GAVIN PUBLISHERS

Educational Research Applications

Research Article

Sinkvist D, et al. Educ Res Appl: ERCA-125.

Five Year Data and Results of Continuous Quality Improvement Using SKURT

David Sinkvist[1], Annette Theodorsson[2], Torbjörn Ledin[3], Elvar Theodorsson[4*]

[1]Division of Community Medicine, Department of Medical and Health Sciences, Faculty of Health Sciences, Linköping University, Primary Health Care in Linkoping, Local Health Care Services in Central Östergötland, County Council of Östergötland, Sweden

[2]Division of Neuroscience, Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, Department of Neurosurgery, Anaesthetics, Operations and Specialty Surgery Center, County Council of Östergötland, Sweden

[3]Division of Neuroscience, Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, Department of Otorhinolaryngology in Linköping, Anaesthetics, Operations and Specialty Surgery Center, County Council of Östergötland, Sweden

[4]Department of Clinical Chemistry and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

[*]**Corresponding author:** Elvar Theodorsson, Department of Clinical Chemistry and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden. Tel: +46101033295, +4613286720, +46736209471; Fax: +46101033240; E-mail: elvar.theodorsson@liu.se

Citation: Sinkvist D, Theodorsson A, Ledin T, Theodorsson E (2017) Five Year Data and Results of Continuous Quality Improvement Using SKURT. Educ Res Appl: ERCA-125.

Received Date: 28 June, 2017; **Accepted Date:** 28 July, 2017; **Published Date:** 03 August, 2017

Abstract

Student rating of teaching is essential for attaining and maintaining high educational quality. A quality improvement system, SKURT, based on digital online weekly combined quantitative, ten-graded scale, and qualitative, open-ended free text, group feedback from medical students was developed. Students rated all educational, non-clerkship, items throughout the entire medical program, spanning eleven terms. The results were semi-publicly available for students and faculty at a Swedish university. This study describes data from five-year use of the system, focusing on how the use of SKURT influenced educational items found to be in the most substantial need for improvements.

Statistically but hardly practically significant improvement in average feedback grade was found during the observation period (average 7.07 in 2009 to 7.24 in 2013 ($p<0.001$)). The medical program was already in 2007recognized as center of excellent quality in higher education. When analyzing the 18 lectures with lowest outcome in the spring 2009 compared to the fall 2013, five were discontinued. The remaining 13 lectures improved significantly ($p<0.001$) 116% from 2.94 (SD 0.92) to 6.34 (SD 2.58).

A weekly group feedback system employing the principles used in SKURT is useful for improving the quality of medical education particularly by improving the items with the lowest ratings.

Keywords: Medical Education; Online Evaluation; Problem-based Learning; Quality Improvement; Rating of Teachers; Student Evaluation

Introduction

A quality improvement system, SKURT, has been in use since 2008at a Swedish university. SKURT is based on digital online weekly combined quantitative, ten-graded scale, and qualitative, open-ended free text, group feedback from medical students. Students rated all educational, non-clerkship, items throughout the entire medical program, spanning eleven terms. The results were semi-publicly available for students and faculty [1]. The system was created to guide formative ratings, quality enhancement, and educational decisions [2].

In this paper, we describe the data from and consequences of the use of the system during the five-year period 2009 - 2013. In the first paper published simultaneously we describe the philosophy, technical solutions and practical application of SKURT.

Methods

General Considerations

SKURT was used to gather student ratings of non-clerkship educational items including but not limited to lectures, seminars, group activities and information session. Clerkship sessions are practical training in wards, primary care etc.

Students' progress through terms, which means that new student cohorts rate recurring items each term. The 5.5-year medical program infers that a large proportion of students were the same during the analyzed period, but rating educational items different terms. Term 6 was an elective period focused on a research project and without educational items to be rated in SKURT.

Teachers could take part of their individual ratings either through an individual report page or by administrators e-mailing them rating data for lectures through a built-in mass e-mailing or individual item-bound e-mailing function. The individual report page functionality was launched in November 2010 and information about accessing it and its functionality was e-mailed to teachers shortly after launch. Only a single reminder of the functionality was e-mailed to the registered teachers in November 2012.

In the clinical terms 2 lecture weeks are separated by 4 weeks of clinical rotations. The last tutorial group sessions could therefore have preceded some educational items of the last lecture week. Students would then have needed to complete the ratings at the next tutorial group session, after the clinical rotation. This led to a longer time period between item date and rating date for some items.

Ethical Considerations

Dealing with feedback is fraught with ethical dilemmas [3], especially when a component of grading is included. The SKURT feedback was intended to focus on form and content of the educational activity, and not on aspects of the personality of the teachers. The students were informed about this and got feedback on the issue when needed. All feedback was screened before publication. Before analyzing the data for the present study all data were anonymized.

Limitations

The initially chosen database structure of SKURT and its practical use limited to some extent the current general analysis of the data. The scheduling functionality, added post-launch, diluted the items meant for rating with items with mainly scheduling purposes. It was also an issue that the terms did not implement the scheduling function at the same time. Lack of clear guidelines, adherence to guidelines and standardization resulted in disparate item entry, rating and grouping on both term administrator and student level. Some tutorial groups rated, without uniformity, items that were not meant for rating. Group activity classes were not sufficiently grouped when scheduling, resulting in inhomogeneous rating practice. These shortcomings represent confounding factors which risk underestimating e.g. response rates when calculated on aggregated level. Lectures were therefore analyzed separately where relevant, as lecture rating and improvement was the main purpose of SKURT at launch and were the item type with most homogenous rating practice.

Export of Database

All items, ratings and logs between 2009-01-01 and 2013-12-31 were exported to Microsoft Excel 2000 format from the MySQL [4] database using PHPMyAdmin [5]. Items were screened for obvious erroneous entries. A total of 410 items marked as copied, hidden and with date before start of semester (indicating items copied but not placed in the new semester), items hidden in schedule without ratings and items lacking either title or teacher and without ratings were removed.

Excel Functions

All statistics and comparisons were calculated using Microsoft Excel for Mac 2011 standard library of functions [6].

Quality Improvement on Item Level

All lectures with average grade below 4.0 points, the spring 2009, were selected and manually tracked the five-year period.

Statistical Analysis

Mean and standard deviation or median and quartiles were used as measures of central tendency and variation, respectively, as appropriate. Two-sided unpaired Student's t-test was used for significance testing in general. Two-sided paired Student's t-test was used when comparing quality improvement on item level.

Results

General Results

Item spanned from 2009-01-19 to 2013-12-20, 13 684 educational items and 71 883 ratings were registered the period. 37% (5 105) of items were lectures, 35% (4 835) group activities, 12% (1 659) tutorial groups and 15% (2085) others. 68% (49 181) of all ratings had an open-ended feedback and 64% (45 936) had a grade.12% (8 913) of ratings were open-ended feedback without grade whilst 8% (5 668) were graded ratings without feedback.98% (5013) of the 5 105 lectures had at least one rating and 85% (31 090) of lecture ratings had an open-ended feedback.

Rated lectures had an average 7.34 ratings with a positive trend in number of ratings per lecture from 6.09 to 8.43, correlating with increase in the number of medical students and tutorial groups by 27% (722 to 919 students) during the period. Number of tutorial groups spanned from 8 to 15 the fall 2013 and response rate for lectures ranged from 63% to 91% with an average for all terms of 77% (Figure 1).

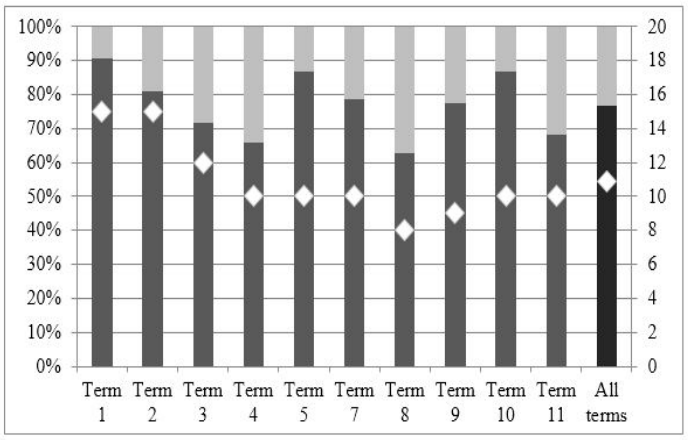

Figure 1: Lecture response rates (left axis) for different terms and number of tutorial groups (right axis) the fall semester 2013.

The average open-ended feedback was 105 (10th percentile 16, 1st quartile 40, median 81, 3rdquartile 143 and 90th percentile 220) characters. The longest feedback was 3 326 characters.

The proportion of positive ratings, grade over 5 points, constituted the overwhelming majority of ratings (Figure 2). The average grade per rating was 7.14 (SD 1.73). There was a statistically but perhaps not practically significant increase from average 7.07 in 2009 to 7.24 in 2013 ($p<0.001$).

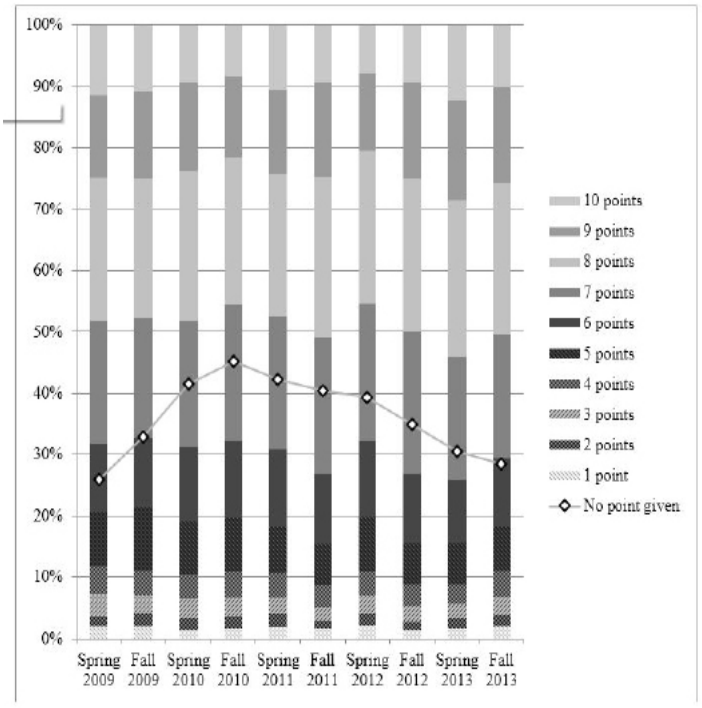

Figure 2: Distributions of rating grades amongst rated items per semester.

512 of 1 353 educational items the spring 2009 were found with an exact matching title the fall 2013. The remaining 972 educational items the fall 2013 were new, revised or renamed.

The average number of days from item date to rating date was 5.62 (SD 3.44) with 10th percentile 0, 1st quartile 1, median 3, 3rd quartile 7 and 90th percentile 12 days.

Screening of Qualitative Feedback

40 (0.06%) feedback responses were edited during the period and all were marked as screened, and published, by at least one term administrator. In five of the terms no editing had taken place and in the remaining five terms there were 4-15 changes made.

A manual analysis of all edited feedback responses revealed that 21 were merely spelling errors, 3 were judged as improper editing and the remaining 16 (0.02%) changes were considered in line with the main purpose of the functionality.

Teacher Reports

629 teachers were assigned at least one educational item.

54% (338) had logged in at least once to their individual rating report page. 1880 logins were registered during the period yielding an average login of 5.6 logins per logged in teacher the six semesters after functionality launch.

10 519 e-mails were sent to 611 unique e-mail addresses belonging to teachers and administrators. 492 of teachers assigned to an item during the period got at least one e-mail summary of students' ratings. 85% (533) of the teachers with educational items the period had either received an e-mail summary, logged in to their individual page or both.

Quality Improvement on Item Level

25 lectures had an average grade less than 4.0 points the spring 2009. Three received low grade because of cancellation or being called off whilst four were actually not lectures and none of them recurred the other semesters. The remaining 18 lectures evolved as depicted in Figure 3.

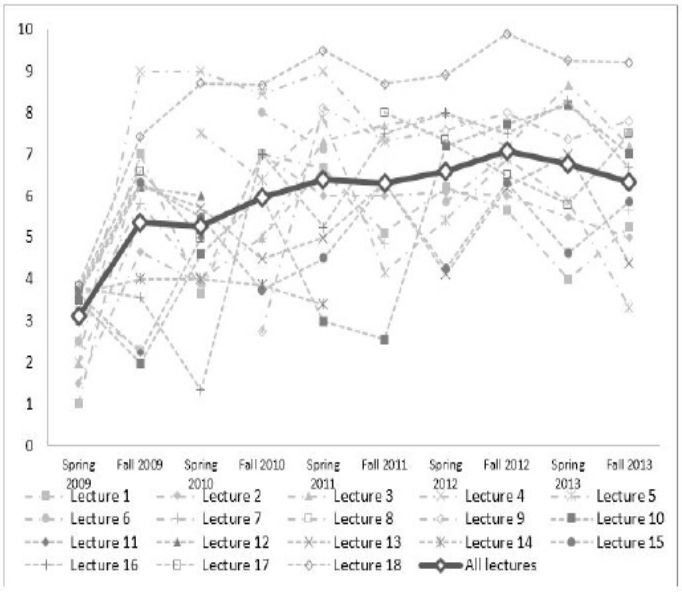

Figure 3: Evolution of lectures with average grade below 4.0 points the spring 2009 including average grade for all 18 lectures.

The 18 lectures were distributed amongst all terms except term 6 and 10. 92% (943) of ratings included open-ended feedback with an average length of 149 characters (10th percentile 34, 1st quartile 71, median 125, 3rd quartile 198, 90th percentile 285 and maximum 1 425 characters).

Lecture 1 had lowest starting grade. Feedback led to an increase in average grade. Feedback included recommendations on how to improve and update the handouts, requests for areas to be clarified and requesting a summary of the most important take home messages related to plans of study including clinical context.

Lecture 2 had an average grade of 3.35 the first three semesters. When the teacher was replaced, including new structure and lecture content, the grade increased to an average 5.94 the following seven semesters. The lecture was a combined lecture and introduction for a seminar assignment. Initial criticism pointed out a low proportion of real lecture content in comparison to seminar information, which did not recur in later semesters even though lecture title still included the introduction.

Lecture 3 initially received low grades including feedback on shortcomings in tutoring style, lecture content and handouts. An additional teacher backing up the first was added and shared teacher responsibility improved the grades slightly. The lecture improved more after recruiting a new pair of teachers. The last four semesters only one of them was assigned the lecture and received positive student feedback and an average grade of 7.76. Feedback included appreciation of a new and revised handout, clinical examples and overall content.

Feedback on the first version of lecture 4 included requesting a more clinical and practical rather than a solely research perspective. The opinion was swiftly adopted and the assigned teacher was changed from a pre-clinical researcher to a clinical physician better suited for aiding the students attaining the learning goals. The initial grade of 2 increased to 9 the following semester and the high grades were maintained with an average of 8.55.

The same teacher as in lecture 1 also improved average grade for lecture 5. Students also here initially requested an updated handout and a clinical perspective on the teaching. The increased average grade from 2.5 to 7.5 was accompanied by compliments, in later semesters, on an appreciated handout. The dip for lecture 5 in the fall 2013 was due to change of teacher that specific semester.

Lecture 6 improved from an initial grade of 2.5 to an average of 7.08 in the following semesters. Feedback included that the lecture felt over loaded with information in a too short timeframe and students requested either an extension of time or abbreviating the lecture with more focus on term-specific goals.

Lecture 7 was, after six semesters with an average grade of 4.89, discontinued for three semesters and recurred in another term where it was expected to be better suited. The lecture was then, mistakenly, held for a class that already had the lecture three terms earlier, as stated by a majority of feedback. The revised lecture did nonetheless receive a slightly higher average grade of 5.67.

Lecture 8 was not repeated after spring 2009. Feedback was not useful, as only a minority of the groups included either feedback or grade. Two of the feedback responses stated that the students did not recall the lecture, which could indicate that the lecture was not held.

Lecture 9 had four semesters of low grades with an average of 2.8. Students gave feedback on teaching style, feeling of too extensive shallow information in too short time and questioned the clinical applicability. Some students suggested change of teacher and the following semester a new teacher was assigned the lecture. The grade rose for the coming six semesters to an average of 7.7 whilst keeping the same lecture title and purpose.

Lecture 10 had six consecutive semesters of an average grade of 3.61 with the assigned teacher lecturing both alone and together with colleagues. After change of teacher the average grade increased to 7.53 for the last four semesters of the period. Students initially gave feedback on the difficulty of the subject, lack of introduction before dealing with difficult concepts and on the teacher's pedagogic skills. The increase in average grade was accompanied with description of the lecture as pedagogical, structured and thorough.

Lecture 11 and lecture 12 were rated as irrelevant, unnecessary, lacking new information and misplaced on the term. Both were discontinued after three semesters.

Lecture 13 had a shared teacher design the first five semesters. Student feedback included that the time allocated was too long and that coordination between teachers could have been better as some identical information and cases were repeated during the same lecture. The lecture was split to two separate lectures for the last five semesters with one keeping the title but with only one of the teachers. Students still found that the, now two, lectures were overlapping and one feedback even included a suggestion that the two lectures maybe could be combined to one. Only a slight improvement of average grade was noted.

Lecture 14 was discontinued after five semesters as students repeatedly gave feedback that the lecture title was not in line with its content and the lecture lacked pedagogical properties and clinical context.

Lecture 15 improved over time and initial feedback from students included lack of time and some factual errors in the handouts. The lecture was split into two parts the last semester and average grade increased slightly from 5.04 the previous semesters to 6.7 (part 1) and 5.0 (part 2) respectively.

Three different teachers were, involved in lecture 16 the first five semesters and attained an average grade of 4.19. Students gave feedback on lecture layout, presentation style and expressed feelings that the lecture and teacher run the risk of becoming uninspiring. The delicate feedback was though combined with several constructive and detailed suggestions for improvements. The average open-ended feedback was 303 characters the first three semesters and response rate was 93% (100%, 80% and 100% respectively). A new teacher was assigned the lecture the fall 2011 and the lecture thereafter received a stable average grade of 7.6. Students gave feedback that the lecture was valuable, informative, pedagogical and clarifying of an otherwise difficult subject.

Students gave feedback on lack of time and unordered lecture layout and handouts for lecture 17. A lack of time was still experienced by some students in later ratings but the handouts and structure were no longer stated as areas of improvement in the feedback. The lecture was called off two semesters because of unrelated circumstances. Nonetheless the average grade increased over time.

Lecture 18 improved drastically to a stable high grade. Students initially gave feedback that the lecture was too early in the term, that they lacked the necessary prerequisites and that therefore the lecture was too difficult to comprehend. In coming terms, the lecture was moved to about a month later in the term and thereafter receiving an average of 8.92.

In summary, all recurring lectures improved and 13 were still part of the curriculum the fall 2013.Increase in average grade was associated with a change of teacher in six (46%) of the lectures.

The average grade for the 18 lectures the spring 2009 improved from 3.01 (SD 0.89) to 6.34 (SD1.56) for the remaining 13 lectures the fall 2013 (Figure 3).The 116% improvement of the 13 items from 2.94 (SD 0.92) the spring 2009 to 6.34 (SD 2.58) the fall 2013 was significant ($p<0.001$).

Discussion

Our results indicate that a weekly group feedback system employing the principles of SKURT can be used to improve the quality of medical education particularly by improving the items with the most substantial need for improvements as perceived by the students.

At an aggregated level the already high average grade did improve statistically but hardly practically over the five-year period. The high starting average grade, the medical program already in 2007 recognized as a center of excellent quality in higher education [7], and that a large number of items were new, revised or renamed the fall 2013 compared to the spring 2009 could contribute to the lack of larger general improvement. A negative correlation between class size and rating results have previously been noted [8] but was not evident in the current data.

The educational items receiving the lowest grades in the spring 2009 improved significantly and practically aided by the use of SKURT. Average feedback grades more than doubled and students' narrative feedback included improvement-focused responses with direct tangible recommendations that were adopted. Lectures receiving low feedback grades had high feedback rates and longer feedback length indicating that students were contributing to improve educational items, which is in line with previous research [9]. When criticisms were harsher the responses were

also longer and more elaborate. Five of the low graded items were removed from the curriculum and new, revised or renamed items were added instead.

A lack of standardization regarding content and nomenclature is a weakness of the current database structure. It diluted the data with items and ratings not in line with the main purpose of the quality improvement system. Group activities were scheduled and rated in disparate ways in different terms and semesters. Ratings were sometimes unrelated to teacher and item, such as criticism aimed at administrators when items were called off.

The miniscule number of edited feedback, even though all items were marked as screened, could indicate that students generally provided feedback that was not fraught with focus on the teachers' personality but rather on substantial issues. The range (0 to 15) of the number of edited ratings each term however indicates inhomogeneous use of the function and probably indicates that some of the administrators did not screen thoroughly enough or did not feel that they had permission to edit responses. To further improve feedback quality the new version of SKURT includes a clearly visible summary of the feedback model, as summarized by Ovando [10], when students rate. A guideline for administrators' permissions and responsibilities when screening and editing feedback is under development. A report button in association with each feedback could further help identify feedback not following guidelines, as unconstructive comments are a risk noted with student ratings [11].

The possibility to assign new teachers to educational items suffering from low grades represents a powerful tool for improvements. In other curricula, commonly based on few teachers assigned to weeklong courses, change of teacher is unlikely and seldom possible. In this case, application of the principles SKURT is built on could then instead guide teachers' professional development and improve practical class pedagogical methods, as was seen in several of the low graded items, with timely feedback continuously during the course enabling a recurrent open feedback loop with continuous improvement of quality [12-14].

The mean duration of time from the educational activity until feedback is given indicates that SKURT is used as regularly as intended. The short time period ensures an effective feedback loop. The item-specific feedback enables development of different pedagogical methods and structures as well as optimizing the recruitment of teachers to the activities. Changes based on specific tangible feedback can be channeled into better horizontal and vertical integration throughout the program and also to better constructive alignment. Even moving lectures, seminars or practical learning activities a month or two in time, to fit better with the learning curve, can result in major improvements.

The high response rates for lecture ratings were without systematic individual or group reminders and with voluntary ratings, even though tutorial groups were mandatory. The format also seems to counteract factors such as forgetting and lack of time noted in previous studies [9,12,13,15,16]. The response rate was 25 percentage points higher than the medical programs average response rate 2007-2012 on the university wide end-term online rating system [17]. It should, however, be noted that the end-term online rating system is based on individual students' ratings and not on group ratings. Weekly ratings and the online format did not result in overuse of students' interest and motivation with accompanying low response rates as previously have been noted and feared [18-22]. The students have the option of giving feedback, grading or both. A majority of ratings received an open-ended feedback and fears from faculty that students would by routine only grade items and not write feedback were unrealized.

The integration of SKURT in the campus culture, involvement of student organizations, improvements based on rating data and student's wide-spread notice of their feedback importance [12,18,21,23-25] could be factors promoting the high response rates. Students soon started using a custom-created verb, "Skurta", with the meaning "Using SKURT", which illustrates the incorporation in campus culture.

Areas of improvements include clearer guidelines and more rigorous control of nomenclature, data structures and input of both items and ratings. Building in reminders of uncompleted ratings could increase response rates further. Other options, than giving unjustified low ratings including frustration over called off items, such as a weekly "General Feedback"-option would increase validity of item ratings. A push-notification system for teachers when new ratings are published would increase the usage of the individual ratings page. Manual cleansing and standardization of the database could enable further aggregated analysis of the data.

In summary the principles applied in SKURT which conveniently can manage vast amounts of data, have been an important tool in improving low graded educational items by providing item specific timely feedback from students. The response rates were high and no signs of wear and tear on the students caused by too much feedback activities were observed.

Acknowledgements

David Sinkvist programmed SKURT and administered all server applications. The three other authors of this manuscript constituted a project group throughout the entire duration of the project, and as heads of the medical program (TL, AT) communicated with teachers (ET, TL, AT), students (DS, ET, TL, AT), institutions (ET) and the Faculty (ET, TL, AT).

Declaration of Interest

The intellectual rights and the copyright to SKURT belong to David Sinkvist.

References

1. Sinkvist D, Theodorsson A, Ledin T, Theodorsson E (2017) SKURT: Quality Improvement System with Comprehensive Weekly Digital Student Group Feedback.Educ Res Appl: ERCA-124.
2. Berk RA (2013) Top five flashpoints in the assessment of teaching effectiveness. Medical Teacher 35: 15-26.
3. Norton LS (2012) Action research in teaching and learning: A practical guide to conducting pedagogical research in universities.Teach Theo Religi 15: 302-303.
4. Bellman E (1999) Tio års erfarenhet som sjukhusdirektör: Stora förändringar bör göras i små steg. Lakartidningen 96: 3186-3189.
5. Blumsohn A, Eastell R (1992) Prediction of bone loss in postmenopausal women-. Eur J Clin Invest 22: 764-766.
6. Microsoft. Excel for Mac. 14.4.1 ed: Microsoft: 2014.
7. Centres of Excellent Quality in Higher Education 2007. Högskoleverket: 2008.
8. Pounder JS (2007) Is student evaluation of teaching worthwhile?: An analytical framework for answering the question. Quality Assurance in Education 15: 178-191.
9. Luks AM (2007) An alternative means of obtaining student feedback. Med Educ 41: 1108-1109.
10. Ovando MN (1994) Constructive feedback: A key to successful teaching and learning. International Journal of Educational Management 8: 19-22.
11. Jones J, Gaffney-Rhys R, Jones E (2014) Handle with care! An exploration of the potential risks associated with the publication and summative usage of student evaluation of teaching (SET) results. Journal of Further and Higher Education 38: 37-56.
12. Nevo D, McClean R, Nevo S (2010) Harnessing information technology to improve the process of students' evaluations of teaching: An exploration of students' critical success factors of online evaluations. Journal of Information Systems Education 21: 99.
13. Anderson HM, Cain J, Bird E (2005) Online Student Course Evaluations: Review of Literature and a Pilot Study. American Journal of Pharmaceutical Education 69: 34-43.
14. Winchester TM, Winchester M (2011) Exploring the impact of faculty reflection on weekly student evaluations of teaching. International Journal for Academic Development 16: 119-131.
15. Stowell JR, Addison WE, Smith JL (2012) Comparison of online and classroom-based student evaluations of instruction. Assessment & Evaluation in Higher Education 37: 465-473.
16. Donovan J, Mader C, Shinsky J (2007) Online vs. traditional course evaluation formats: Student perceptions. Journal of Interactive Online Learning 6: 158-180.
17. Läkarprogrammet, Linköpings universitet, Självvärdering 2013. Faculty of Health Sciences, Linköping University; 2013. Report No.: 411-155-13.
18. Winchester MK, Winchester TM (2012) If you build it will they come?; Exploring the student perspective of weekly student evaluations of teaching. Assessment & Evaluation in Higher Education 37: 671-682.
19. Palmer S (2012) The performance of a student evaluation of teaching system. Assessment & evaluation in higher education 37: 975-985.
20. Cleary M, Happell B, Lau ST, Mackey S (2013) Student feedback on teaching: Some issues for consideration for nurse educators. Int J Nurs Pract 19 Suppl 1: 62-66.
21. Schiekirka S, Reinhardt D, Heim S, Fabry G, Pukrop T, et al. (2012) Student perceptions of evaluation in undergraduate medical education: A qualitative study from one medical school. BMC medical education 12: 45.
22. Stead DR (2005) A Review of the One-Minute Paper. Active Learning in Higher Education the Journal of the Institute for Learning and Teaching 6: 118-131.
23. Alderman L, Towers S, Bannah S (2012) Student feedback systems in higher education: a focused literature review and environmental scan. Quality in Higher Education 18: 261-280.
24. Wright SL, Jenkins-Guarnieri MA (2012) Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use. Assessment & Evaluation in Higher Education 37: 683-699.
25. Youssef LS (2012) Using student reflections in the formative evaluation of instruction: a course-integrated approach. Reflective Practice 13: 237-254.