Statements will usually contain a reference to the time frame designated for student mastery of a learning objective. However, this is not relevant for CSA gain calculations as students are asked to self-rate their performance before and after exposure to a teaching module. Thus, the time frame is predefined by the duration of the teaching module.

Here are some additional recommendations for the wording of self-rating statements:

- Statements should be written in the first person (active voice), hence they should all begin with 'I …'
- Statements should allow for a graded response on a six-point scale ranging from 'fully agree' to 'completely disagree'. Statements forcing students to choose between two options (i.e., dichotomous items) are not recommended.
- Statements for learning objectives that can be quantified should contain the word 'all'; inclusion of the maximum value in the statement is not needed (example: 'I can list/name **all** risk factors for the development of coronary artery disease.' – rather than 'I can list/name all five risk factors for the development of coronary artery disease.')
- Qualitative learning objectives and objectives related to concepts rather than factual knowledge might be best captured using phrases such as 'I can elaborate **in detail** on… / I can provide a **detailed** explanation of… / I can **fully** describe… / I can elucidate…**in detail**.'
- Statements referring to psychomotor learning objectives (e.g., practical skills) should adhere to the following structure: 'I can safely/correctly/reliably perform…'
- The use of verbs such as 'believe', 'feel', 'think' should be avoided.
- Statements should be short and precise. Please remember only to cover <u>one single aspect</u> in each self-rating statement. (good example: 'I can fully describe the breathing sounds associated with chronic-obstructive pulmonary disease.' – bad example: 'I can

recognize and fully describe the typical breathing sound associated with chronic-obstructive pulmonary disease'. In the latter example, two different types of learning objectives have been merged: recognizing a typical clinical finding is a clinical skill while describing a breathing sound is a complex cognitive exercise. The statement should be aligned to the actual learning objective.)

Deliberate practice is key to the preparation of high-quality statements used for CSA gain calculations. As this tool has never been used in English before, the above recommendations regarding the wording of self-rating statements will certainly need to be revised. Discussing your own 'home-grown' evaluation statements with colleagues might prove helpful in further carving out the essentials of the learning objective you are trying to assess. In fact, at Göttingen Medical School, self-rating statements undergo a formal peer review process before being used in outcome-based evaluation activities. Please find a few examples of how this process enhances the quality of self-rating statements below (these were translated from German into English):

1) Original wording: I know which emergency actions need to be taken in a case of suspected acute arterial occlusion.

   Revised wording: I can name all emergency actions that need to be taken in a case of suspected acute arterial occlusion. (Reason for revision: Actually naming the actions is more specific than just 'knowing' – or simply believing to know – them.)

2) Original wording: I have understood the genetic causes of hypertrophic cardiomyopathy.

   Revised wording: I can elaborate in detail on the genetic causes of hypertrophic cardiomyopathy. (Reason for revision: lack of specificity in the original statement)

3) Original wording: I know which spirometry parameter indicates the presence of bronchial obstruction.

   Revised wording: I can reliably recognise the presence of bronchial obstruction from a lung function test. (Reason for revision: While the factual knowledge addressed in the original statement is a prerequisite for being able to recognise bronchial obstruction, applying this knowledge to make a diagnosis is more relevant for the overall goal of medical education.)

4) Original wording: I can name the distinct stages of wound healing in detail.

   Revised wording: I can list *all* distinct stages of wound healing. OR I can describe the distinct stages of wound healing *in detail*. (Reason for revision: The first statement lacks specificity as to the type of learning objective (quantitative or qualitative). If this was a quantitative learning objective, just being able to name the stages would be enough to check the 'fully agree' box – however, in this case, the addition 'in detail' is not needed. If this was a qualitative learning objective, students would be required to describe the appearance of

wounds in different stages of the healing process – in this case, the expression ‘in detail’ is an important qualifier, but the word ‘name’ does not adequately represent the nature of the learning objective.)

Literature

1. Gibson KA, Boyle P, Black DA et al. Enhancing Evaluation in an Undergraduate Medical Education Program. Academic Medicine 2008; 83: 787-793 10.1097/ACM.0b013e31817eb8ab.

2. Griffin BW. Instructor Reputation and Student Ratings of Instruction. Contemp Educ Psychol 2001; 26.

3. Naftulin DH, Ware JE & Donnelly FA. The Doctor Fox Lecture: A Paradigm of Educational Seduction. J Med Educ 1973; 48: 630-635.

4. Prave RS & Baril GL. Instructor ratings: Controlling for bias from Initial student interest. . J Educ Bus 1993; 68: 362-366.

5. Blumberg P. Multidimensional Outcome Considerations in Assessing the Efficacy of Medical Educational Programs. Teach & Learn Med 2003; 15: 210-214.

6. Kern DE, Thomas PA, Howard DM et al, *Curriculum development for medical education - A six-step approach*. 1998, Baltimore and London: The John Hopkins University Press.

7. Raupach T, Brown J, Anders S et al. Summative assessments are more powerful drivers of student learning than resource intensive teaching formats. BMC Medicine 2013; 11: 61.

8. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. Acad Med 1991; 66: 762-9.

9. Falchikov N & Boud D. Student Self-Assessment in Higher Education: A Meta-Analysis. Rev Educ Res 1989; 59: 395-430.

10. Cross KP. Not can, but will college teaching be improved? New Directions for Higher Education 1977; 1977: 1-15.

11. Fitzgerald JT, White CB & Gruppen LD. A longitudinal study of self-assessment accuracy. Med Educ 2003; 37: 645-9.

12. Lam TCM. Do Self-Assessments Work to Detect Workshop Success? Am J Eval 2009; 30: 93-105.

13. Schiekirka S, Reinhardt D, Beibarth T et al. Estimating Learning Outcomes From Pre- and Posttest Student Self-Assessments: A Longitudinal Study. Acad Med 2013; 88: 369-375.

14. Raupach T, Munscher C, Beissbarth T et al. Towards outcome-based programme evaluation: using student comparative self-assessments to determine teaching effectiveness. Med Teach 2011; 33: e446-53.

15. Raupach T, Schiekirka S, Munscher C et al. Piloting an outcome-based programme evaluation tool in undergraduate medical education. GMS Z Med Ausbild 2012; 29: Doc44.

16. Schiekirka S, Anders S & Raupach T. Assessment of two different types of bias affecting the results of outcome-based evaluation in undergraduate medical education. BMC Med Educ 2014; 14: 149.

17. Raupach T & Schiekirka S, *Innovationen im Medizinstudium: Stärkung der outcome-basierten Lehre durch Lernzielkataloge und lernzielbezogene Evaluationsinstrumente*, in *Neues Handbuch Hochschullehre*, B. Berendt, et al., Editors. 2013, Raabe Fachverlag für Wissenschaftsinformation: Berlin. p. J2.19.

18. Schiekirka S, Reinhardt D, Heim S et al. Student perceptions of evaluation in undergraduate medical education: A qualitative study from one medical school. BMC Med Educ 2012; 12: 45.

19. Manthei RJ. The response-shift bias in a counsellor education programme. Br J Guid Counsell 1997; 25: 229-237.

20. Doran GT. There's a S.M.A.R.T. way to write managements's goals and objectives. Management Review 1981; 70: 35.