Quantitative and Qualitative Evaluation of Transforming to Flipped-Classroom from Instruction Teaching using Micro Feedback Loops

Andre Mastmeyer

Aalen University, Germany

and University of Luebeck, Germany

Author Note

Andre Mastmeyer, Aalen University, Studies in Digital Health Management, Germany,

and Institute of Medical Informatics, University of Luebeck, Germany

Contact: andre.mastmeyer@hs-aalen.de

Version 0.5 - Thursday, August 27, 2020

Abstract

Recently, the institutionalized transformation of frontal instruction classrooms into active learning spaces to foster the concept of (inter-)active learning has gained increasing attention. To investigate the impact of elements of active learning on learning reception of students in an advanced small sized MSc STEM course (<25 students), a traditional instructor teaching style class was transformed to flipped-classroom teaching. Before and after each lecture, anonymized evaluation Likert items from the students were recorded. Thus, both teaching styles for every given lecture were covered equally. In both classrooms, some didactic and methodological elements were kept constant, while others were changed when flipped-classroom took over semester midterm. Qualitative and quantitative results indicated that the flipped-classroom format generated greater learning effects as well as classroom enjoyment, fostered students' self-regulated learning, enhanced group interaction, stimulated group activity and guaranteed a more synergistic learning behavior.

Keywords: STEM, active learning transformation, frontal instructional teaching, flipped-classroom teaching, micro-feedback loops.

Quantitative and Qualitative Evaluation of Transforming to Flipped-Classroom from Instruction Teaching using Micro Feedback Loops

Introduction

In a SoTL (Miller-Young & Yeo, 2015) project at the University of Luebeck, Germany, the transformation of a formerly frontal lecture to flipped-classroom teaching was studied. In instructional teaching classrooms, the teacher traditionally is giving the lecture orally on his own during the whole-time span of the lecture using several kinds of media such as chalkboard, flipchart and PowerPoint presentations. Yet increasingly, interactive elements being used are quiz-elements serving as "time to think" breakpoints. These can be augmented with electronic media (Hwang, Lai, & Wang, 2015), e.g. for in class voting. They can be seen as a first step towards more interaction in the classroom, but are shallower than a complete human and responsible engagement of the student in the course of a lecture.

Modern educational aims must try to engage and enlighten students on a deeper cognitive level, as well as let them develop skills useful in their own professional career or eventual teaching or scientific career (Auster & Wylie, 2006). Traditional teaching (TT) formats are being criticized for good reasons: Their perception by students is autocratic, sometimes eccentric and they hardly pay attention to the multidimensional diversity of students and diverse learning types and styles. TT goes along with traditional passive learning. Most valuable students' self-activity is postponed to homework, seminars or on-site exercises, which could be seen as inefficient use of invested time during lectures. Typically, students are apt to sit in the classroom bored off and listen to the teacher (as reader) with limited attention while occasionally taking notes or using smart phones (Steuter & Doyle, 2010). So, having observed the flaws of passive learning for decades, probate alternatives to passive learning are sought that involve a greater commitment of the teacher (as coach) and students (coachee) as well to their synergistic enlightenment. On the one hand, coaching and supporting

students' active contributions in the course of lectures is clearly much more challenging than just reading a lecture frontally – it also comprises social learning about the target group. On the other hand, students also have to invest more energy and time into active learning formats.

Two increasingly popular alternatives of active learning are problem-based Learning (PBL) (Hung, Jonassen, & Liu, 2008) and flipped-classroom learning (FC) approach (Davidson, Major, & Michaelsen, 2014). With regards to PBL, several universities already have transformed whole study programs to be able to advertise and dominantly use the new format.[1,2] Thus, with PBL being the already more established method, FC needs more research to follow the success of PBL. Currently, there is no university marketing with FC teaching around the globe. However increasingly, institutions adopt the method into their curricula. The difference between FC and PBL is how and which topics and holistic human cognitions and actions are focused. FC encourages students to take over the responsible role of the teacher conveying curricular learning goals. PBL puts emphasis on problem solving in small case studies in a study discussion and work group format (Tawfik & Lilly, 2015). In both formats students actively and creatively take part in the course lectures and act in a social context as required in diversity- and gender-oriented teaching and especially later in their professional life later on.

As indicated also from another point of view, the use of active learning methods actually gets mandatory in the future classroom: The distraction of students by the new media environment is an elephant in the room to be discussed and addressed. Facebook, Twitter and the like are a challenging distraction, especially with large classes.

Here, in the proposed flipped-classroom course format, the students were encouraged to bring their own devices such as laptops and use them for the benefit of the lecture. However, there is no holy grail, and flipped-classroom is just one child of the active learning philosophy (Figure 1). Other active learning strategies may be pursued (Jensen, Kummer, & Godoy, 2015) following the

same core principles with equal success. They fit with the paradigm of constructivism in learning, where teachers always allow students to construct knowledge in their own way for themselves while advising them (Piaget, 2013). The foundation for constructivism was laid out by the Russian psychologist Lev Vygotsky, who stated that social interaction and learning precedes cognitive development and empowerment (Crawford, 1996).

Learning by tutoring has been analyzed favorably early by Bloom (Bloom, 1984) (Figure 1). The next evolutionary step is, let students take over the class - however, supervised. In the literature, positive aspects of FC teaching were shown e.g. for organic chemistry courses (Fautch, 2015). In this work, the focus is a densely monitored transformation of a traditional to a flipped classroom (<25 students) in the STEM[3] disciplines during the semester term and how this change is received by the MSc students using qualitative and quantitative evaluation.

In perspective, active learning methods may also be customized successfully to large classrooms of 45-85 students (Klegeris & Hurren, 2011). Also, the productive use of new media inside the active classroom leads to the "active digital classroom" concept (Hwang et al., 2015).

This work contributes to the research field of SoTL by proposing re-conceptualizing means and a new process and evaluation framework to transform TT to new teaching formats (here: FC), while especially continuously monitoring the reception of the transformed teaching format before and after each lecture, and comparing FC vs. TT lecture reception. Additionally, TAP (teaching assessment protocol) interviews were conducted mid and end semester term (Angelo & Cross, 2012). The work was carried out in the SoTL (Miller-Young & Yeo, 2015) program at the University of Luebeck, Germany.

This paper is organized as follow: 1. The instructional TT and FC course format is presented in the material and methods section. 2. The study design section lays out the continuous and final

evaluation methodology. 3. The results section shows the results following the study design. 4. The discussion and conclusion close the paper.

Materials and Methods

The lectures topic was "Virtual Reality in Medicine". This was a course with technical STEM[3] character, where data structures and algorithms for the visuo-haptic virtual reality simulation of medical interventions such as needle punctures were taught (Mastmeyer, Fortmeier, & Handels, 2016; Mastmeyer, Wilms, Fortmeier, Schröder, & Handels, 2016). Lecture time was two hours (core lecture time: 90 min.) in the middle of the week. Another 90 minutes beyond this study and same to both teaching formats were spent end of the week in on-site exercises elaborating the content of the lecture. Four homework letters with a return time of two to three weeks were issued during the semester. Students teamed up in pairs to solve them. 50% of the homework scores granted admission to the final oral exams. A new micro-feedback loop per lecture was implemented: A questionnaire was handed out to the students to be filled in two parts. The first part was answered before the lecture, the second afterwards. Some question items were paired. Semester mid-term, the educational objectives were TAP assessed (Angelo & Cross, 2012).

The aim of a 31 items questionnaire was to get instant feedback at the beginning and end of each lecture. All questions asked before the lecture, have counterparts after the lecture to allow for instant comparison how a lecture was perceived and allows adaptations and steering of the course format and content by the teacher. At the time, the questionnaire was handed out in paper format and digitized to tables by a scanning system.[6]

The Traditional Instructional Course Format (TT)

TT was given in seven weekly lectures. PowerPoint slides with the dual function of serving as script for home studying were read in TT format by the teacher using a laser pointer. Activating quiz

elements were used as breakpoints to trigger thinking and motivating oral students' contribution to the course.

The Flipped-Classroom Course Format (FC)

New totally activating teaching was introduced in the second set of six weekly lectures. The upcoming lectures PowerPoint slides were distributed by an institutional Moodle-web-system to the students one week ahead. Study groups were defined to team up and identified a team speaker. At the time, they were congruent to the homework teams. They were provided a weekly doodle-poll (Figure 2) to vote for the part of the lecture they want to take over next week.

The classroom featured a projection screen, a projector, and tables arranged in a U-formation with 24 seats. Generally using self-coaching capabilities of group dynamics, a number of four groups with six heads each could be regarded as maximum still enabling to deliver effective coaching. At the opening of the U, the projection screen was installed. A long whiteboard was available at one side of the U. The student with the teaching role stood between the top of the U and the whiteboard pointing with a laser pointer to the screen. With the advent of bringing in own devices for lecture talk preparation, the teacher and the learners created a collaborative learning situation, a so called active classroom (Meyers & Jones, 1993).

The time ahead of the lecture with the students already possessing the lecture material served for questionnaire part 1 (B), final material-related questions and the emotional preparation to take over the role of the teacher. The study groups' workload was standardized by assigning an equally fixed number of slides to each group. Group number separation slides for the groups were inserted into the lecture material by the teacher, i.e. title slides with title "FC Group 1" to "FC Group N" (Figure 2). The days after the just held lecture were used by the study groups to vote their next week's part. Particularly, the lecture format was augmented by these detailed elements:

1. Questionnaire Part 1 (TT and FC) – **B**efore each lecture: The **B**-part of the questionnaire (Table 1) consisting of 10 Likert items was filled in. These items were repeated and mixed for dispersion with other questions only occurring in questionnaire part 2 (A).

2. Instant Feedback Loop Part 1 (TT and FC): Educational goals had been defined by the students at home and were written to the whiteboard in the first 15 minutes before the actual lecture start (c.t.-time[4]). The teacher occasionally refined the delivered goals and discussed and refined them concisely with the students.

4. Off-site Flipped-Classroom preparation and Student Presentations (FC only): Each study group was in charge of their slide section and home-made preparation of the lecture slides presentations. The task of the teacher was coaching the study groups and providing tips and useful insight into difficult parts of the slides and how to present them skillfully. Thus, the teacher's role was coach, backup and member of all groups, stepping in for a missing group member or if (ever) an oral delivery went insufficiently. After a group presentation, a one to two minutes discussion and feedback panel were held.

5. Continuous Small Oral Exam Simulations (TT and FC): On some occasions using questions arising during the course of the lecture or stemming from quiz-elements (integrated into the slides), students were picked by the teacher to contribute to the questions answer on the whiteboard. This way, a transparent situation simulation regarding the final oral exam was provided to the students. Furthermore, the active contribution and the switch to the FC format was also encouraged in the TT lectures.

6. Instant Feedback Loop Part 2 (TT and FC): The defined educational goals from 2. were reviewed in a Q&A again at the end of the lecture. Thus, the students and the teacher got instant feedback, how the lecture was digested. The teacher was then able to eventually re-elaborate on some picked topics in the 90 minutes on-site exercises at the end of the week.

7. Questionnaire Part 2 (TT and FC) – **A**fter the lecture: The list of 20 **A**-questions (Table 2) to be answered in the 15 minutes **after** the lecture was handed out to the students and filled in while the lecturer was de-installing the laptop and beamer presentation. For dispersion, ten A-questions were not paired to former part 1 B-questions in this part of the questionnaire.

Study Design

The 12 head study population consisted of 8 male and 4 female master students. The age distribution was 25.17±1.95 and they had quite a history of numerous lectures in TT format. To compensate to some degree for a small but yet sufficient number of subjects, the number of questions was set high and some redundancy in the questions was introduced.

The questionnaire that was handed out to the students partly before and after the lecture is shown in Table 1 and Table 2. Questions addressed **cognitive**, **affective** and **didactical** categories. To enable **objectivity** and personal data protection, study member identification a pseudonym signature was used and the factor **Student** is not pursued in monitoring and to the end of this study. Each student filled in the answers using a memorized nick name re-used in each lecture. The (formerly frontal) lectures slide material was enough for max. 24 students in FC teaching format, i.e. min. ca. two slides per speaker. Here, the FC student group size was two, which is preferable as it coincided with the homework submission groups and off-site lecture preparation was considered homework - instead of more homework letters.

Regarding statistical **reliability**, the micro-feedback questionnaires in two parts per lecture delivered enough Likert data samples (4692) on a six level Likert scale (Matell & Jacoby, 1971) to conduct statistical procedures meaningfully with enough power for the main factors (format, time point), i.e. vs. a small number of degrees of freedom in the model formulation (Ryan, 2013). For increased **validity**, a Likert scale with an even number of items was chosen, to provoke a clear decision in the middle of the scale. Obviously, dedicated monitoring questions concerning the FC

format were not answered in TT lectures (A29, A30) except lecture 9. Especially, the negative question A29 served as general monitoring question and sequence bias test against stereotypical answering.

The second questionnaire part repeated some questions (pairs) from Table 1 and introduced new items (Table 2) for distraction and more valuable information. The before-after question pairs are shown in Table 3. The paired Likert items define comparison pairs to be analyzed post-hoc.

The purely qualitative, standardized TAP questions (T) and a supplementary question (S) were worded:

T00. What helps learning most in this course?

T01. What makes your learning more difficult?

T02. What suggestions do you have for improving on the obstacles?

S04. What do you think about using FCL right from the beginning of the lecture term?

The answers to the questions moderated without the lecturer present were noted, condensed, counted and weighted in the TAP vote and are presented synoptically in the results.

Main Effects Analysis

Two main two-level factors were used: Flipped-Classroom (**FC**=Y or N with >2088 samples), and before and after the lecture (**B/A**=B or A with >1560 samples) from multiple questions.

The measured variable was the Likert score between 1 (fully disagree) and 6 (fully agree). In this work, the ordinal scale motivated the author to use the generalized linear model family in SPSS while keeping SPSS's default independent structure working correlation matrix and an "ordinal logistic" model type (Lawal & Lawal, 2003). To further cope with the taken repeated measurements structure, in a "Generalized Estimating Equations" (GEE) framework (Loeys, Cook, De Smet, Wietzker, & Buysse, 2014) **Student** was chosen as subject variable, while **Lecture** and **Question** served as within-subject variables for ordering the Likert score measurements. Main effects and two-

way variable interactions were considered along the research questions and for model sensitivity reasons. Iterative elimination of highest insignificant interactions from the set of all two-way interactions was carried out (Carey, 2013). The most cascaded (*) and insignificant interactions were eliminated first, and then the model fit was run again after each term elimination step.

Covariate factors instead of fixed factors were chosen when an ordering could be applied and the number of value realizations was greater than four. The pursued research questions to be answered with a focus on the dichotomic **FC** fixed main factor were:

1. Is there statistical difference in the main, or covariate effects?:

FC,

B/A, or

Lecture,

Question.

2. Is there a two-way interaction effect (*) especially for **FC** with the other variables, e.g.:

FC * B/A,

FC * Lecture,

etc.

The main effect factors therefore are "before", "after" (**B/A**) the lecture, "traditional" vs. "flipped" (**FC**), and factor interaction between **B/A** and **FC**, i.e. especially if the resulting Likert scores were higher with **FC**=Y.

In summary, an univariate multifactorial general linear model was used with repeated measurements (general estimating equations, GEE) consisting of subject and within subject variables. The model was chosen along the research questions and with respect to all two-factor interactions with high sample power. Instead of standard deviations, the broader error bars with confidence intervals (CI-95%) are shown in the descriptive profile plots.

Detailed Assessment of B/A-Question Pairs

In the paired question comparison stage, SPSS's Kolmogorov-Smirnov and Shapiro-Wilk (Yazici & Yolacan, 2007) tests for normality for question groups B00 to A30 were carried out first. Based on the normality test results, the Friedman ANOVA followed by post-hoc Wilcoxon-Signed-Rank tests for the outlined question pair differences (Table 3) and for **B/A** question pairs was used. Post-hoc and most conservatively, Bonferroni correction is alternatively given for a more conservative judgement in the narrative of results and discussion of this text (Armstrong, 2014). More sensitively and considering the independence of each question pair, the involved test groups are pre-selected and independent with their own H_0 and used only once each, i.e. not in a complete two-combination family of tests as in usual combined post-hoc testing, where there is a joint group null hypothesis. Still for the conservative reader, the strict Bonferroni levels p_B are also shown in the tables. Many important p-values for the affective question domain still were significant also with conservative levels.

The statistical software platforms were IBM SPSS 26 (Armonk, NY, USA)[5] for testing and descriptive plotting. JASP 0.11.1 was used for some descriptive plotting, where convenient. Linear model based descriptive 2D profile plots' error bars are CI-95%.

Results

Qualitative Observations

As expected, the teacher observed the students partially as passive listeners during the first half of the semester. They were occasionally playing on their smart phones. During small exam simulations carried out in both traditional and flipped format, the students took a big step in presentation skills, especially when the flipped-classroom was activated. They were shyer and unsecure dealing with the small exam simulations in the first part of the semester. When actively integrated in the course of

the lectures, the use of smart phones and the shyness vanished. They developed a keen interest in providing their share to the lectures without the teacher interrupting.

One student's significant and astonishing quote in one of the first FC lectures was:

> "Please do not interrupt my talk in the middle, I have my own way of presenting these thoughts and I have been preparing and thinking through this at home."

On this occasion, the teacher was happy to hear about the commitment of this student and from now on only jumped in at the end of the student's presentation, eventually triggering a small discussion or adding some missing details if needed. Note, the need for discussion or additions diminished towards later lectures in the FC half of the semester as the students grasped the elements important to the teacher quickly. Finally, even difficult slides with up to one-minute video footage or difficult system overviews were presented fearlessly and successfully by the students.

In the first FC lecture (7), the TAP interview with vote counting **(N)** was answered by students as follows (abstentions were not counted)

T00. What helps learning most in this course?

(4) "Interactive lecture."

(2) "Small group size."

(2) "Practical examples."

(2) "Discussions are also possible during the lecture."

T01. What makes your learning more difficult?

(3) "Slides as the only way to prepare."

(2) "Often no references."

(2) "Different use of (technical) terms.

(2) "Sometimes incomprehensible explanations."

T02. What suggestions do you have for improving on the obstacles?

(3) "(Lecturer should) indicate own sources (in the handout slide material)."

(3) "Link to further material (clickable link)."

(2) "Lecturer should step in if not all information has been given by the student."

S04. What do you think about using FCL right from the beginning of the lecture term?

(3) "Without additional material some slides are too much (difficult)."

(2) "Flipped-Classroom also means more time and less content."

(2) "Too much workload (if fellow students are absent)."

Quantitative Analysis

Main Effects Analysis

The final significant variables and interactions kept in the GEE model were:

FC,

B/A,

Question,

Lecture

and the interactions:

FC * B/A,

FC * Question,

FC * Lecture,

BA * Question.

The results show significance and corresponding high Wald-Chi-Squares (effect weights) for the main effects and their interactions (Table 4).

The questions with index >9 (A) were filled in after each lecture, others (B) before. Generally, there is a tendency to better Likert scores towards the last question (Figure 4).

Interactions with **Question** were significant (Table 4). Thus, in conclusion there are significant

effects for **Question**. Individual question pairs will be compared later in this section for the factors **FC** and **B/A** (cf. details in Table 5 and Table 6).

Flipped-Classroom (FC) and Interactions with other Variables

The comparison of frontal vs. flipped-classroom teaching (**FC**=Y) shows highly significant results (Table 4, Figure 5 and Figure 6). This is also detailed post-hoc per question in Table 5 and Table 6.

More importantly as main effects are outshined by interactions, the interaction of **FC*B/A** was significant (Table 4). The profile plots in Figure 5 and Figure 6 with CI-95% error bars show a strong uptrend and angulation (interaction) with increased confidence (smaller error intervals).

The interaction of **FC*Lecture** is significant (Table 4). The intergroup separation is shown in Figure 7. **FC*Question** was significant, too and the interaction profile is shown in Figure 8. So, individual questions were supposedly answered much better with **FC**=Y.

The remaining interaction **B/A*Question** (Figure 9) was significant, but is not in the research focus of this work and is rather only kept for significant effect model formulation completeness.

B/A

There is significance to higher scores after (A) the lecture, especially interacting with **FC**=Y. This is shown in Figure 5. Variable interactions show significance (Table 4 and Figure 5). Concluding and obviously for the variable itself, there is a significant effect for B/A.

Lecture and Question

FC lectures, were rated differently by the subjects. This is depicted in Figure 3. There is a clear trend towards better scores with FC teaching. Interaction with **Lecture** and the variable itself is significant. **Question** was involved in significant interactions, so it is a significant factor even if it is not as a main effect in Table 4 (Carey, 2013).

Comparison of Individual B/A Question Pairs and Splitting by FC=Y/N

The question group data samples were tested for normality with the Kolmogorov-Smirnov and Shapiro-Wilk test which both failed highly significantly. Friedman ANOVA yielded $p<0.001$, so significantly differently answered question pairs were present. Wilcoxon-Signed-Rank tests were used as post-tests (not a post-hoc test family) for all **FC** question pairs (Table 5) and selected cross test pairs as shown in Table 3 (Table 6). Bonferroni correction $p_B<0.05/10$ is provided for conservative readers (Armstrong, 2014). As there is an independent one to one mapping of groups, i.e. by the predefined pairing, every group is used only once in the comparisons (Table 6). So, the results are shown in Table 6 also with normal significance levels. The layout of Table 6 corresponds cell-wise to Table 3, augmented by three-fold in-cell information in separate lines:

1st cell line: overall B/A difference,

2nd cell line: B/A subgroup difference with **FC**=N, and

3rd cell line: B/A subgroup difference with **FC**=Y.

Some question pairs show significant differences before vs. after the lecture (Table 6).

Further splitting the total group into **FC**=N and **FC**=Y subgroups, yields a Friedman ANOVA result of $p<0.001$ for the total and the **FC=J** group. Afterwards, different answers B01-A11, B07-A20 are detected by Wilcoxon-Signed-Rank tests. For the **FC**=Y subgroup, a $p<0.001$ is achieved with the Friedman ANOVA test. As detected by Wilcoxon-Signed-Rank tests, differences B02-A12, B03-A16, B04-A17, B05-A18, B06-A19, B07-A20, B08-A21 and B09-A24 are significant (all except the first two question pairs). This is summarized in Figure 9 and Table 6.

Discussion and Conclusion

The transformation of frontal to flipped classroom teaching was a success story and allows for fruitful comparison of the two teaching formats.

Obviously for **Question**, different questions provoked different Likert responses which is due to the different topic of the questions and the interactions, such as **Question*B/A** (Figure 4, Figure 5 and Figure 9). **Lecture** was a significant factor, also due to its significant interactions, too. As the content of the lecture was different and also the format varied, clearly this is a significant factor. More relevant than the conclusions regarding the main effects are the interactions itself.

The main research question here regards the factor **FC** and its interactions. In Table 5 the significance levels are shown for the classrooms with **FC**=Y vs. **FC**=N (cf. Figure 8). Bonferroni correction $p_B<0.05/31$ is only optionally applied here (Armstrong, 2014), as there is an independent one to one mapping of groups, i.e. by the predefined pairing, every group is used only once in the set of comparisons. B00 to B02 as well as A10 to A13 and A28 concerning the learning goals or the "red thread" show significance. A-questions were answered slightly better. In this part of the study, the interpretation is that the FC format activates the students' positive course experience and their own thinking about the didactic targets, even if the learning goals discussion was kept in both formats. B03, B06 as well as A16 to A19 (n.s., but more confident votes for **FC**=Y in Figure 8) being about future benefits show uptrends, too. Hence, the feeling of the students with regards to FC lectures is positively inclined. A13 and A15 indicate that the FC lecture was better tailored to the learning goals and prepared the students better for the homework. A20 was about the skill to convey the learning goals to others and in an exam situation.

Moreover, significance was found for all interactions with **FC**. This leads to the important conclusion of **FC** itself and its interactions being significant model effects (Table 4). The Wald-Chi-Square weight of **FC** is also dominant in this analysis. The significant items A23 and A24 reflect on the didactic methods used and their explanation by the teacher or students' on-the-job grasping their sense. With the mind-activating FC format, this could be successfully achieved easily. The paired question of A24, i.e. B09 in Table 6, was not significant as being asked before the lecture and

382 students being indifferent at this time (Likert: 3 or 4). The questions concerned about didactics, i.e.
383 the learning environment and the emotional atmosphere: A25 to A26, showed significant benefits
384 for the FC format. Thus overall, the new format was well received by the students also in terms of
385 human factors (soft-skills). A27 and A29 (n.s., but confident vote for **FC**=Y in Figure 8) confirmed
386 that a trend towards feeling burdened by switching to the new format was perceived (A29). This
387 could be attributed to the perceived increased workload in students' activated minds compared to
388 TT.

389 The interactions analyze the inclines in Likert scores depending on joint variables. The
390 obvious side interaction without an important research question in this paper **B/A*Question** was
391 significant and thus was kept in the model fit (Figure 9). The important interaction **FC*BA** was
392 significant (Table 4, Figure 5 and Figure 6). This could be attributed to the increased confidence
393 (smaller error intervals) with **FC**=Y and the higher knowledge gain after a lecture at the same time.
394 Second, along the research lines, the interaction **FC*Lecture** stands out (Table 4, Figure 7). The
395 significant interaction **FC*Question** contributes with a relatively low interaction to the model
396 (Table 4, Figure 8). Figure 8 shows many upwardly directed and crossing connections, and some
397 significant differences for individual questions were indicated (Table 5). The negative monitoring
398 question A29 is the only item resisting in the lower Likert values. So, the students received the new
399 format, when better knowing it, as a burden. This fact could be attributed to behavioral change
400 towards a new kind of teaching, some starting problems as indicated in the qualitative TAP results
401 and the necessary rewiring of students' brains – stated exaggerated: FC is not watching TV with
402 learning procrastination to the end of the semester.

403 Looking at Table 6 with selected pair-wise comparisons of question pairs using rank-based
404 methods and Figure 9, there are many significant differences found for **B/A**, even if Bonferroni
405 levels apply (Table 6: First cell line #s). This attracted further investigation, especially for the

406 subgroups concerning **FC**. B00-A10 was not significant overall as the discipline of announcing
407 learning goals and discussing them at the end of the lecture was kept in both teaching formats.
408 Except the first two questions pairs, all paired items were significant and the task was to clarify what
409 was the origin of the significance. Further splitting the groups into **FC**=N and **FC**=Y subgroups
410 revealed more details. While in the subgroup **FC**=N, **B/A** comparisons were only rarely significant
411 (B1-A11, B7-A20), consistent significance was detected for the **FC**=Y, **B/A** comparisons except the
412 first two pairs (third cell line). The interpretation here was that the factor **FC** causes a steep incline
413 in Likert scores as also can be seen in Figure 6, Figure 7 and Figure 8. Thus, supposedly the personal
414 gain for the students was clearly higher using the FC teaching format.
415 This can also be discussed looking qualitatively at some of the individual **B/A** cross
416 compared questions:
417 B02-A12: The learning objectives are firmly anchored in my memory.
418 There was a significant difference before and after the lecture, especially when the flipped-
419 classroom format was used. This supposedly means, FC helped the students to memorize and think
420 about the teaching objectives in a deeper way.
421 B03-A16: I feel I am well prepared for the exercises.
422 B04-A17: I have the feeling that I am well prepared for the homework.
423 B05-A18: I feel I am well prepared for the exam.
424 B07-A20: I can convey the content to others, especially in the exam situation.
425 The felt preparation level for exercises mainly from the affective question category,
426 homework and final exam is significantly higher with **FC**=Y. This is a good sign of the effectively
427 increased and active thinking triggered in the students. For B7-A20 and **FC**=N, there was
428 significance. So, also the traditional lectures did a good job. However, for **FC**=Y highly and

Bonferroni significant differences showed out. This again underlines the better impact of FC lectures for the final exam situation.

B06-A19: I have the feeling that the event has brought me something for the future (e.g. further study, job, personality development).

B09-A24: I can profitably take didactic methods from the event for my (possibly) own teaching (e.g. as a tutor, later as a lecturer)

The perceived use of the lecture is clearly higher for FC lectures. The conveyed skills do not only cover the reproductive and deductive use of a knowledge base, but also soft skills such as team work, presenting difficult technical topics in front of a group and discussion with a group. There was also a significant effect for FC classrooms regarding the teaching of the new didactic methods just by using them and explaining them occasionally.

B08-A21: I think about my own learning and control my learning activities myself.

While in a passive **FC**=N setting, no self-monitoring of learning activities was triggered before vs. after the lecture, there is a significant impact for FC classrooms. This proves the activation of students once more.

As indicated, Bonferroni correction or similar post-hoc test methods were only optionally applied here to make the discussion more explanatory and sensitive (lively) vs. the observed effects. The adjustment of significance levels for test sets is part of current discussions in the statistics research community (Armstrong, 2014). The conservative reader might want to apply the Bonferroni levels, which are provided as well in Table 5 and Table 6 (#).

From the teacher's and a student's perspective, the students' amount of (home-)work was greater than in traditional lectures (A29) – classroom preparation was added to the regular homework exercises. In the author's opinion, a fine-tuned balance has to be found here and it might be advisable to reduce the regular homework load inherited from a TT format during a

transformational semester. In this study, many students got their admission for the oral exam from the reduced homework results in the middle of the semester.

In conclusion, the comparison of a traditional to a flipped-classroom was a success. Remarkably, the students were very motivated and once insisted that the teacher does not interrupt during their presentation to finish their delivery. Statistical significance for the main factors and their interactions was proven using a repeated-measurements generalized linear model (GEE) with a correction term for "ordinal logistic" scales (i.e. the Likert scale) regarding questions mainly from the cognitive (facts) and affective levels (emotions) vs. the introduced FC teaching. Significant factor constituents were then analyzed in detail with rank-based statistical methods as the normality assumption for general linear models (ANOVA) was hurt. Finally, the FC classroom format reaches out to offer significant benefits, i.e. increased mind activation levels, to the students, possibly even if a mixed format lecture course (traditional & flipped) is offered. In this class of course, a quality driver was the final oral exam at the end of the semester.

For the future, it would be interesting to implement the lecture with larger scale classrooms, to investigate the topic of pure FC format lectures more closely, and which results show with written exams at the of the semester violating constructive alignment, which is sometimes inevitable.

Acknowledgements: Bettina Jansen-Schulz (TransferConsult science consulting) and Anke Timmann (TAP interviews), Linda Brüheim (questionnaire assist), and SoTL project group, especially Daniel Wiswede and Jonas Obleser, all also affiliated to University of Luebeck, Germany.

Footnotes

[1] https://www.maastrichtuniversity.nl/education/why-um/problem-based-learning

[2] https://mdprogram.mcmaster.ca/mcmaster-md-program/overview/pbl---problem-based-learning

[3] STEM: Science, Technology, Engineering, and Mathematics academic disciplines

[4] c.t.: from latin, cum tempore, 15 minutes, colloquial "the academic quarter"

[5] https://www.ibm.com/analytics/spss-statistics-software

[6] https://www.cogniview.com/pdf-to-excel/pdf2xl-ocr

[7] https://tatp.utoronto.ca/teaching-toolkit/ci-resources/cdg/lesson-design/active-learning/

Tables

Table 1

Part I of the questionnaire asked before (B) the lecture. Colour coding: blue: cognitive level (facts); yellow: affective level (emotions); green: didactic level (learning environment).

Lecture No.: _____ Flipped-Classroom: y / n Pseudonym:_____	fully [illegible]	disagree	rather [illegible]	rather	agree	fully agree
Likert-value	1	2	3	4	5	6
B00. The learning objectives were defined together at the beginning of the lecture.						
B01. I can explain the learning goals in my own words.						
B02. The learning objectives are firmly anchored in my memory.						
B03. I feel I am well prepared for the exercises.						
B04. I have the feeling that I am well prepared for the homework.						
B05. I feel I am well prepared for the exam.						
B06. I have the feeling that the event has brought me benefits for the future (e.g. further study, job, personality development).						
B07. I can convey the content to others, especially in the exam situation.						
B08. I think about my own learning and control my learning activities myself.						
B09. I can profitably use didactic methods from the event for my (possibly) own teaching (e.g. as a tutor, later as a lecturer).						

489 Table 2

490 The questionnaire part II asked after (A) the lecture.

Lecture No.: _____ Flipped-Classroom: y / n Pseudonym:_____	fully ...	disagree	rather ...	rather	agree	fully agree
Likert-value	1	2	3	4	5	6
A10. The learning objectives were discussed at the end.						
A11. I can explain the learning goals in my own words.						
A12. The learning objectives are firmly anchored in my memory.						
A13. The lecture is tailored to the educational objectives.						
A14. The learning objectives are covered or mirrored in the exercises and home exercises.						
A15. The homework exercises were solvable with the material of the lecture and exercise (without using google).						
A16. I feel I am well prepared for the exercise.						
A17. I have the feeling that I am well prepared for the homework.						
A18. I feel I am well prepared for the exam.						
A19. I have the feeling that the event has brought me benefits for the future (e.g. further study, job, personality development).						
A20. I can convey the substance to others, especially in the exam situation.						
A21. I think about my own learning and control my learning activities myself.						
A22. The used didactic methods were explained.						
A23. I know why certain didactic methods were used.						
A24. I can profitably use didactic methods from the event for my (possibly) own teaching (e.g. as a tutor, later as a lecturer).						
A25. The atmosphere has encouraged learning.						
A26. The environmental conditions were appropriate for the learning of the group.						
A27. The learning success is in good proportion to the extent of self-activity.						
A28. The entry into the lecture is facilitated by an overview ("red thread").						
A29. The flipped-classroom method is not an additional burden for me.						
A30. Flipped-classroom helps me prepare for the oral exam format.						

491

492 Table 3

493 The corresponding items asked before (B) vs. after (A) each lecture.

A \ B	0 The learning objectives were defined together at the beginning of the lecture.	1 I can explain the learning goals in my own words.	2 The learning objectives are firmly anchored in my memory.	3 I feel I am well prepared for the exercises.	4 I have the feeling that I am well prepared for the homework.	5 I feel I am well prepared for the exam.	6 I have the feeling that the event has brought me something for the future (e.g. further study, job, personality development).	7 I can convey the content to others, especially in the exam situation.	8 I think about my own learning and control my learning activities myself.	9 I can profitably use didactic methods from the event for my (possibly) own teaching (e.g. as a tutor, later as a lecturer).
10	x									
11		x								
12			x							
16				x						
17					x					
18						x				
19							x			
20								x		
21									x	
24										x

494

Table 4

Tests of Model Effects conducted with general linear model (GEE) and repeated measurements. p<Sig. Table from IBM SPSS 26.

Tests of Model Effects

	Type III		
Source	Wald Chi-Square	df	Sig.
FC	56.247	1	.001
BA	6.632	1	.010
Question	.100	1	.752
Lecture	12.295	1	.001
FC * BA	17.244	1	.001
FC * Question	9.548	1	.002
FC * Lecture	32.882	1	.001
BA * Question	4.770	1	.029

Dependent Variable: Likert score

Model: (Threshold), FC, BA, Question, Lecture, FC * BA, FC * Question, FC * Lecture, BA * Question

Table 5

Comparison of questions for the factor **FC**=Y vs. N using Wilcoxon-Signed-Rank tests. N.s.: not significant; *: $p<0.05$; **: $p<0.01$; $p_B<0.0016$ (two-tailed).

Q \ FC	$p<$	Significance
B00 - B00	0.017	*
B01 - B01	0.011	*
B02 - B02	0.017	*
B03 - B03	0.018	*
B04 - B04	0.058	n.s.
B05 - B05	0.058	n.s.
B06 - B06	0.028	*
B07 - B07	0.14	n.s.
B08 - B08	0.079	n.s.
B09 - B09	0.147	n.s.
A10 - A10	0.011	*
A11 - A11	0.033	*
A12 - A12	0.007	**
A13 - A13	0.035	*
A14 - A14	0.084	n.s.
A15 - A15	0.011	*
A16 - A16	0.016	*
A17 - A17	0.026	*
A18 - A18	0.016	*
A19 - A19	0.009	**
A20 - A20	0.026	*
A21 - A21	0.119	n.s.
A22 - A22	0.071	n.s.
A23 - A23	0.023	*
A24 - A24	0.048	*
A25 - A25	0.014	*
A26 - A26	0.018	*
A27 - A27	0.033	*
A28 - A28	0.016	*
A29 - A29	0.317	n.s.
A30 - A30	1.000	n.s.

Table 6

The corresponding before (B) vs. after (A) items and post test result: n.s.: not significant; *: $p<0.05$; **: $p<0.01$; #: $p_B<0.005$ (Bonferroni); ***: $p<0.001$ (two-tailed). First cell row: overall B/A; second row: **FC**=N subgroup; third row: **FC**=Y subgroup. Statistical test results with IBM SPSS 26.

A \ B	0	1	2	3	4	5	6	7	8	9
10	n.s. n.s. n.s.									
11		n.s. * n.s.								
12			***# n.s. **							
16				**# n.s. **						
17					***# n.s. ***#					
18						***# n.s. ***#				
19							***# n.s. **#			
20								***# * **#		
21									** n.s. **	
24										* n.s. *

Figures

Bloom's Learning Pyramid Turned over:

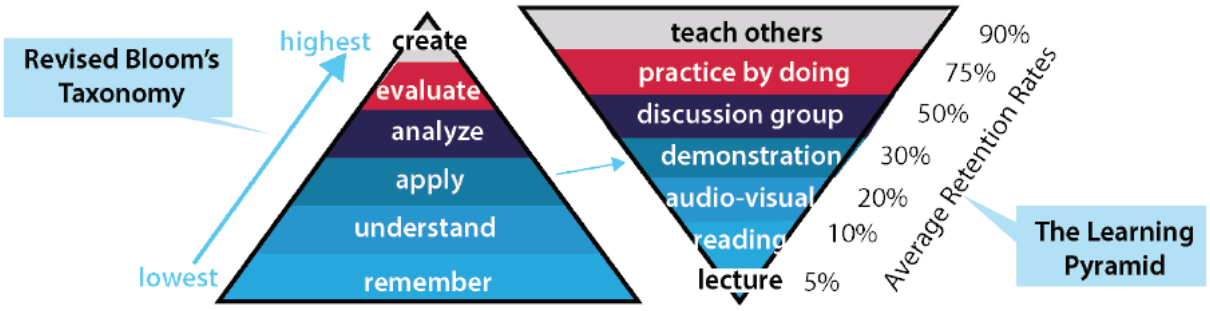


Figure 1: Bloom's taxonomy revisited[7] (Bloom, 1965, 1984). Students in action promote better recall of knowledge. The active learning method "FC teaching" exactly fosters the beneficial areas as well as the approach of constructive learning (Piaget, 2013), where students are responsible how they "do" achieve the goals under some supervision.

Flipped-Classroom Voting Poll Set-up:

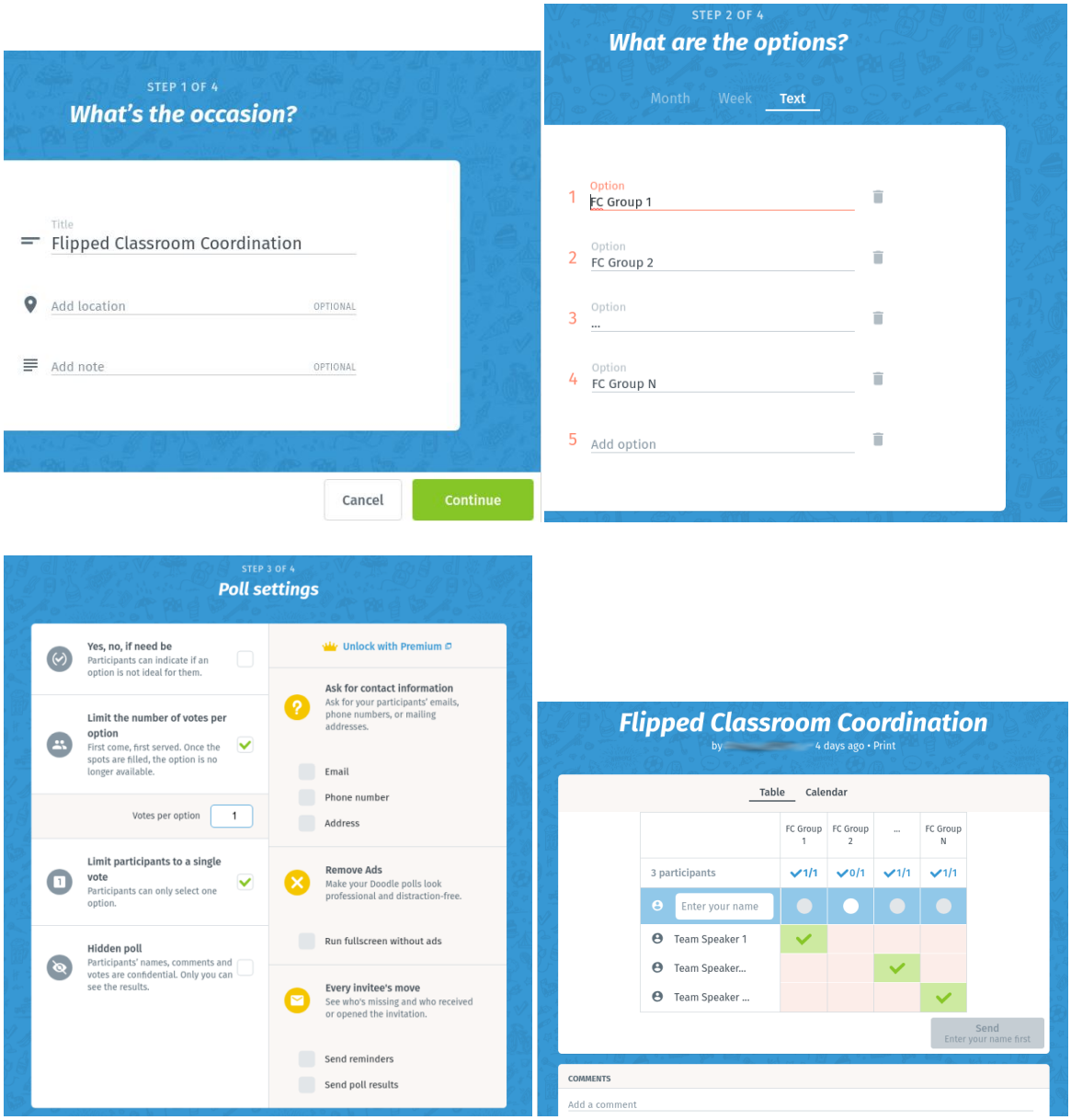
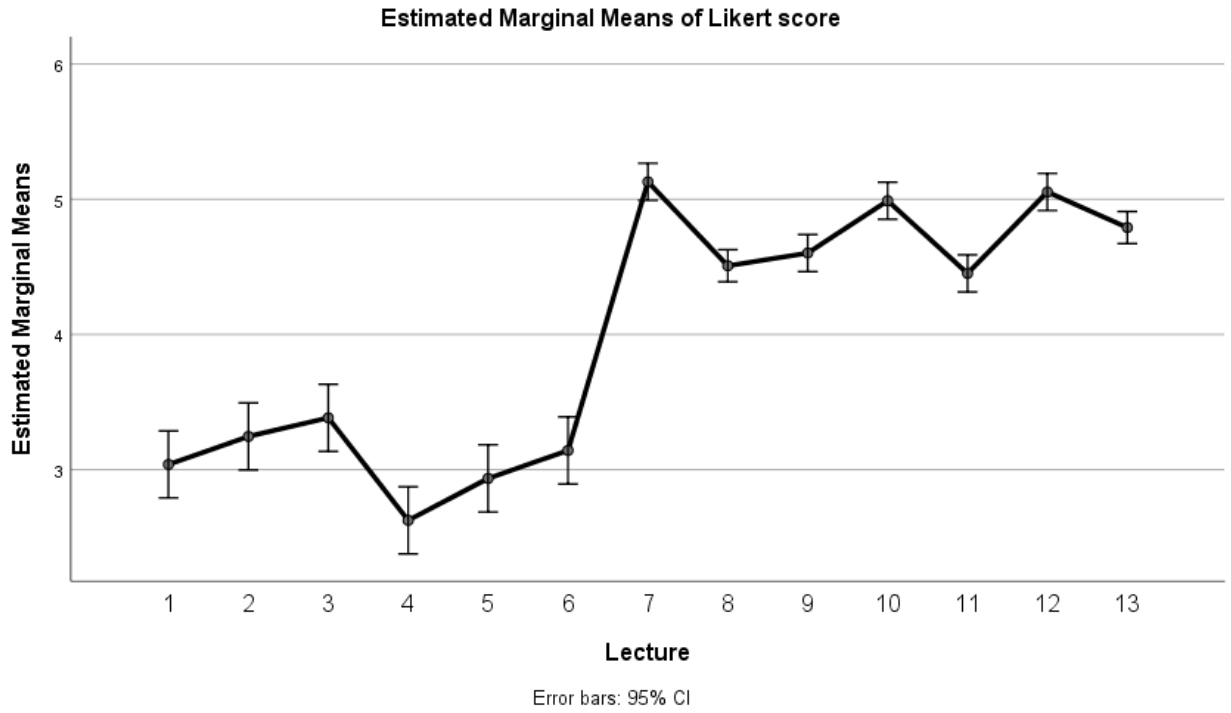


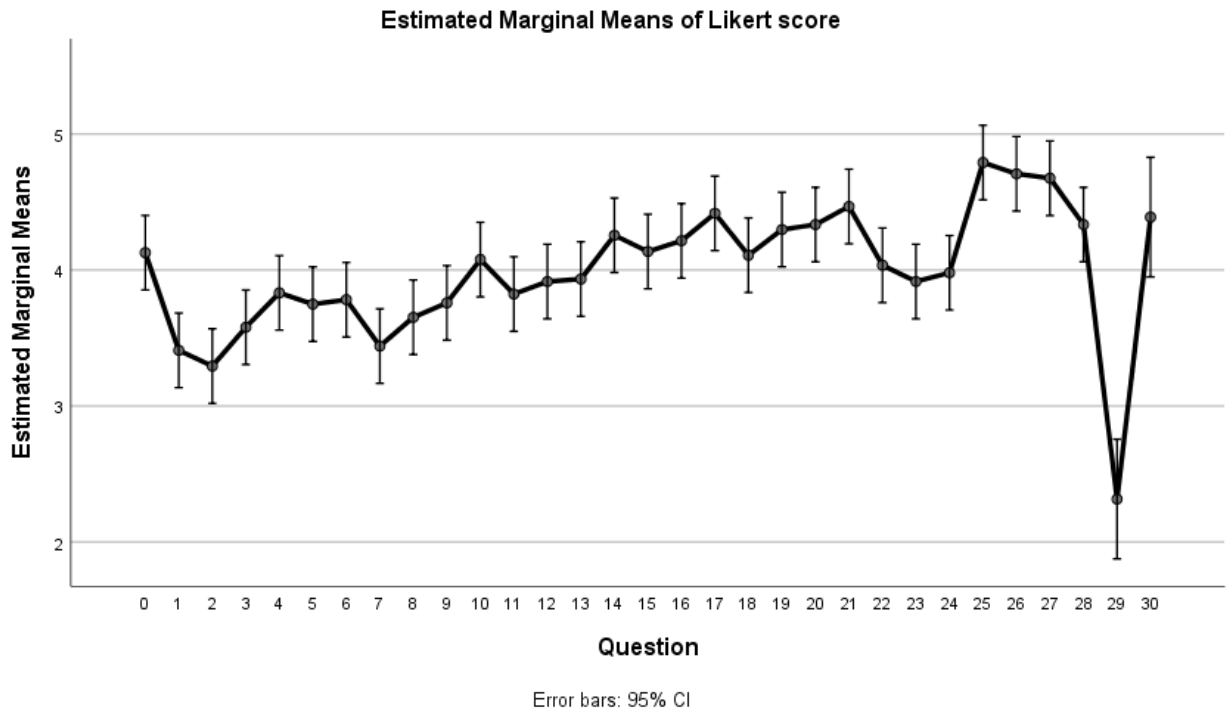
Figure 2. flipped-classroom voting pol set-up with doodle.com. Reading direction: Top left to bottom right. The link to the poll: https://doodle.com/anonymous can be used in the invitations by email to the team speakers of the study groups.

524 Likert Scores during the Course of Lectures:



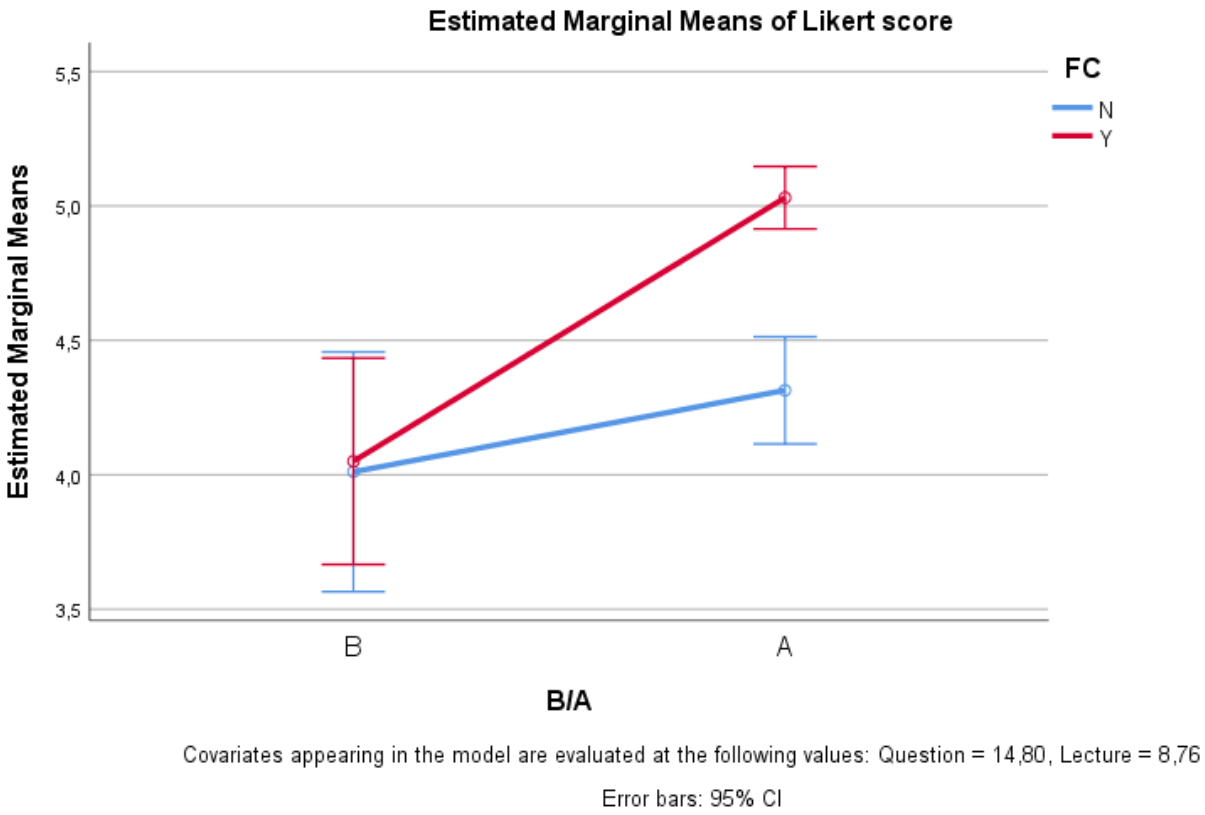
525
526 *Figure 3*. Likert scores provided after the FC started with lecture 7. There is a trend to better Likert scores to
527 the end of the semester. Underlying model from *Table 4*. Plots are done with JASP.

528 Likert Score vs. **Question**:



529
530 *Figure 4*. Questions answered before and after (>9) each lecture. There is a mild trend towards higher scores
531 at the end of the lectures except negative question A29, which aims at FC as additional burden for students.
532 Underlying model from *Table 4*. Plots are done with JASP.

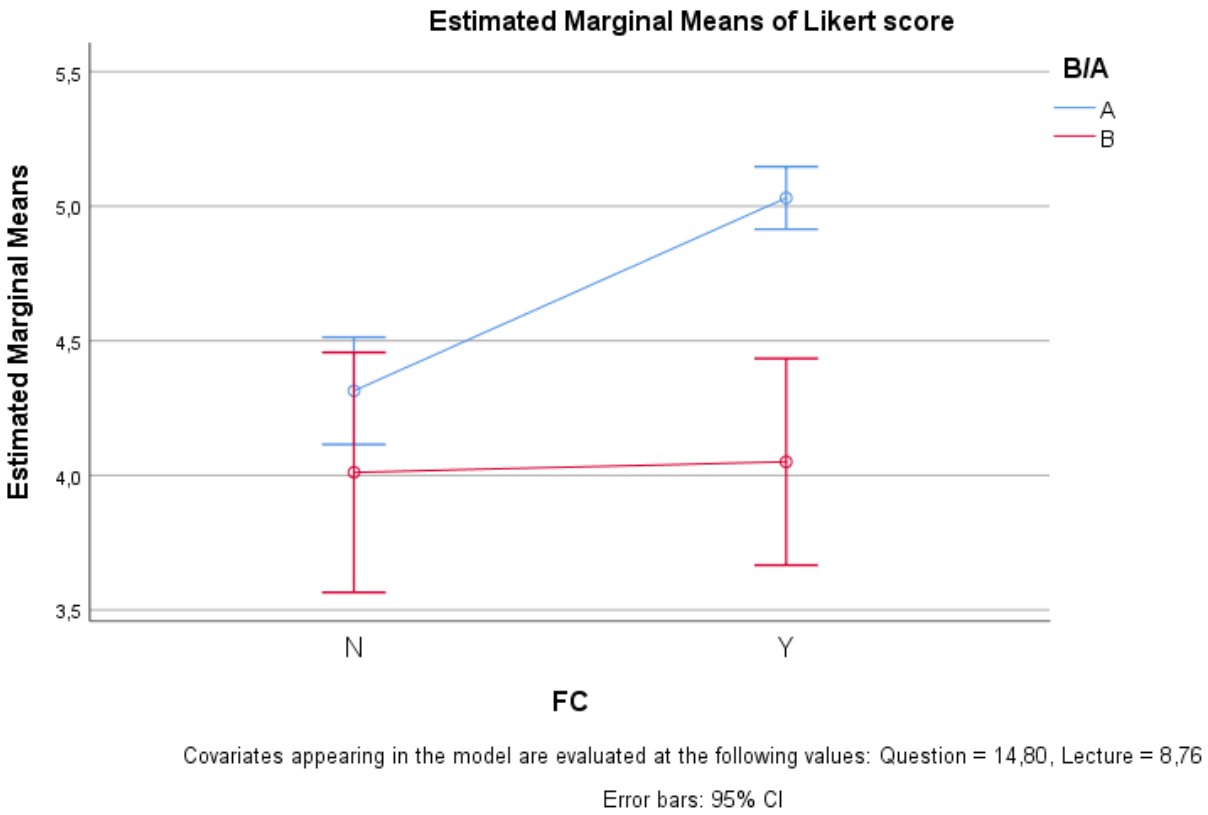
533 The Interaction Plot of the Factors **B/A*FC** vs. Likert Score:



534

535 *Figure 5*. Likert scores provided after the before/after (B/A) the lecture. There is a trend to better Likert
536 scores at the end of a lecture and an interaction with FC. Underlying model from *Table 4*. Plots are done with
537 SPSS.
538

539 The Interaction Plot of the Factors **FC*B/A** vs. Likert Score:



540
541 *Figure 6*. Significant interaction of factors **FC*B/A**: In the generalized linear model preferred here, there is a
542 trend to better Likert scores for FC lectures and after a lecture. White dots=0=before the lecture. Black
543 dots=1=after the lecture. Underlying model from *Table 4*. Plots are done with SPSS.

Descriptive plot of the Factors **FC*Lecture** vs. Likert Score:

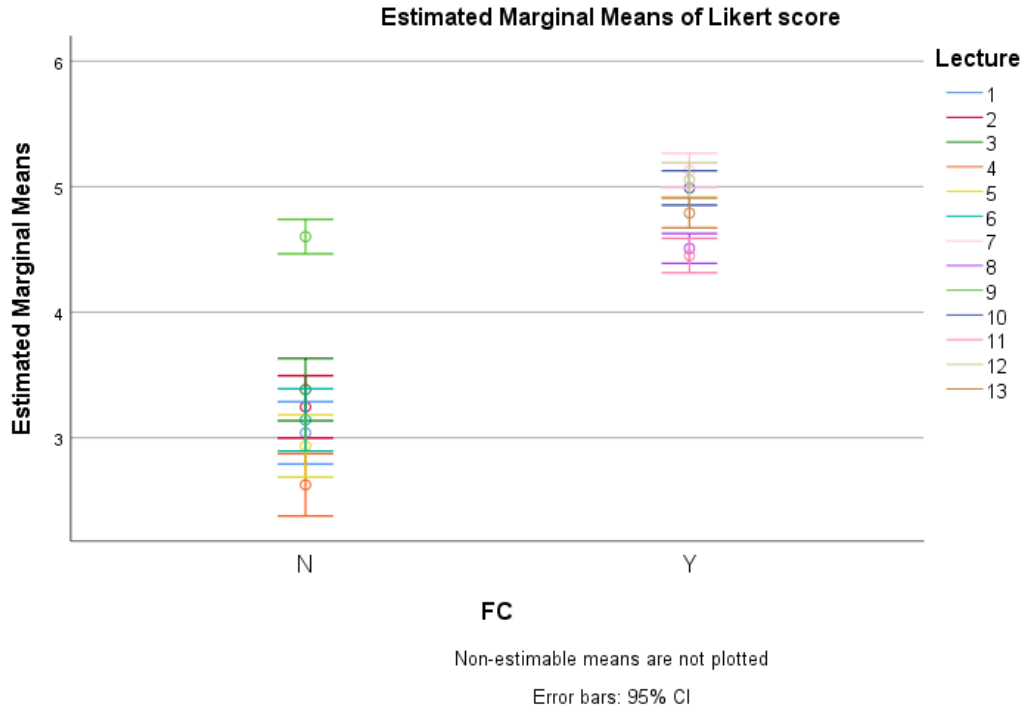


Figure 7. Significant changes for the variable **FC*Lecture**. There is a strong trend to better Likert scores for FC lectures. Lecture 9 (grass green) was given in TT format, once in the FC period. Underlying model from *Table 4*. Plots are done with SPSS.

Descriptive Plot of the Factors **FC*Question** vs. Likert Score:

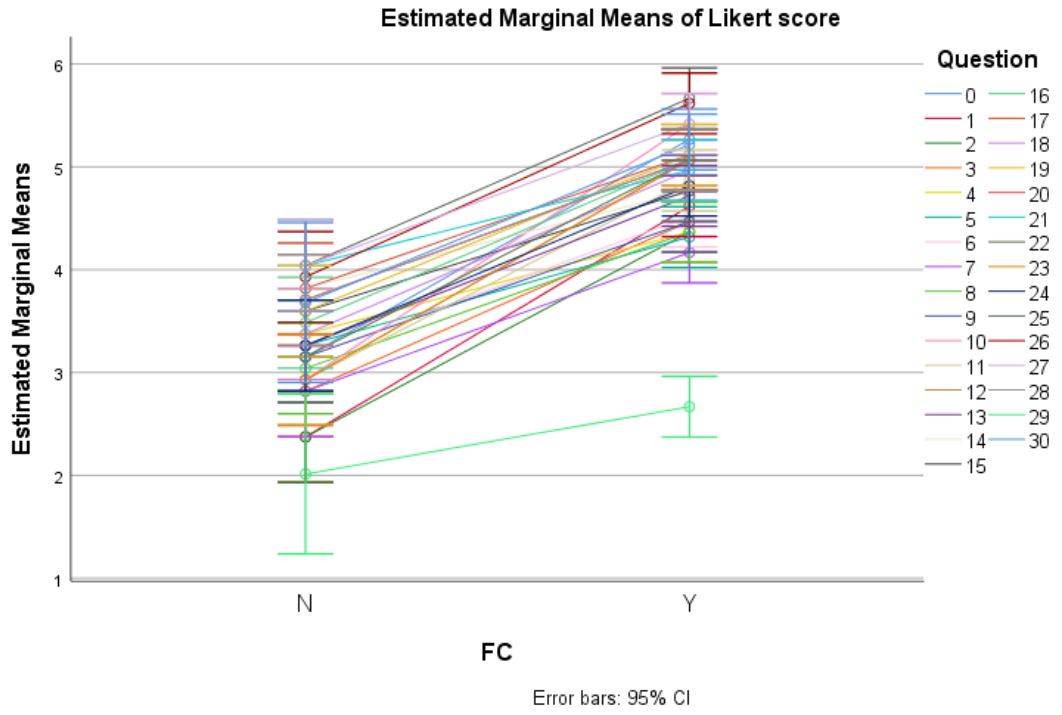


Figure 8. Significant influence of factor combination **FC*Question**. In the generalized linear model preferred here, there is a trend to better Likert scores for FC lectures. Only the negative question A29 is a low outlier. Underlying model from Table *4*. For uncluttered plots, see Appendix *Figure 10*. Plots are done with SPSS.

Descriptive Plot of the Factors **B/A*Question** vs. Likert Score:

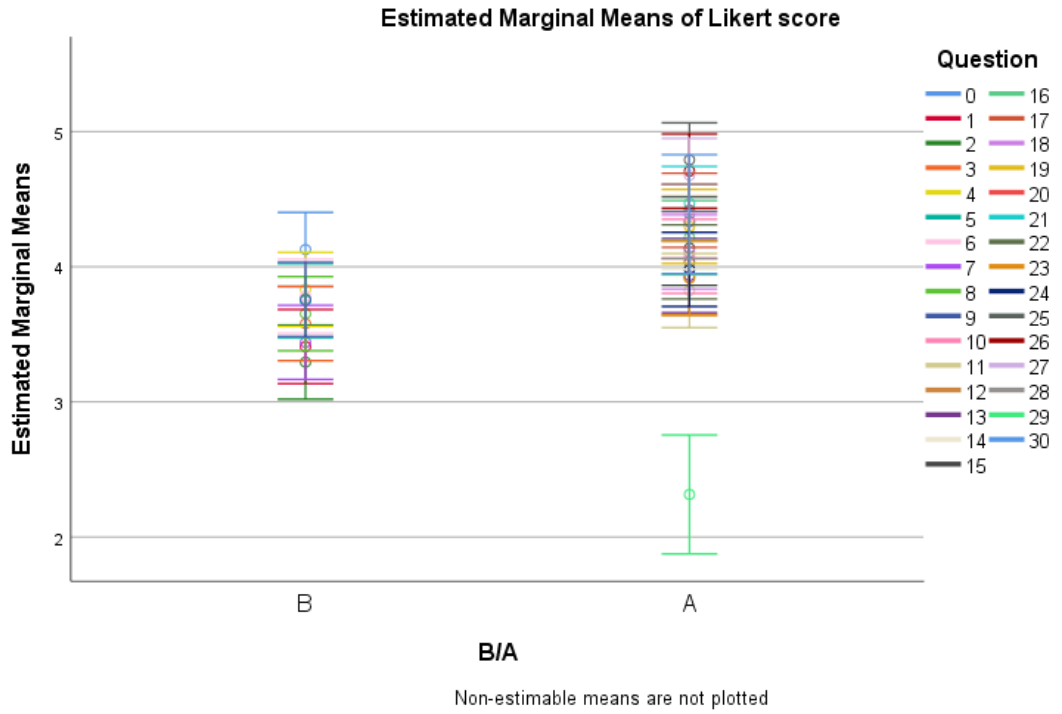


Figure 9. Significant influence of variables **B/A*Question** augmented by **FC** in the lower diagrams. In the linear model, there is a trend to better Likert scores after lectures. Confidence intervals indicate the spread of the Likert scaled answers to the questions (vertical lines) in the per question stacked plots. The bottom right outlier A29 (green) was found again to be differently answered here. Underlying model from Table *4*. For uncluttered plots, see Appendix *Figure 11*. Plots are done with SPSS.

Appendix

Detailed descriptive Plots of the Factors FC*Question vs. Likert Score:

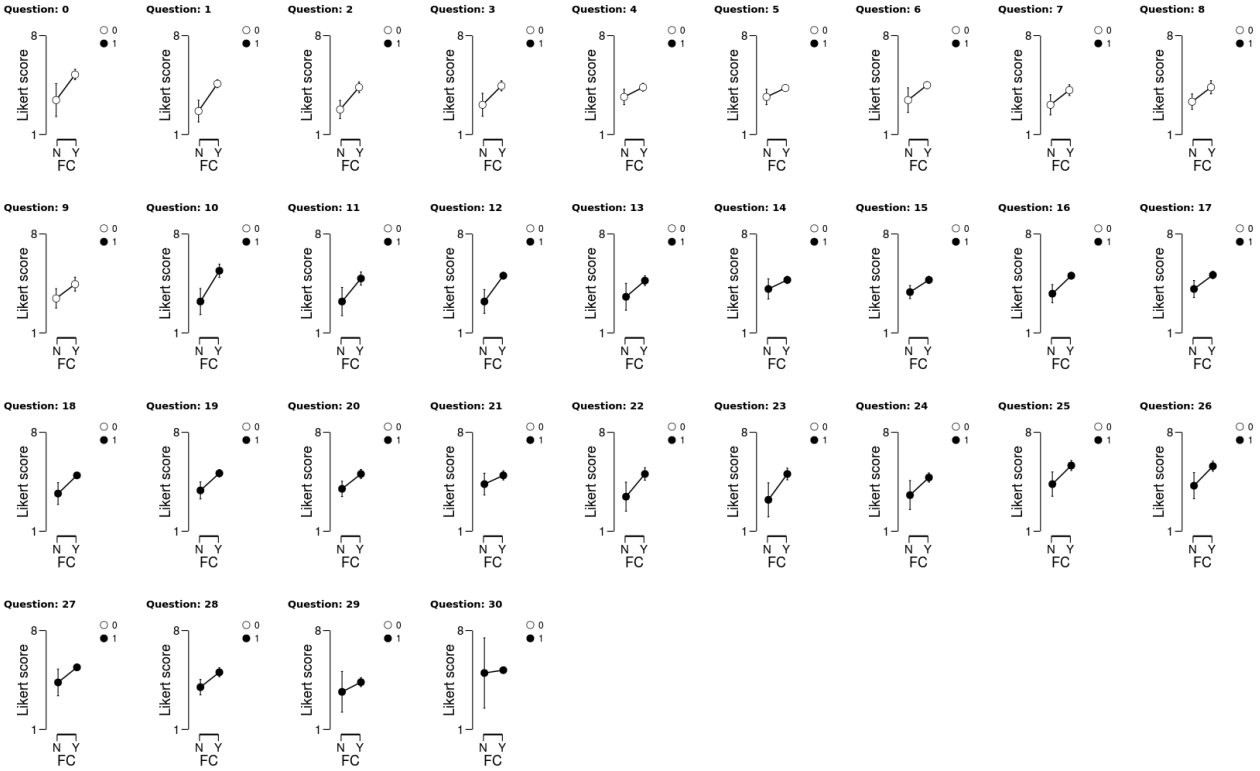


Figure 10. Significant influence of factor combination **FC*Question**. In the generalized linear model (GEE) preferred here, there is a trend to better Likert scores for FC lectures: $p<0.02$. Confidence intervals are omitted in the overview above to increase readability of the plot. Plots are done with JASP.

Detailed descriptive Plots of the Factors B/A*Question vs. Likert Score:

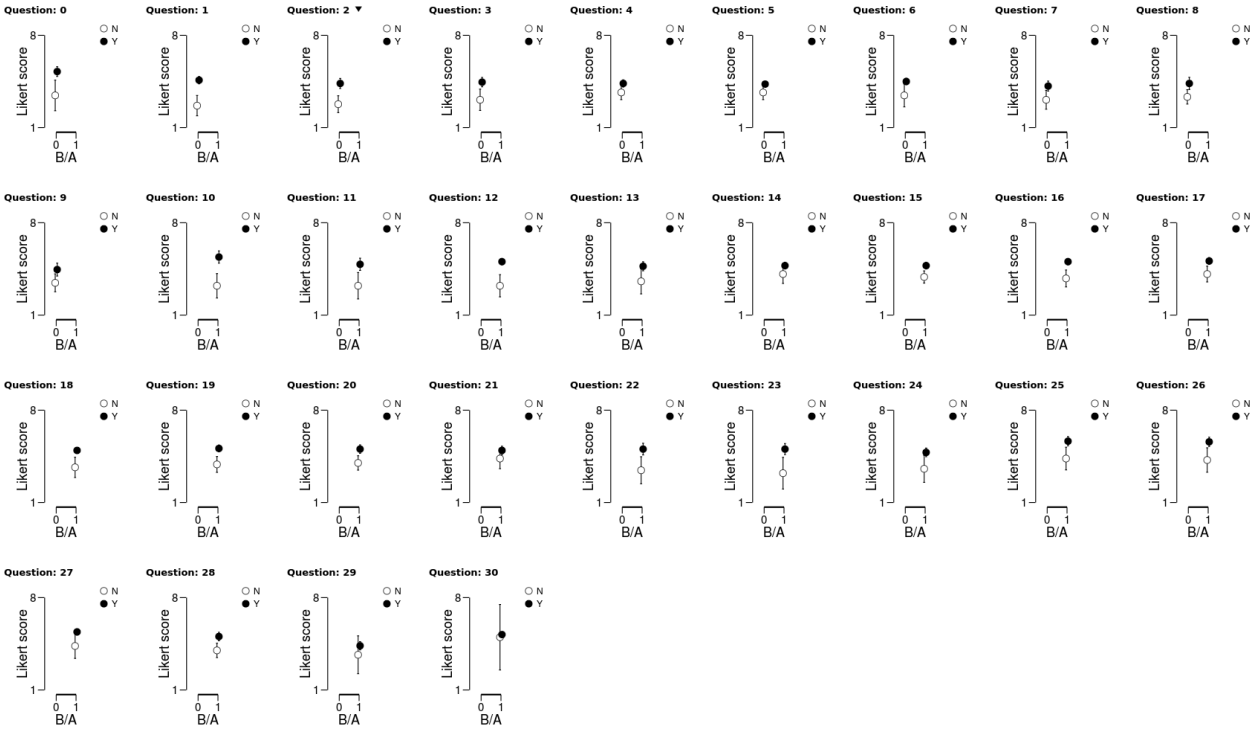
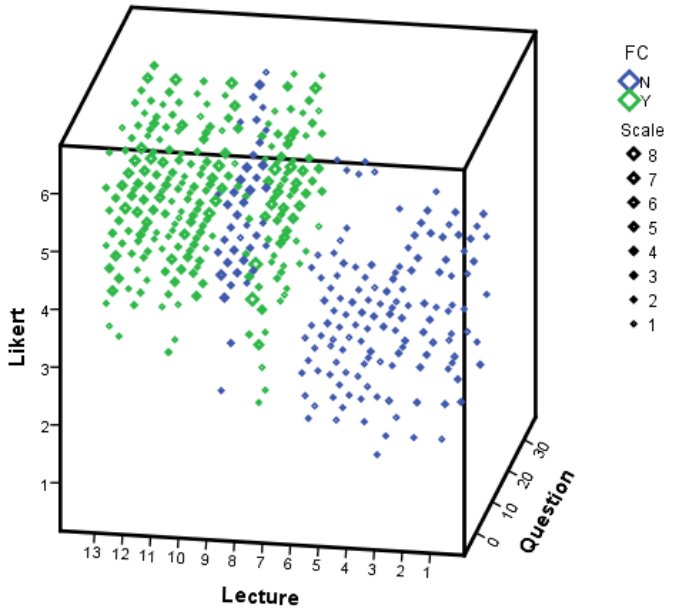


Figure 11. Significant influence of variables **B/A*Question** augmented by **FC** in the lower diagrams. In the generalized linear model (GEE), there is a trend to better Likert scores after lectures: $p<0.029$. Confidence intervals (omitted above) indicate the spread of the Likert scaled answers to the questions (vertical lines) in the per question plots below. White dots: FC=N, black dots: FC=Y. The outlier B00 was not found to be differently answered to A10 (bordered circles). Plots are done with JASP.

578 3D Scatter Plot of the Variables Lecture vs. Question vs. Likert Score:



579

580 *Figure 12.* 3D scatter plot of the variables **Lecture** vs. **Question** vs. Likert score. Lecture 9 in traditional
581 teaching format inside the cluster cloud of the FC lectures is clearly visible. For better display, binning with 8
582 scale levels was used in SPSS.

583 References

584 Angelo, T. A., & Cross, K. P. (2012). *Classroom assessment techniques*: Jossey Bass Wiley.
585 Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological*
586 *Optics, 34*(5), 502-508.
587 Auster, E. R., & Wylie, K. K. (2006). Creating active learning in the classroom: A systematic
588 approach. *Journal of Management Education, 30*(2), 333-353.
589 Bloom, B. S. (1965). *Bloom's taxonomy of educational objectives*: Longman.
590 Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as
591 effective as one-to-one tutoring. *Educational researcher, 13*(6), 4-16.
592 Carey, G. (2013). *Quantitative Methods in Neuroscience*.
593 http://psych.colorado.edu/~carey/qmin/qmin.php.
594 Crawford, K. (1996). Vygotskian approaches in human development in the information era.
595 *Educational studies in mathematics, 31*(1-2), 43-62.
596 Davidson, N., Major, C. H., & Michaelsen, L. K. (2014). Small-group learning in higher
597 education—cooperative, collaborative, problem-based, and team-based learning: an
598 introduction by the guest editors. *Journal on Excellence in College Teaching, 25*(3&4), 1-
599 6.
600 Fautch, J. M. (2015). The flipped classroom for teaching organic chemistry in small classes: is it
601 effective? *Chemistry Education Research and Practice, 16*(1), 179-186.
602 Hung, W., Jonassen, D. H., & Liu, R. (2008). Problem-based learning. *Handbook of research on*
603 *educational communications and technology, 3*, 485-506.
604 Hwang, G.-J., Lai, C.-L., & Wang, S.-Y. (2015). Seamless flipped learning: a mobile technology-
605 enhanced flipped classroom with effective learning strategies. *Journal of Computers in*
606 *Education, 2*(4), 449-473.
607 Jensen, J. L., Kummer, T. A., & Godoy, P. D. d. M. (2015). Improvements from a flipped
608 classroom may simply be the fruits of active learning. *CBE—Life Sciences Education,*
609 *14*(1), ar5.
610 Klegeris, A., & Hurren, H. (2011). Impact of problem-based learning in a large classroom
611 setting: student perception and problem-solving skills. *Advances in physiology education,*
612 *35*(4), 408-415.
613 Lawal, B., & Lawal, H. B. (2003). *Categorical data analysis with SAS and SPSS applications*:
614 Psychology Press.
615 Loeys, T., Cook, W., De Smet, O., Wietzker, A., & Buysse, A. (2014). The actor–partner
616 interdependence model for categorical dyadic data: A user-friendly guide to GEE.
617 *Personal Relationships, 21*(2), 225-241.
618 Mastmeyer, A., Fortmeier, D., & Handels, H. (2016). *Random forest classification of large*
619 *volume structures for visuo-haptic rendering in CT images.* Paper presented at the Proc.
620 SPIE Medical Imaging: Image Processing.
621 Mastmeyer, A., Wilms, M., Fortmeier, D., Schröder, J., & Handels, H. (2016). Real-Time
622 Ultrasound Simulation for Training of US-Guided Needle Insertion in Breathing Virtual
623 Patients. In *Studies in health technology and informatics* (Vol. 220, pp. 219): IOS Press.
624 Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale
625 items? Study I: Reliability and validity. *Educational and psychological measurement,*
626 *31*(3), 657-674.

Meyers, C., & Jones, T. B. (1993). *Promoting Active Learning. Strategies for the College Classroom*: ERIC.

Miller-Young, J., & Yeo, M. (2015). Conceptualizing and communicating SoTL: A framework for the field. *Teaching and Learning Inquiry, 3*(2), 37-53.

Piaget, J. (2013). *The construction of reality in the child*: Routledge.

Ryan, T. P. (2013). *Sample size determination and power*: John Wiley & Sons.

Steuter, E., & Doyle, J. (2010). Re-Vitalizing the First Year Class through Student Engagement and Discovery Learning. *Journal of Effective Teaching, 10*(1), 66-78.

Tawfik, A. A., & Lilly, C. (2015). Using a flipped classroom approach to support problem-based learning. *Technology, Knowledge and Learning, 20*(3), 299-315.

Yazici, B., & Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation, 77*(2), 175-183.

#