Improving teacher training using empirical methods:
Evaluation of a teacher training seminar on heterogeneity in the classroom

Prof. Dr. Claas Wegner
Bielefeld University, Germany

Lars Wattenberg
Bielefeld University, Germany

Alexander Maar
Bielefeld University, Germany

Dominik Gardow
Bielefeld University, Germany

Abstract

This article investigates the participating students' evaluation of a teacher training seminar about heterogeneity in the classroom at Bielefeld University, Germany. The evaluation made use of a short questionnaire and an end-of-term test of the students' knowledge on the subject matter. The questionnaire is tested for objectivity, reliability, and validity and is shown to be a suitable tool. The test is shown to have certain weaknesses, but ways of improving it for future use are suggested. The method of evaluation used in this study can be adapted for other university courses. The importance of using empirical methods such as the one employed in this study for the improvement of teacher training is stressed. The seminar in question is shown to be largely successful and to contribute to the development of central diagnostic skills defined by the German conference of ministers of education.

Key words: teacher training, classroom heterogeneity

Please address all correspondence to: Alexander Maar, Bielefeld University, alexander.maar@uni-bielefeld.de

Introduction

The results of the first PISA-survey conducted by the OECD in 2000 and the subsequent surveys received great attention in Germany. The country's relatively weak results were shocking to many educational experts and the general public alike, and they cast severe doubt on the proper functioning of the educational system. Despite some criticism of the methodology of the PISA-surveys and their interpretation, the so-called 'PISA-shock' resulted in discussions on various levels of politics and academia, and entailed plans for minor as well as major reforms (Grek, 2009, pp. 29-30).

One element of the educational system that has since been deemed to be in need of improvement is teacher training. It is seen as crucial for successful school education, as is also suggested by the comparatively high PISA-scores of some countries like Finland that focus on good teachers and teacher training programmes (Dobbins & Martins, 2012, p. 34). As a consequence, teacher training in Germany is now in the process of being made more

comparable across the different federated states (*Länder*), and a greater focus is laid on empirical measures to ensure the quality of such programmes.

Comparability across the states and the adherence to certain nation-wide standards are difficult to achieve in Germany, since educational policy is not the responsibility of the federal government, but is governed by the sixteen states individually. This results in a vast number of different systems of school education and teacher training, which some commentators describe as a "patchwork" (Lange, 2007, p. 159)[1]. However, the *Kultusministerkonferenz* (conference of ministers of education, KMK for short) tries to achieve a degree of unity and comparability within the educational system(s). In 2004, the KMK defined a set of competences, or "standards", that students studying to become teachers have to acquire in the course of their teacher training programmes (KMK, 2004a). These competences are to be implemented by the state governments and the individual universities.

This study conducted at Bielefeld University investigates a teacher training seminar titled "Theory-seminar: Dealing with heterogeneity in teaching natural sciences". The course was offered in summer semester (SS) 2011, winter semester (WS) 2011/12, and in summer semester 2012. It is part of the biology education project *Kolumbus-Kids*, which aims at fostering the abilities of pupils interested and talented in natural sciences. At the same time, *Kolumbus-Kids* also serves as a teacher training programme. In accordance with the goal of creating empirically sound teacher training programmes, the effectiveness of the seminar in preparing future teachers to deal with heterogeneous groups of pupils[2] and to recognise certain types of pupils is measured. This corresponds with one of the competences defined by the KMK, the diagnosis-competence, especially competence number 7 in the set *evaluating*: "Teachers diagnose their pupils' requirements for learning and learning processes; they support pupils in a target oriented manner and advise pupils and their parents" (KMK, 2004a). Furthermore, this study assesses the quality of the tool that was used for evaluating the seminar, as it was specifically adapted to fit the course's conditions. Since empirical methods can only produce valuable results if the tools used are objective, reliable, and valid, ensuring these qualities is of central importance for improving teacher training.

Teacher training at Bielefeld University

Modern teacher training needs to meet university standards. The core of university education is the application of scientific and scholarly methods; it is, in other words, *research* (Terhart, 2007, p. 212). As Terhart (2007) points out, giving up this principle, the connection of teacher training programmes and current research in the future teachers' respective disciplines, would mean a return to a 19th century attitude, when, in Germany, teachers were not educated at universities, but in special "teachers' seminars" (p. 212).

In Bielefeld, research at a university level and the acquisition of the skills needed for teachers' future work in schools are connected by using explorative learning and through increased cooperation with the regional centres for teacher training, which are responsible for the second part of teacher training that takes place after university. The cooperation is realised in the "practical semester": during their master's course at university, students spend one semester working at a school and reflecting on this practical experience. The critical reflection of their role and their actions is one of the main goals of explorative learning and involves

[1] All quotations from German sources are translations by the authors of this article.
[2] To avoid confusion, the term "students" refers to university students *only*, while secondary school students are referred to as "pupils".

connecting theoretical knowledge and scientific methods acquired at university with practical knowledge and experience acquired in- and outside university.

Explorative learning means that students acquire theoretical knowledge and insights into scientific methods by (semi-)independently carrying out projects, which includes developing research questions and hypotheses, collecting and presenting evidence, drawing conclusions, etc. (Huber, 2009, p. 11). According to Fichten (2012), this type of learning, albeit not a new concept, has experienced a "renaissance" and has become increasingly common in teacher training programmes (para. 1.1). At Bielefeld University, *Kolumbus-Kids* is one of the projects that tries to put the principles of explorative learning into practice. It is a project that aims at fostering gifted secondary school pupils' abilities in the area of natural sciences, and biology in particular. Courses for those pupils are organised and taught by university students studying to become teachers. The theoretical seminar that is evaluated in this paper accompanies this practical work done by the trainee teachers. The emphasis of *Kolumbus-Kids* is on connecting theory and practice. In the "theory-seminar", this is achieved by simulating and reflecting typical situations that can occur in school.

The tool used for the evaluation

The tool used for evaluating the seminar's success consisted of two parts: a questionnaire filled out by the participating students and a test used to verify the students' self-assessment in the category "learning outcome".

The questionnaire has four dimensions: "form and structure", "instructor's traits", "quantity and relevance", and "learning outcome". It uses a 5-point Likert scale on which students have to express their level of agreement with certain statements about the course. The questionnaire also contains an open-question-section at the end (items 21-23). In total, it consists of 24 items making it rather short. While short questionnaires are usually less reliable (Zumbach et al., 2007, p. 317), they are also more likely to be filled out completely than longer ones and are easier to work with. The questionnaire is based on the tool developed by Zumbach et al. (2007), but it has been adapted to the *Kolumbus-Kids* theory-seminar by adding additional, course-specific items. The construct validity was re-verified after these changes.

The test was used to check whether the students' self-assessed "learning outcome" correlated with a factual increase of certain competences (items 19 and 20). Only by verifying the students' self-assessment with another, independent measure it can be assured that the seminar's aims were actually reached, since students could erroneously assume to have acquired a certain skill without this being the case. Also, their answers in the questionnaire could be influenced by their personal feelings towards the course or the instructor. Therefore, the test is a valuable addition to the questionnaire.

Questions and hypotheses

As mentioned before, the aims of this study were to measure the seminar's success and to assess the quality of the tools used for the measurement. Both aims are important for improving teacher training. Consequently, the following working hypotheses were developed and tested:

- WH_1: The *Kolumbus-Kids* theory-seminar enables students to distinguish various types of pupils. The seminar increases or develops the students' diagnosis-competence.
- WH_2: The students' self-assessment in the area "learning outcome" – items 19 and 20 in particular – is supported by the results of the test.
- WH_3: The questionnaire is sufficiently reliable.

- WH_4: The questionnaire has construct validity i.e. it actually measures the quality of the seminar.

The following two sub-hypotheses derive from WH_4:

  - $WH_{4.1}$: The items specified for a certain dimension exhibit a high correlation only with their respective dimensions.
  - $WH_{4.2}$: The items of one dimension exhibit a high correlation with one another.

Participants

Before the results of the questionnaire are presented, a brief overview of the participants is given (items 1-4): In total, 61 questionnaires were filled out. Out of these, 19 were part of a retest in summer semester 2011, which might lead to certain distortions in the statistics presented below.

The majority of participants (82 %) were female (item 1). This uneven distribution of genders persisted, with a degree of fluctuation, over the course of the three semesters that are taken into account here. This observation can be partly explained by the fact that the majority of teacher trainees are female.

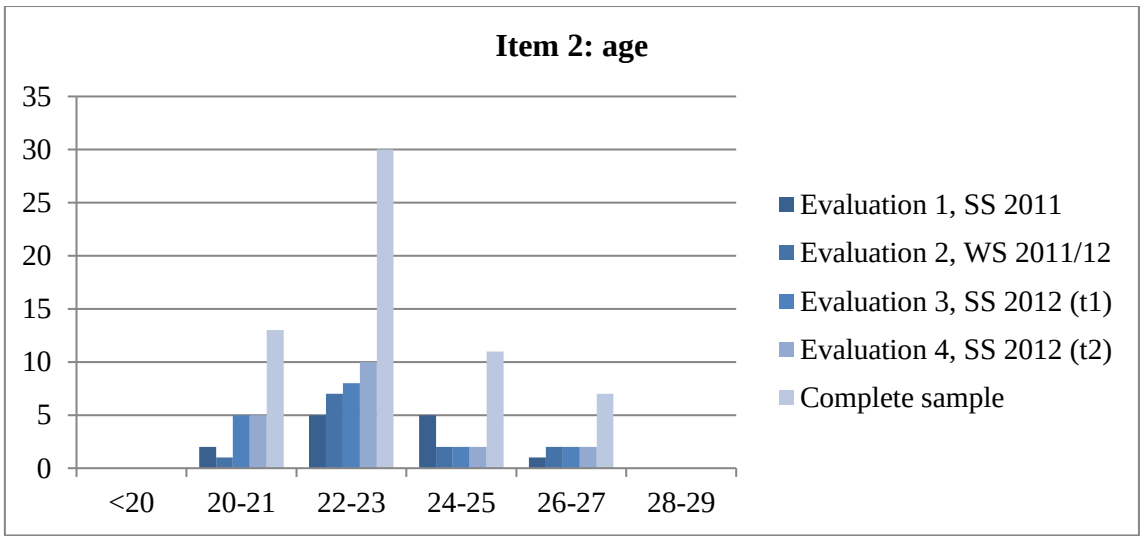

Fig. 1. Item 2: Age groups of the participants. Evaluations 1-4.

The arithmetic mean of the participants' age group is 3.2, which means 22-25 years of age. None of the participants were younger than 20 or older than 27 years.

None of the participants were in their first semester, but apart from that, the distribution was quite broad (item 3). Students in their fourth (26) or sixth (15) semester made up the majority of participants (67.2 %), while students of other semesters up to the tenth were present in much smaller numbers. The fact that most students were at least in their third semester might have influenced the results of the questionnaire. One could assume more experienced students tend to be better at evaluating seminars, as they have come in contact with a number of different ways of teaching. However, the strength of this effect could not be measured in this study.

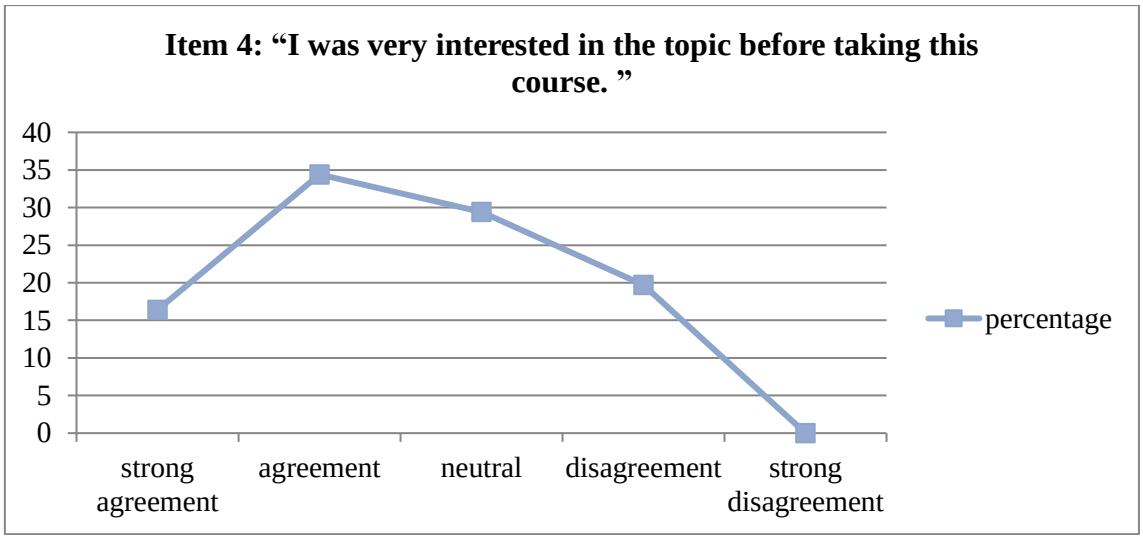

Fig. 2. Item 4: "I was interested in the topic before taking this course." Absolute numbers and percentages. Evaluations 1-4 combined.

Item 4 shows relatively high levels of interest in the topic. Almost all participants were interested in the topic of the seminar to some extent, while 50.8 % expressed either strong agreement or agreement with the aforementioned statement. This has to be kept in mind when interpreting the results, because thematic interest can act as a confounding variable and influence the students' perceptions of the course and the instructor (Rindermann, 2003, p. 239).

Results of the evaluation

After having introduced some basic information on the composition of the sample, the results of the evaluation (items 5-20 and 24) are presented in this section.

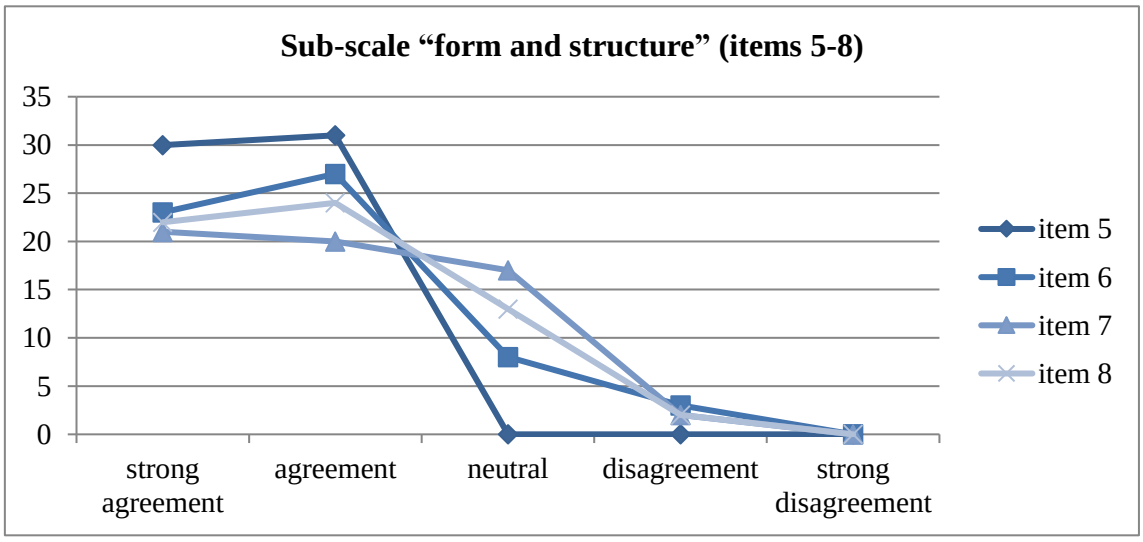

Fig. 3. Sub-scale "form and structure" (items 5-8). Absolute numbers. Evaluations 1-4 combined.

For the sub-scale "form and structure", the students' feedback was mostly positive or very positive. When coding strong agreement as 1.0, agreement as 2.0, and so on, arithmetic means range from 1.51 to 2.0. Item 5, "Content was adequately made comprehensible through

the use of examples, visualisation, etc.", exhibits the lowest standard deviation (.504), with all students expressing either strong agreement (31) or agreement (30). The most negative feedback in this sub-set of items was given for item 7, "The aims of the seminar were clearly defined". The answers for this item exhibit an arithmetic mean of 2.0 and a standard deviation of 0.883. The results fall in line with answers to open questions included in the study (items 21-23), in which some students expressed that the aims should have been defined more clearly. One student, for example, stated that "[m]ore precisely defined aims would have been interesting". However, most students (41) still strongly agreed or agreed with the statement, while 17 were neutral and only 2 disagreed.

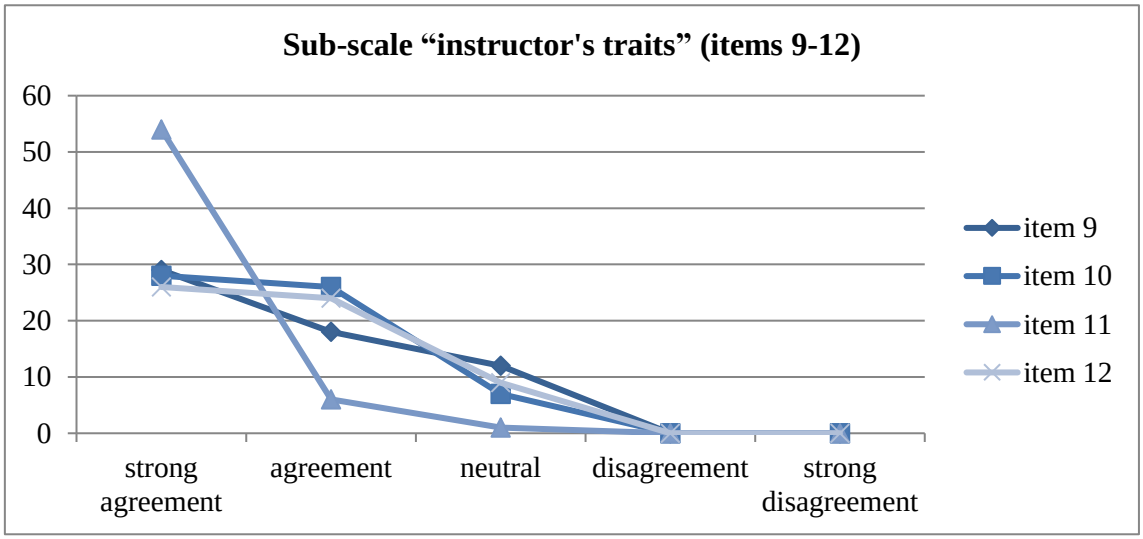

Fig. 4. Sub-scale "instructor's traits" (items 9-12). Absolute numbers. Evaluations 1-4 combined.

No student expressed disagreement with any of the items relating to the instructor's character traits and behaviour. The item that received the most positive response was item 11, "In dealing with students, the instructor was friendly and approachable". For this item, the arithmetic mean was 1.13, and the standard deviation was 0.386. The results for items 9, 10, and 12 were similar to one another, with means ranging from 1.66 (item 10) to 1.71 (items 9 and 12).

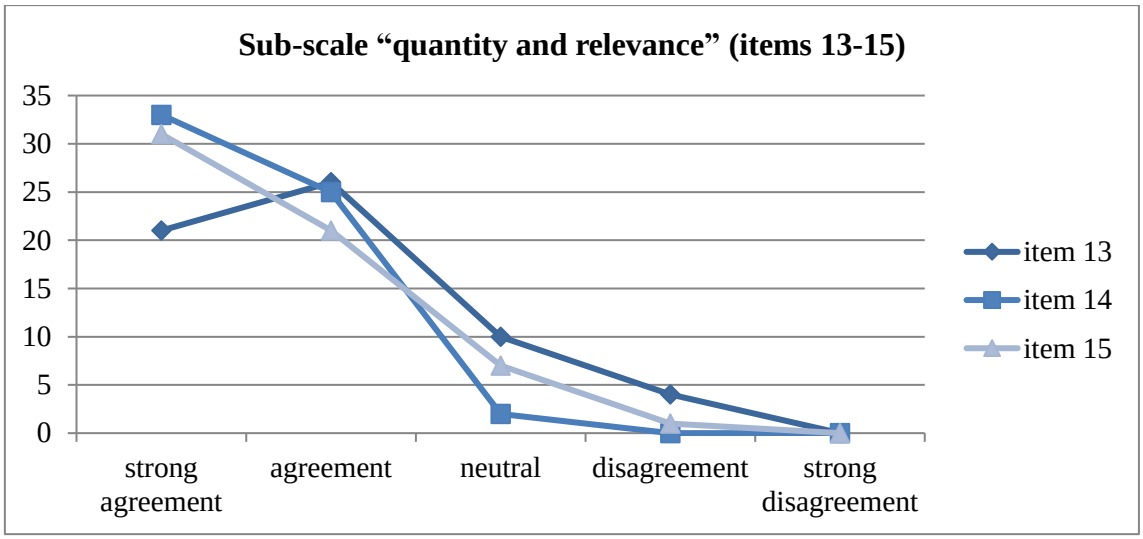

Fig 5. Sub-scale “quantity and relevance” (items 13-15). Absolute numbers. Evaluations 1-4 combined.

For this sub-scale, again, the results were mostly very positive or positive. However, a significant number of students (14) were either neutral towards or disagreed with the course’s content being relevant (item 13), resulting in a mean of 1.95 and a median of 2.0 as opposed to 1.0 for items 14 (pace) and 15 (quantity).

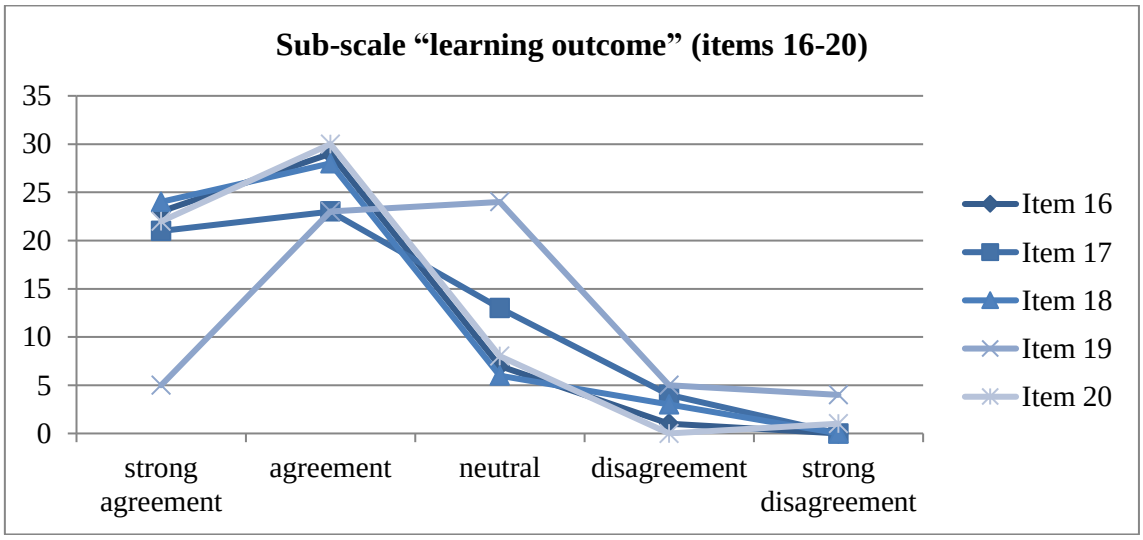

Fig. 6. Sub-scale “learning outcome” (items 16-20). Absolute numbers. Evaluations 1-4 combined.

The sub-scale “learning outcome” yielded mixed results. It is the only sub-scale in which some students expressed strong disagreement, with items 19 and 20. Item 19, “Do you think that what you learned in this seminar enables you to recognise and, if necessary, solve problems that can occur in heterogeneous groups of learners?”, received the least positive feedback of all items in the questionnaire, with four participants expressing strong disagreement, and another five students disagreeing. This results in a mean of 2.67 and a median of 3.0. The other items in the sub-scale received rather positive feedback.

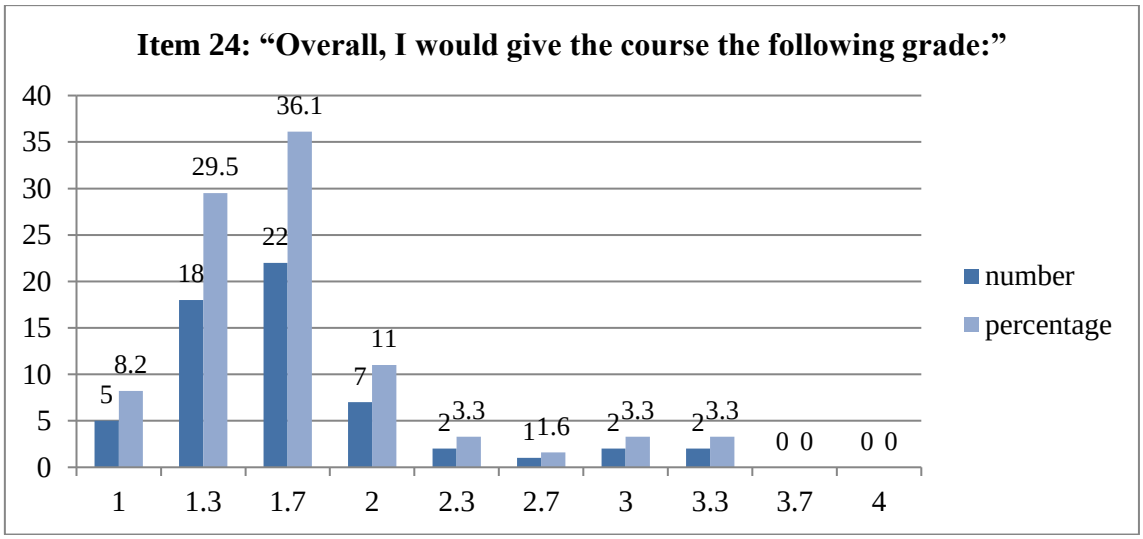

Fig. 7. Item 24: "Overall, I would give the course the following grade:" Absolute numbers and percentages. Evaluations 1-4 combined.

For item 24, students were asked to decide on an overall grade for the course using the grading system that is used at German universities: 1.0 is the best grade, 4.0 is a passing grade, and anything higher than that is a failing grade. A clear majority of the participants graded the course 1.0, 1.3, or 1.7, which can be considered very good grades. The best grade (1.0) was given by five students (8.2 %). The most common grade was 1.7 with 22 students (36.1 %). The worst grade was 3.3, which was given by only two students. This is still a passing grade, albeit not a good one.

Discussion of the questionnaire results

As could be seen, most feedback for the course was rather positive, or at least neutral. Especially the instructor was regarded as positive. The overall grading of the course (item 24) was overwhelmingly positive as well. On the whole, the students' impressions seem to have been a good one. However, item 19, which measured the students' self-reported ability to apply what they had learned, or were supposed to learn, in the course, yielded rather mixed results that might cast some doubt on the course's success. While 28 students thought that participating in the course had enabled them to solve problems that result from heterogeneity in the classroom, a slim majority of 29 thought this to be only partly true, and four students did not feel ready to deal with such problems at all.

While it is one of many, item 19 is one of the most important items in the questionnaire, because it represents one of the seminar's main aims, namely to detect and solve potential problems within heterogeneous learner groups. If this aim is not attained by the majority of students, according to their own judgement, then the positive feedback on many other aspects of the course, for example the aforementioned positive evaluation of the instructor's personality, is not worth much. After all, factors like the students' appreciation of the course atmosphere are only valuable inasmuch as they support learning outcome and are not primary goals in themselves. Therefore, the student's responses to item 19 must lead to a critical view of the course and are in need of explanation. To find such an explanation, one must turn to the open questions (items 21-23) and to the results of the test that was supposed to verify the reliability of the students' self-assessment for items 19 and 20. As shown below, a number of students criticised in the open question section that the course was one-sided and focused almost

exclusively on gifted pupils, while other groups of learners, such as children with disabilities or learning difficulties, were left out, resulting in an incomplete picture of heterogeneity in the classroom.

Open questions

In the open questions section (items 21-23), students could utter praise, criticism and suggestions not covered by the standardised closed questions. The questions were "What did you particularly like about the seminar?" (item 21), "What did you not like?" (item 22), and "Which suggestions for improvement and which further comments do you have?" (item 23). Listing all answers would require too much space and would not be of great utility. Therefore, the presentation is limited to answers that occur frequently, relate to the study's working hypotheses or help explain some of the more surprising results in the closed question section. As the answers are based on the opinions of individual participants, their significance for evaluating the seminar must not be overemphasised. However, they are a valuable addition to the picture of the seminar's success. Since the seminar was conducted in a similar fashion each semester, the answers are not discussed separately for every evaluation.

One element that was mentioned as positive by multiple participants was the filming of small simulated teaching sessions carried out by the students. One student declared that this allowed him or her to judge his or her "outward appearance" while teaching. Another student liked the "videotaping with subsequent discussion". A student in summer semester 2012 expressed that she found it helpful to "have the possibility of watching [her] way of speaking, facial expressions, and gestures". The use of the "SMART Board" was also positively mentioned by some students. These examples might be one factor in explaining the positive responses to item 5, which, among other things, dealt with visualisation.

A point of criticism that occurred repeatedly in the open question section was that the course was seen as one-sided and as not covering the full range of heterogeneity in the classroom. One student suggested including ADHD, as the he or she was of the opinion that behaviour of children with ADHD can be similar to that of *underachievers*. Another participant criticised "the focus on heterogeneity of achievement, and on the upper end of the scale; [furthermore] the focus on the *Gymnasium*"[3]. The "focus on gifted education" was also seen as negative by yet another student, while, as one participant writes, "average or 'weak' pupils were not really dealt with". One suggestion was to include "factors like gender and migration" and to discuss the conditions at comprehensive schools as well. Furthermore, "children with learning difficulties, let alone children with disabilities" were not discussed, according to one student. These are just some examples, but more could be named. The seminar's relatively narrow focus was, therefore, perceived as problematic not only by individual students, but at least by a certain number of the participants. The course's focus on gifted pupils and the various 'types' of these that can occur in the classroom can be explained by that fact that the course is a part of the *Kolumbus-Kids* project, which has the explicit goal of fostering the abilities of gifted learners and is not a project for the complete range of pupils. However, the name of the seminar ("Dealing with heterogeneity in teaching natural sciences") might suggest a fuller picture of this phenomenon than is actually delivered, causing a discrepancy between the students' expectations and the reality of the seminar. One might argue that a more fitting name and greater transparency regarding the seminar's focus would have been able to avoid some of the criticism mentioned above.

[3] The *Gymnasium* is the upper track of the German school system, which traditionally focuses on higher cognitive and academic skills.

Results of the test

The test was administered three times, with a total of 39 participants. To test whether the students' self-assessment for items 19 and 20 corresponded with an actual development of their competences, the participants were presented with two cases: *Benjamin* and *Lisa*.

Task 1 was to name and briefly characterise the types of pupils that were described in each of the cases. For *Benjamin*, the correct answer was to characterise him as a typical *underachiever*. An academic *underachiever* is a child or adolescent who exhibits "a severe discrepancy between [his or her] expected achievement and his or her actual achievement which is not attributable to any diagnosed learning disabilities" (Figg et al., 2012, p. 54).

The test in evaluation 4 yielded the best results, with all students correctly characterising Benjamin as an underachiever, but in the first and second evaluation, the results were mixed, with only 58 % and 50 % correct answers, respectively. The rest of the students categorised Benjamin as either a gifted or highly gifted pupil. The vastly varied test results for case *Benjamin* in different semesters suggest that *underachievers* were discussed in more detail in summer semester 2012 (evaluation 4), enabling students to recognise them more reliably.

For case *Lisa*, task 1 was similar, with the difference that three pupils were to be classified: Lisa was supposed to be characterised as a typical *quiet pupil*, Markus as a *class-jester* and Rocco as the *class's favourite*.

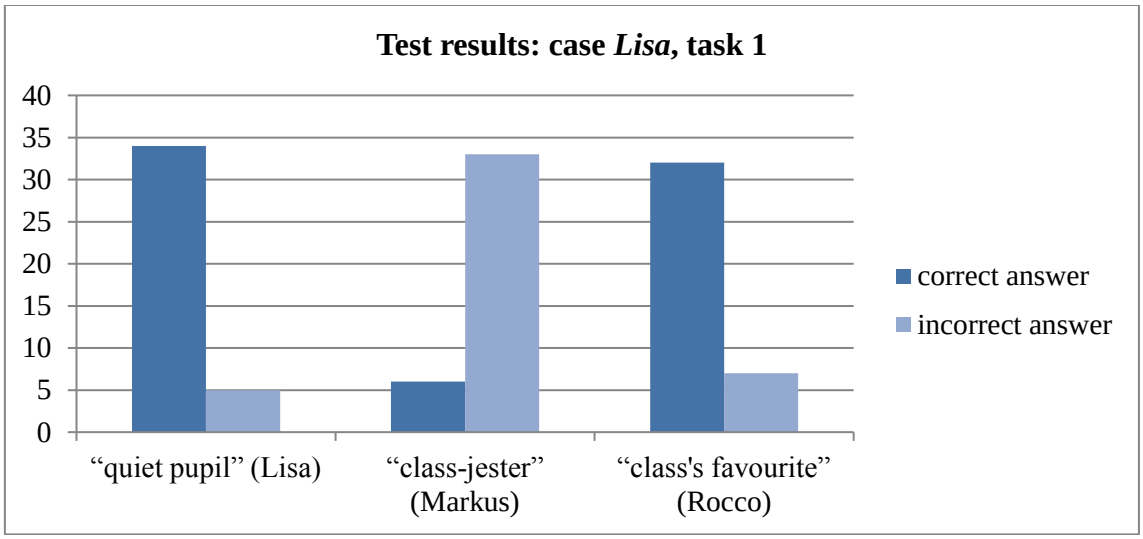

Fig. 9. Test results: case *Lisa*, task 1. Absolute number of correct and incorrect answers. Evaluations 1-4 combined.

Most participants were able to correctly identify Lisa's and Rocco's types, although they did not always use the correct terminology. Many students did not call Lisa a quiet pupil, but used words like "shy" or "anxious" instead, while Rocco was described as the class's "star" or even "king". If the description of the typical behaviour was fitting, answers of this kind were counted as correct. For Markus, the results were completely different: Only 6 out of 39 students were able to correctly classify him as a *class-jester*, while the rest wrongly characterised him as an *underachiever* or *troublemaker*. The differences between these concepts were apparently not obvious to most participants.

For item 20 in the questionnaire, a majority of participants expressed either strong agreement (36.07 %) or agreement (49.18 %), while the responses of eight students were neutral, and one expressed strong disagreement. So, most students thought the course had led them to develop a good or very good ability to recognise certain types of pupils in a heterogeneous group. The test paints a more complicated picture than the questionnaire alone: While most students seem to have developed the abilities to recognise and describe some types of pupils, they had difficulties with certain types and mistook them for others. One might argue that the participants of the seminar did indeed acquire a set of skills that helps them to classify different types of typical pupils, but some of them seem to have overestimated their new-won competence in the self-assessment. In general, however, the test results were mostly satisfying.

The answers to the remaining questions, questions 2-4 in case *Benjamin* and 2.1-2.4 in case *Lisa*, are more difficult to evaluate, as they were more open. Students' answers diverged quite clearly from one another and did not always conform with the pre-formulated expected answers, but without necessarily being incorrect. Question 3 (case *Benjamin*), "What exactly is the risk associated with the continuing social isolation of a pupil?", for example, was not answered in the expected way in most cases. Rather, many of the students offered descriptions that fit the question without really naming the particular danger that the question was aiming at, that is the risk of damaging the child's interpretation of his or her role and the difficulties this causes for the process of socialisation. For question 2.2 (case *Lisa*), "What could be causes for Markus' jester-like behaviour? Which of these causes seems most probable to you in this case?", the majority of participants thought that Markus' behaviour was caused by a lack of challenge, resulting in boredom, a lack of interest, and a craving for attention. None of the students, not even those who had correctly identified Markus as a *class-jester*, wrote that his behaviour could just as well result from psychological issues as from his academic performance.

Question 4 (case *Benjamin*), "How would you proceed to improve Benjamin's situation? Outline your plan to achieve your goals!", received many unexpected answers as well, but a lot of these were plausible nonetheless. Many creative and suitable ideas convey the impression that the students were indeed able, at least in theory, to deal with such a situation. But there were also unfitting or plainly wrong answers, casting some doubt on this assumption.

In total, the students' answers in the more open section of the test leave a mixed impression of their ability to deal with heterogeneous groups of learners and to solve problems that can occur in those groups. This is reflected in the participants' meagre self-assessment of their abilities in this area (item 19). In this case, the results of the test seem to verify the results of the questionnaire.

Assessing the quality of the questionnaire

This study also aims at assessing the quality of the questionnaire used for the evaluation, which is important for two main reasons: First of all, the evaluation of the seminar can only be used to improve future teaching if the method used is actually valid, reliable, and objective, as only working with a tool that fulfils these criteria can produce meaningful results. Secondly, developing good evaluation tools is absolutely necessary for teacher training that is based on empirical foundations, which is the kind of teacher training that is required for the modern educational system. While Zumbach et al. (2007) already tested the quality of the questionnaire, arriving at medium reliability and construct validity (p. 321, 323-24), the tool was used in an adapted form in this study, making it necessary to repeat the assessment of its quality. This is supposed to ensure that the adaptation to the seminar's specific requirements was not detrimental to the questionnaire's quality. The criteria for its qualities are as follows.

Objectivity means that multiple independent testers arrive at the same results when they use the testing tool (Bortz & Döring, 2006, p. 195; Moosbrugger & Kelava, 2008, p. 8). As the questionnaire is delivered in a paper-and-pencil format and the researcher is not in any way involved in the testing procedure itself, there is no reason to assume that the evaluation is not objective. Therefore, objectivity is not to be discussed any further.

Reliability describes the precision of the measurement. A higher precision means a smaller influence of random measurement errors (Moosbrugger & Kelava, 2008, p. 8). In this study, reliability is determined by calculating internal consistency and test-retest-reliability. Reliability corresponds to WH_3.

Validity means that the items are logically connected with the phenomena (factors) that are supposed to be assessed (Bortz & Döring, 2006, p. 200). In other words: The questionnaire is valid only if it really measures what it is supposed to measure. Here, the validity is analysed with a focus on the sub-scale "learning outcome" (items 16-20) by using factor analysis. Validity corresponds to WH_4, $WH_{4.1}$, and $WH_{4.2}$.

Reliability of the evaluation

Reliability can be expressed as $Rel = \frac{s_T^2}{s_X^2}$, with s_T^2 being the empirically measurable variance and s_T^2 being the actual variance that excludes all measurement errors. As reliability is measured variance divided by actual variance, fewer measurement errors mean a greater reliability (Bortz & Döring, 2006, p. 196; Moosbrugger & Kelava, 2008, pp. 115-116). A greater reliability therefore means a greater independence of the test from external and random factors. Since the degree of measurement errors cannot be known, the actual variance is an unknown variable and has to be estimated. There are a number of methods to estimate the true variance and thereby the reliability. In this study, internal consistency analysis and test-retest reliability are used. The reliability of the overarching construct and the reliability of the individual sub-scales are determined using internal consistency analysis, on the basis of a sample of n=61. Test-retest-reliability is employed to determine the reliability of the evaluations 3 (t1) and 4 (t2) in summer semester 2012. A correlative comparison of the two evaluations is performed.

The most common method of internal consistency analysis is Cronbach's alpha (Mutz & Daniel, 2008, p. 36). Cronbach's alpha is described by the formula $\alpha = \frac{n\bar{r}}{1+\bar{r}\,(n-1)}$, with n being the total number of items and $\bar{r}$ being the average intercorrelation – the mean of all possible split-half coefficients (Rindermann, 1998, p. 79) – of the items. When applying this method, all items are treated as individual tests and are correlated with one another. The mean of all correlations is the average reliability of the complete scale.

Internal consistency of the overarching construct

Items 5-20 are taken into account, because these measure the teaching quality of the seminar. Before the overall internal consistency can be calculated, the sub-scales "form and structure", "instructor's traits", "quantity and relevance", and "learning outcome" are analysed separately. The advantage of using a multi-dimensional model like this is that it allows one to analyse individual dimensions of teaching instead of only regarding the seminar's quality as a whole.

The internal consistency analysis using Cronbach's alpha yields the following results (n=61, evaluations 1-4):[4]

- For the sub-scale "form and structure", the mean intercorrelation of the 4 items is $\bar{r}$=0.201. Therefore, α=0.501.
- For the sub-scale "instructor's traits", the mean intercorrelation of the 4 items is $\bar{r}$= 0.264. Therefore, α=0.581.
- For the sub-scale "quantity and relevance", the mean intercorrelation of the 3 items is $\bar{r}$=0.137. Therefore, α=0.322.
- For the sub-scale "learning outcome", the mean intercorrelation of the 5 items is $\bar{r}$=0.4097. Therefore, α=0.775.

According to Bortz and Döring (2006), reliabilities between 0.8 and 0.9 are medium, while a reliability coefficient over 0.9 is considered high (p. 199). A good measuring tool should have a reliability coefficient of at least 0.7 (Moosbrugger & Kelava, 2008, p. 11). Judging by these criteria, the sub-scales "form and structure", "instructor's traits", and "quantity and relevance" do not exhibit sufficient reliability. This can in part be explained by the relatively small sample. As has been mentioned, Zumbach et al. (2007) found a much higher reliability while using a sample of n=648 (p. 323).

The sub-scale "learning outcome", in contrast, had a slightly better reliability than Zumbach et al. found in their study: α=0.775 as opposed to α=0.73. The reason for this slightly improved reliability coefficient might be the addition of two seminar-specific items (items 19 and 20). The increased reliability of this sub-scale is advantageous, as the scale serves as the backdrop against which the results of the test are discussed. Ensuring the reliability of the data is a necessary prerequisite for making meaningful comparisons.

Reliability of the overarching construct "quality of the seminar"

Case processing summary			
		N	%
Cases	Included	57	93.4
	Excluded	4	6.6
	Total	61	100.0

Table 1. SPSS-output: Overview of cases included and excluded in the calculation of the internal consistency of the overarching construct "quality of the seminar" (items 5-20).

Reliability statistics		
Cronbach's alpha	Cronbach's alpha based on standardized items	N of items
.817	.818	16

Table 2. SPSS-output: Cronbach's alpha of the overarching construct "quality of the seminar" (items 5-20).

As was expected because of the greater number of items (16), the internal consistency of the overarching scale was significantly higher than that of the individual sub-scales. According to the classification by Bortz and Döring (2006), a Cronbach's alpha value of 0.817

[4] The internal consistency was calculated using SPSS and additional calculations performed by the authors of this study.

means that the questionnaire measures the quality of the seminar with an adequate reliability. If the relatively small sample of n=61 and the low number of items are taken into account, this is a more than acceptable result.

Test-retest reliability

In summer semester 2012, evaluation 3 (t1) and evaluation 4 (t2) were conducted five weeks apart, with 17 and 19 participants, respectively. The outcomes can be compared in the diagrams below. Items 1-3 and 21-23 are not taken into consideration, as they are not relevant for the calculation of the test-retest reliability. Item 24 is depicted, but only for illustration, as it, does not count towards the degree of test-retest reliability.

Test-retest reliability is another measure for the reliability of a measuring tool. It is defined as follows:

$$rel_{test-retest\ method} = \frac{S_T^2}{S_x^2} = \frac{cov(t_1, t_2)}{S_{t1} \times S_{t2}} = r_{t1t2}$$

Test-retest reliability is determined by using the same testing tool a second time after a certain time span. The reliability is high when there is a high correlation between both measurements (Moosbrugger & Kelava, 2008, p. 117). It is based on the assumption that the actual values of the participants should not change between the first and the second measurement and that the influence of confounding variables should be constant if the same measurement procedure is used. If this is the case, then the correlation of the values measured in the different measurements is equal to the unknown value of true variance divided by measured variance (Moosbrugger & Kelava, 2008, p. 117).

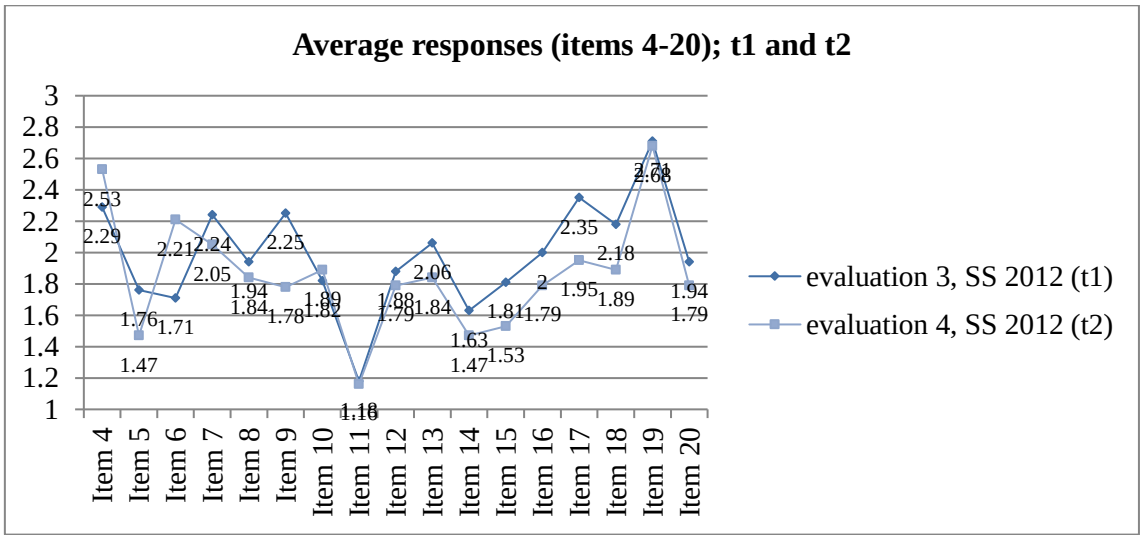

Fig. 10. Average responses (items 4-20). Agreement coded as 1-5. Comparison of evaluations 3 (t1) and 4 (t2).

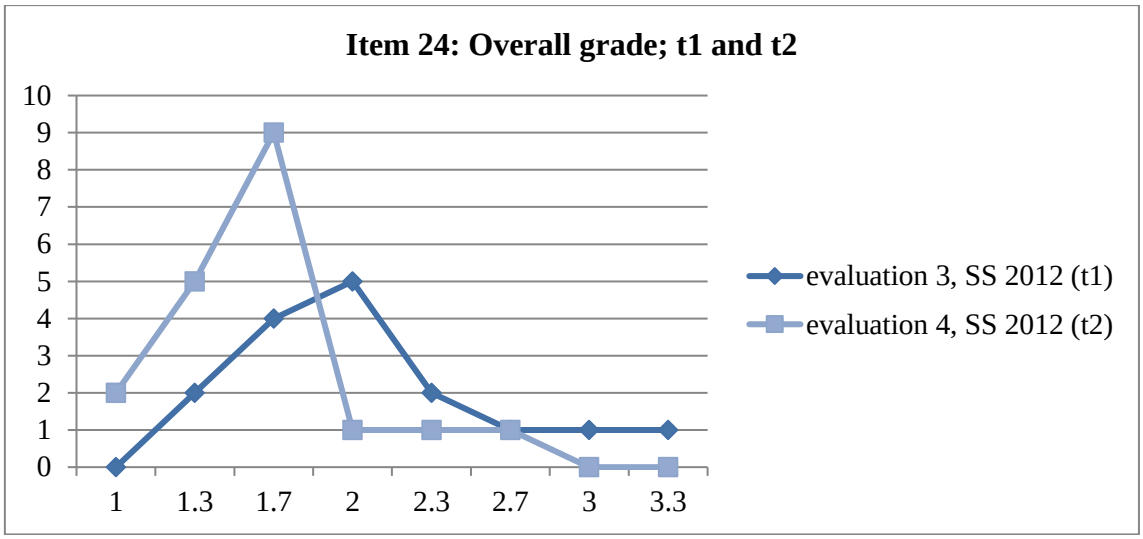

Fig 11. Item 24: Overall grade. Comparison of evaluations 3 (t1) and 4 (t2).

At first glance, the charts seem to indicate a correlation between evaluation 3 (t1) and evaluation 4 (t2), showing that the participants' judgement remained mostly unchanged over the time between the two evaluations. It seems that the true variance and the degree of measurement errors were relatively constant, which would mean that the evaluation questionnaire is based on scales that test stable attributes, for example personal values and attitudes, but this impression is misleading, since the test-retest reliability calculated for the overarching scale "quality of the seminar" for the repeated evaluation in summer semester 2012 is merely 0.429.

Correlations			
		Quality of the seminar; evaluation 3, SS 2012 (t1)	Quality of the seminar; evaluation 4, SS 2012 (t2)
Quality of the seminar; evaluation 3, SS 2012 (t1)	Pearson correlation	1	.429
	Significance (2-tailed)		.098
	N	16	16
Quality of the seminar; evaluation 4, SS 2012 (t2)	Pearson product-moment correlation coefficient	.429	1
	Significance (2-tailed)	.098	
	N	16	18

Table 3. SPSS-output: Test-retest-reliability (Pearson product-moment correlation coefficient) of the overarching scale "quality of the seminar". Evaluations 3 and 4.

This result is both unexpected and unsatisfactory. The questionnaire does not seem to be based on constant values and error variance. The true values and the error variance rather seem to be subjected to unsystematic changes, causing insufficient test-retest reliability. "When

the true values change unsystematically (different for different people) between the moments of measurement, then the correlation of the test values between the moments of measurements 1 and 2 is decreased" (Moosbrugger & Kelava, 2008, p. 116-117).

The low test-retest reliability indicates that the items of the questionnaire are not stable over time, which means test-retest reliability is not a suitable measure for reliability of the questionnaire. This becomes even more apparent when one compares internal consistency and test-retest reliability. For t_1, Cronbach's alpha is 0.734, while it is 0.804 for t_2. Both, evaluation 3 (t_1) and evaluation 4 (t_2), have medium reliability, based on their internal consistency, which is a significantly better result than the meagre test-retest reliability of 0.429 would have one expect. According to Rindermann (1997), test-retest reliability is lower than the consistency coefficient in most cases (p. 22). The results of the internal consistency analysis are probably closer to the actual reliability. The visible contrast of the results obtained through the use of each method casts some doubt on whether test-retest reliability is a suitable measure for estimating the actual reliability of the tool used in this study.

Validity of the evaluation

The validity analysis of the evaluation consists of two parts: determining the construct validity of the evaluation questionnaire and analysing the potential influence of confounding variables and biases on the students' judgement. Validity is arguably an even more important measure of quality, for a highly reliable measuring tool can be useless when it measures something other than what it is supposed to measure (Bortz & Döring, 2006 p. 200).

Construct validity of the questionnaire

The questionnaire is construct valid if the dimensions of the questionnaire – the sub-scales – are used by the participants as the basis of their judgement of the overarching construct "quality of the seminar" (Pohlenz, Grindel & Köpke, 2006 p. 238). In other words, construct validity describes an agreement of the dimensions of quality defined in the measuring tool and the students' judgement of quality. If these agree, one can conclude that the items adequately represent the sub-scales and therefore the overarching construct "quality of the seminar".

Factor analysis is employed to test whether there is a sufficient correspondence between the empirical data and the overarching construct "quality of the seminar". This method is based on the principle that items with a high degree of inter-correlation form a factor, or sub-scale. The factors, on the other hand, should only have weak correlations with one another. Since Zumbach et al. (2007) have already determined the four dimensional structure of the questionnaire through factor analysis (pp. 321, 323), only a confirmatory factor analysis of the added items 19 and 20 and the sub-scale "learning outcome" is performed. Ideally, items 19 and 20 should correlate with the dimension as a whole as well as with the other items in the sub-scale (items 16-18).

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin measure of sampling adequacy.		.649
Bartlett's test of sphericity	Approximate chi-square	93.739
	Df	10
	Significance	.000

Table 4. SPSS-output: Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity. Performed for the sub-scale "learning outcome" (items 16-20).

Before a factor analysis of items 19 and 20 can be performed, the data is checked for substantial correlations. By calculating the Kaiser-Meyer-Olkin (KMO) measure and the sphericity according to Bartlett, the suitability of the data for factor analytical testing is determined. The KMO-coefficient of 0.649 measured here indicates an acceptable suitability of the data for factor analysis (Bühl, 2014, p. 628). The significance of 0.000 means that the null hypothesis – all correlations between the 5 items of the sub-scale equal 0 – can be refuted (Bühl, 2014, p. 628). Therefore, it is certain that there are correlations between the five items, which justifies factor analytical testing of the sub-scale "learning outcome".

Factor analytical testing using SPSS reveals that two factors can explain 72.779% of the variability within the data for the sub-scale "learning outcome", which is why a two-factor solution is proposed. As can be seen in the table below, factor one accounts for a greater degree of variability than factor two for all items in the sub-scale. These results were to be expected for items 16-18, because the questionnaire had been subjected to factor analytical testing by Zumbach et al. (2007). However, the new items 19 and 20 strongly correlate with factor one as well, with coefficients of 0.709 and 0.741, respectively. While some of the items show medium correlations with factor two as well, which means it would be possible to split up the sub-scale "learning outcome" into two separate sub-scales, it seems reasonable, due to the consistently higher correlations with factor one, to keep the five items as one sub-scale. The confirmatory factor analysis thus shows that the added items do not negatively affect the construct validity of the sub-scale learning outcome, which means the quality of the questionnaire is not reduced by the addition of the two items.

Component matrix		
	Component	
	1	2
Item 16: "If possible, the content of the seminar was up-to-date."	.671	.466
Item 17: "I think the learning effect of the seminar was great."	.772	-.469
Item 18: "The seminar increased my interest in the topic."	.733	-.525
Item 19: "The seminar enabled me to recognise and, if necessary, solve problems that can occur in heterogeneous groups of learners."	.709	.051
Item 20: "I think the seminar increased my awareness of characteristic types of students."	.741	.537

Table 5. SPSS-output: Component matrix depicting the correlations between items 16-20 and the two components resulting from factor analytical testing.

Confounding variables

As has been mentioned, thematic interest (item 4) can act as a confounding variable. This fact has to be taken into account if one aims at measuring the quality of a seminar rather than the students' motivation to attend it, which is often caused by an interest in the topic, according to Spiel and Gössler (2000). Other confounding variables are thought to have only small systematic influences on the results of evaluations.

In order to ensure that the students' thematic interest did not confound the results, the Pearson product-moment correlation coefficient is calculated, testing for correlations between item 4 and items 5-20. The calculations reveal correlation coefficients ranging from -.223 to 0.225. As all correlations are weak, it can be concluded that the students' interest in the topic

had little or no influence on their judgment of the seminar's quality, which means the results of the evaluation are not invalidated by this potential confounding variable.

Summary of the results

The data gathered using the questionnaires shows that the participants regarded the *Kolumbus-Kids* theory-seminar mainly as good or even very good. This includes items that dealt with the course and the instructor in the sub-scales "form and structure", "instructor's traits", and "quantity and relevance", as well as the self-assessment sub-scale "learning outcome". The overarching construct "quality of the seminar", and especially the sub-scale "learning outcome" have been shown to exhibit medium internal consistency, meaning they are sufficiently reliable. Test-retest reliability, however, does not appear to be an appropriate measure for judging the tool's reliability, which means that multiple measurements in the same semester are not necessary. Furthermore, the questionnaire is construct valid and the potential confounding variable "thematic interest" does not influence the measurement. The results of the evaluation tests show that students who participate in the course acquire diagnosis-competences. The tasks relating to item 20 were solved in a satisfactory manner, which means the largely positive self-assessment of the students is mostly supported by the test results. The empirical data supports the claim that the course contributes to the development of the diagnosis-competences of the participants, in line with the KMK's demands (KMK, 2004a/2008).

Students were also supposed to learn how to recognise and solve problems in heterogeneous groups of learners (item 19). In this area, the students' self-assessment was much more careful, which reflects the test results. These were decisively weaker for the tasks relating to item 19.

Due to the design of the measuring tool, it is not possible to distinguish pre-existing knowledge from knowledge acquired in the seminar. Therefore, one cannot say whether the results of the test, especially of the questions relating to item 19, are to be attributed solely to the seminar's quality or whether they were influenced by the students' background knowledge. This must be considered a drawback of the measuring tool used in this study.

Discussion and conclusion

In retrospect, the test design appears partly flawed. On the one hand, the open questions[5] led to widely divergent answers that made the evaluation and the gathering of precise data quite difficult. This runs contrary to the aim of making the evaluation process easy and efficient, which was the original reason for using the adapted short questionnaire by Zumbach et al. (2007). On the other hand, the test design did not make it possible to sharply distinguish knowledge and competences acquired in the course from pre-existing knowledge. This aspect was unfortunately neglected when designing the evaluation tool. For future use, the test should therefore be changed to contain precise, course-specific multiple-choice questions. This would make such a distinction possible, enabling one to identify and assess the actual learning outcomes.

It is not possible to draw any certain conclusion concerning item 19 of the questionnaire. Although the self-assessment seems to correlate with the test results to some degree, one cannot

[5] Questions 2-5 (case: Benjamin) and 2.1-2.4 (case: Lisa).

say for certain whether the knowledge used to answer the questions relating to item 19 was actually acquired in the *Kolumbus-Kids* theory-seminar. Therefore, it is not possible to decide whether the course enables students to better recognise and solve problems that occur in heterogeneous groups of learners. This is a serious shortcoming of the evaluation design.

For item 20, the situation is different, as the questions relating to this item[6] were more specific and demanded knowledge that could only have been acquired in the course. The questions dealt with certain characteristic types of pupils that were discussed in the theory-seminar, thus limiting the range of possible answers. The answers given by the students show that course-specific knowledge was indeed acquired by most participants, confirming the positive self-assessment in this area. From this, one can conclude that the *Kolumbus-Kids* theory-seminar allows students to develop a certain level of diagnosis competency, which means WH_1 is at least partly confirmed by the results of the questionnaire and the test. This is a gratifying outcome, as diagnosis-competency is a central element of the teaching profession, which was stressed in the Bremen Declaration in 2000 and in the KMK's resolutions of 2004 and 2008. The fact that the theory-seminar "Dealing with heterogeneity in teaching natural sciences" improves the students' diagnosis-competences means that it makes an important contribution to fulfilling the KMK's demands concerning competences. The evaluation of the seminar has shown that participation in the course leads to the development of competency 7 of the set *evaluating*.

WH_2 can be confirmed as well, as the test results fit the students' self-assessment for item 19 and item 20. For item 19, however, it should be kept in mind that it is not certain whether knowledge was actually acquired in the course or whether students relied on pre-existing knowledge and common sense to answer the questions. Therefore, the results should be interpreted with some caution.

It could be argued that some questions in the test were too difficult for the students, which would explain the divergent answers. Determining whether the difficulty of the test was appropriate is not an easy task, but the test was devised after consulting with the instructor teaching the course. Therefore, the tasks reflect his ideas of adequate difficulty.

One goal of this study was to determine the quality of the course, while the other one was to assess the quality of the tool used for evaluating the course's quality: The questionnaire has been shown to be of adequate quality, with working hypotheses WH_3, WH_4, $WH_{4.1}$, and $WH_{4.2}$ having been confirmed.

For the overarching scale "quality of the seminar", Cronbach's alpha is 0.817, which means a medium degree of reliability. Considering the relatively small number of items, this is a satisfactory consistency coefficient. Calculating the test-retest reliability was not useful, meaning it does not make sense to have two evaluations in one semester. The low test-retest reliability of evaluations 3 and 4 of only 0.429 shows the temporal instability of the questionnaire's items. Therefore, the questionnaire is only suitable for being used once per semester, at the end of term.

Construct validity was only tested for the sub-scale "learning outcome", because this sub-scale was adapted by adding items 19 and 20, meaning the construct validity calculated by Zumbach et al. (2007) for the original questionnaire had to be re-verified. Another reason for

[6] Question 1 (case *Benjamin*) and question 1 (case *Lisa*).

focusing on this sub-scale was that it is the one that the test relates to, meaning it is especially important to ensure that the changed sub-scale is suitable for making comparisons. The calculations reveal that the participants indeed based their self-assessment on the dimension "learning outcome". This is shown by the fact that items 16-20 have high correlations with one another and have high correlations with the same factor.

While the test should be reworked in order to solve the problems mentioned above, there are no reasons why the questionnaire would have to be changed for future use. In its current form, it is sufficiently valid and reliable and can therefore be used as a tool for evaluating the quality of seminars, thereby laying the foundation for further improving teacher training through the use of empirical methods.

References

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer Medizin Verlag.

Bühl, A. (2014). *SPSS 22: Einführung in die moderne Datenanalyse* (14th ed.). Hallbergmoos: Pearson Deutschland GmbH.

Dobbins, M., & Martens, K. (2012).Towards an education approach à la finlandaise? French education policy after PISA. *Journal of Education Policy*, 27(1), 23-43.

Fichten, W. (2012). Über die Umsetzung und Gestaltung Forschenden Lernens im Lehramtsstudium: Verschriftlichung eines Vortrags auf der Veranstaltung 'Modelle Forschenden Lernens' in der Bielefeld School of Education 2012. Retrieved from http://www.uni-oldenburg.de/fileadmin/user_upload/diz/download/Publikationen/Lehrerbildung_Online/Fichten_01_2013_Forschendes_Lernen.pdf

Figg, S. D., Rogers, K. B., McCormick, J., & Low, Renae (2012). Differentiating low performance of the gifted learner: Achieving, underachieving, and selective consuming students. *Journal of Advanced Academics*, 23(1), 53-71.

Grek, S. (2009). Governing by numbers: the PISA 'effect' in Europe. *Journal of Education Policy*, 24(1), 23-37.

Huber, L. (2009). Warum forschendes Lernen nötig und möglich ist. In L. Huber, J. Hellmer, F. Schneider (Eds.), *Forschendes Lernen im Studium: Aktuelle Konzepte und Erfahrungen*. Bielefeld: Universitätsverlag Webler.

KMK (2004a). Sekretariat der ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Ed.) (2004), Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

Lange, H. (2007). Föderales Handeln in einer nicht-föderalen Gesellschaft? Föderalismusreform und Bildungspolitik. *Erziehungswissenschaft. Mitteilungen der Deutschen Gesellschaft für Erziehungswissenschaften*, 18(35), 137-164.

Moosbrugger, H., & Kelava, A. (Eds.) (2008). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer Medizin Verlag.

Mutz, R., & Daniel, H.-D. (2008). Nutzung von Lehrevaluationsdaten für die Qualitätssicherung der Evaluationsinstrumente am Beispiel der Universität Zürich. *Beiträge zur Hochschulforschung*, 30(2), 34-55.

Pohlenz, P., Grindel, E., & Köpke, A. (2006). Zur Validität von Evaluationsergebnissen. Qualitätsdimensionen der Lehrerausbildung im Lichte zentraler Testgütekriterien. In W.

Schubarth, P. Pohlenz (Eds.): *Qualitätsentwicklung und Evaluation in der Lehrerbildung: die zweite Phase: das Referendariat*. Potsdam: Universitätsverlag Potsdam.

Rindermann, H. (1997). Die studentische Beurteilung von Lehrveranstaltungen: Forschungsstand und Implikationen für den Einsatz von Lehrevaluationen. In R.-S. Jäger, R.-H. Lehmann, G. Trost (Eds.), *Tests und Trends. Jahrbuch der Pädagogischen Diagnostik*, 11, 12-53.

Rindermann, H. (1998). Übereinstimmung und Divergenz bei der studentischen Beurteilung von Lehrveranstaltungen: Methoden zu ihrer Berechnung und Konsequenzen für die Lehrevaluation. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 19(2), 73-92.

Rindermann, H. (2003). Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. *Zeitschrift für Evaluation*, 3(2), 233-256.

Spiel, C., & Gössler, P. M. (2000). Zum Einfluß von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende. *Zeitschrift für Pädagogische Psychologie*, 14 (1), 38-47.

Terhart, E. (2007). Universität und Lehrerbildung: Perspektiven einer Partnerschaft. In R. Casale (Ed.), *Bildung und Öffentlichkeit: Jürgen Oelkers zum 60. Geburtstag*. Weinheim: Beltz.

Zumbach, J., Spinath, B., Schahn, J., Friedrich, M., Krögel, M. (2007). Entwicklung einer Kurzskala zur Lehrevaluation. In M. Krämer, S. Preiser, K. Brusdeylins (Eds.), *Psychodidaktik und Evaluation*. Vol. 4. Göttingen: V & R unipress.