122nd ASEE Annual Conference & Exposition
June 14 - 17, 2015
Seattle, WA
Paper ID #13350

Multiple Choice Questions that Test Conceptual Understanding: A Proposal for Qualitative Two-Tier Exam Questions

Mr. Dion Timmermann, Hamburg University of Technology

Dion Timmermann studied electrical engineering at Hamburg University of Technology, Hamburg, Germany. In his master thesis he worked on simulation methods for the signal and power analysis of high speed data links. He currently pursues his Ph.D. in the Engineering Education Research Group at Hamburg University of Technology, where he investigates students understanding in introductory electrical engineering.

Prof. Christian H Kautz, Hamburg University of Technology

Christian H. Kautz has a Diplom degree in Physics from University of Hamburg and a Ph.D. in Physics (for work in Physics Education Research) from the University of Washington. Currently, he leads the Engineering Education Research Group at Hamburg University of Technology.

Multiple-Choice Questions that Test Conceptual Understanding: A Proposal for Qualitative Two-Tier Exam Questions

1 Introduction

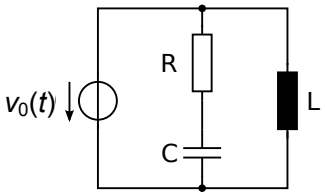
The goal of most university courses is for students to not only learn factual knowledge but to acquire conceptual understanding of the topics taught. Consequently, a course's assessment should at least in part evaluate this conceptual understanding.[1] To achieve this, there are multiple assessment methods that could be used, as for example essays or oral exams. However, many of these methods require a very high time investment on the part of the instructor, which is, in many cases, simply not possible. For large classes, multiple-choice tests are among the most efficient types of assessment. Although much care has to be taken in their development, machine-based scoring of multiple-choice tests can significantly reduce an instructor's work load, freeing up time for more face to face interaction with students. However, one main point of criticism of multiple-choice tests is that they only test factual knowledge and not conceptual understanding.[2,3]

Ideally, tests for conceptual understanding and multiple-choice questions could be brought together. One possible solution is the use of two-tier multiple-choice questions, where the first tier requires a fact-based response and the second tier a reasoning for that response. This type of questioning was probably first used by Tobin and Capie in 1981 in a test on logical thinking.[4] In 1985 Treagust presented an approach for the design of such questions for standardized tests on student misconceptions.[5] Since then, this question format has been used in several concept inventories and similar tests.[6,7,8] Instead of standardized tests, we started using such questions in exams. The results are very promising. Not only do these questions improve the quality of the exams by requiring a higher level of thinking, they also give us access to a large pool of data about the conceptual understanding of students at our institution, and thus allow us to monitor the prevalence of student misconceptions.

This paper aims to promote this type of two-tier multiple-choice questions, especially for exams, by giving examples and guidelines for their development. We will also show some properties (difficulty and discrimination) of eight sample questions and try to gauge the influence of the second tier statistically.

Exercise 2 (8 points)

The circuit at right contains an ideal alternating voltage source $v_0(t)$ with $v_0(t) = \hat{V}_0 \cdot cos(\omega t + \varphi_0)$. The voltages at the circuit elements are denoted by $v_R(t)$, $v_L(t)$ und $v_C(t)$; the corresponding currents by $i_R(t)$, etc. The current through the voltage source is denoted by $i_0(t)$.



Two sinusoidal signals have the same phase, if (choosing the respective sign) they reach their maximum at the same time.
The values for R, L, C, and ω are neither zero nor infinity, but otherwise unknown.

Please select for each question the correct answer as well as the corresponding reason.

Question 1 (2 points, if both items are correct)

Item 1.1	**$v_L(t)$ and $v_0(t)$**
A	are in phase,
B	are not in phase,
C	can have the same or a different phase,
Item 1.2	**because**
a	the inductance and the source are connected in parallel.
b	the inductance and the source are connected in series.
c	at the inductance the voltage leads.
d	between source and inductance, there is a branch with R and C.
e	the phase relation depends on the value of the capacitance.

Figure 1: A two-tier multiple-choice question in the format suggested in this paper, used in an exam in 2014.

2 Description and design of our two-tier multiple-choice questions

2.1 Basic structure of our two-tier multiple-choice questions

There are several different ways to formulate two-tier multiple-choice questions. The type of question used by our group is very similar to that used by e. g. Treagust[5], Lawson[6], and Chandrasegaran[7]. However, their questions were used in standardized tests and could be improved iteratively after being tested on students, while we use our questions in exams, where we do not have the possibility to test them on students beforehand. While the types of questions are very similar, our reasoning and design approach differs because of the different usage.

In the following, we are going to describe a specific type of two-tier question that we have found useful in the context of examinations. This description will be illustrated by the sample question presented in Figure 1. Further similar questions can be found at the end of the article in Figures 3, 4, and 5. In designing our questions we follow four central principles:

The first tier (item) asks students to apply their conceptual understanding to make a prediction about a specific situation. Thus, we do not ask students to recite a law or rule but rather to apply a law or rule to a specific situation. The goal of the exam question shown in Figure 1 was to assess if the students knew that the voltages of circuit elements connected in parallel are equal, and thus in phase. While we could have asked for this directly, students were asked the question shown in Figure 1, which requires them to *apply* their knowledge. This has two positive effects: On the one hand, we do not only expect students to be able to recite the law in question, but rather to be able to apply it. Thus, it is only logical to also test for this. On the other hand, this makes it easy to generate a large amount of different questions about the same topic without the possibility for students to memorize the correct answer. One can simply exchange the situation the students have to make a prediction about.

In the second tier (item), students have to choose one of several statements to justify their prediction made in the first tier. Ideally, the items in this second tier, i. e. the reasons offered to the students, are simple, factual statements that are either true or are believed to be true by many students. In Figure 1, reasons *a* and *d* are true statements. From informal classroom observations we knew that some students have an incorrect understanding of series connections, which would lead them select reason *b*. From previous research we also knew that a large fraction of students has an incorrect understanding of phase relations at reactive circuit elements.[9,10] This prompted us to add reason *c*. When grading, only a question where both, the first and the second tier, are marked correctly will be counted as correct. In our opinion, this combination of two items allows to better test students' conceptual understanding than with just the first item. This will be investigated in the following sections.

Ideally, there is no obvious relation between the items in both tiers, i. e. for each reason (second item) there is not just one answer (first item) it could apply to. This becomes easier to do if the reasons are factual statements. In Figure 1, reasons a, b, and d are simple statements about the topology of the circuit. Reason c is an oversimplified statement about the behavior of one of the circuit elements in question. There are two benefits from this: Students cannot simply select one reason or answer they know is correct and then select the respective item linked to it. While this increases the difficulty of the question and requires higher level thinking, it also increases the number of possible answers to the question, which is ideally the product of the number of answers and the number of reasons.

Distractors should be based on known student misconceptions. This rule is true for all multiple-choice questions.[5] Using known student misconceptions usually increases the difficulty of the question, but also increases the value of the answers for the teacher. It is easier to select the correct answer if all other choices are unattractive. However, if a student decides not to select a common misconception, one can be quite certain he does not hold it. Additionally, the percentage of students that have a certain misconception is usually surprisingly constant, at least among similar populations. This allows to predict the difficulty of a question, which is especially helpful when designing exams where it is usually impossible to test the questions beforehand. An example for this can be seen in Tables 1 and 2, which we will discuss in more detail later.

2.2 Tips for the design of such questions

There are many guides to the design of multiple-choice questions.[11,12,2,3] While much of the advice given in these is also true for two-tier questions, some aspects are different. In the following, we will point out the most important aspects for designing two-tier multiple-choice questions, focusing especially on the differences to single-tier questions.

Ideally, 2 answers and 4 to 5 reasons should be provided. For single-tier multiple-choice questions, 3 to 5 possible answers are suggested.[11,3] If a two-tier question has n_a answers and n_r reasons for students to select from, there are $n_\mathrm{a} \cdot n_\mathrm{r}$ possible combinations – given there are no obvious connections between answers and reasons. Thus, if one wanted to have 3 to 5 possible answers in total, each two-tier question had to have two answers and two reasons. Neither this, nor the other extreme of 5 answers and 5 reasons seems like a good solution to us. We prefer to give students four to five reasons to choose from. The number of answers to choose from should be lower. For each answer, there should be at least two reasons "that make sense". If there is a strong relation between answers and reasons, providing only two answers for students to choose from would be the logical consequence. The format of 2 answers and 4 to 5 reasons is also used by Treagust.[5]

Ideas for reasons can be generated from previous open-ended questions or literature. To generate statements that contain the possible reasons for students to select, Treagust first asked only single tier multiple-choice questions but required students to give a written reasoning for each answer.[5] While this approach is very useful for standardized tests, the first tiers of exam questions can hardly be given to students before their use in an exam. However, different questions about the same concepts but with different problem set-ups can be used. Alternatively, one can browse through the abundance of papers on known student misconceptions.

Reasons can contradict each other. When writing one-tier multiple-choice questions with three or more answers, one should not provide two answers that are the negation of each other. This allows a student to exclude all but these two answers by logic. Consider the question *"Which of the following properties should one-tier multiple-choice questions fulfill?"*. If, among the answers there are the choices *"Two answers may be the negation of each other."* and *"No two answers must be the negation of each other."* one of these two answers has to be true, making it unnecessary to consider any other possible answers. However, for the reasons of a two-tier multiple-choice question, this is different. If two reasons are the negation of each other, still neither could be the correct reason as they do not explain the correct answer.

Answers and reasonings that are too precise should be avoided. Many of the known student misconceptions sound absurd if they are formulated with precision, for example because they arise from students not differentiating between two related concepts like velocity and acceleration. Because of this, we often formulate the incorrect reasons quite vaguely. If the correct reason was formulated with precision, it would in many cases be the statement with the highest number of words. In these cases, a fine balance has to be found between not being obvious by too much precision, but still being precise enough so the reason is still correct.

Questions that require reasoning chains should be avoided. We found that it is difficult to ask two-tier multiple-choice questions with reasoning chains, i. e. where the reason consists of two or three statements that build upon each other or where several conditions have to be

Assertion		Reason
In a small open economy, if the prevailing world price of a good is lower than the domestic price, the quantity supplied by the domestic producer will be greater than the domestic quantity demanded, increasing domestic producer surplus.	BECAUSE	In a small, open economy, any surplus in the domestic market will be absorbed by the rest of the world. This increases domestic consumer surplus.

(a) True; True; Correct reason
(b) True; True; Incorrect reason
(c) True; False
(d) False; True
(e) False; False

(The correct answer is (d).)

Figure 2: Example of an assertion-reason question. Reproduction of Figure 2 from [13]

met in order for a law to apply. For these problems, other types of two-tier questions might be more suitable, as for example assertion-reason questions. One example of such a question from a publication by Williams [13] is reprinted in Figure 2. With these questions, two statements are given. One concerning a specific situation, and a general one. For both statements, students have to decide individually if they are true. If a student decides that both statements are true, he also has to indicate if the general statement is the reason for the specific one. This format of two-tier question allows to ask more complex questions, as the decisions students have to make are more limited. But as Williams points out, long statements require a high level of reading comprehension. [13]

3 Analysis of the sample questions

Standardized tests and concept inventories usually are tested and evaluated extensively before their actual use as a measuring instrument. This is done by analyzing the results of test cohorts of students. However, exams can usually not be shown to students before their administration. Thus, except for their face validity, exams can not be analyzed beforehand. However, after its administration, the results of an exam can be used to analyze it. From this analysis, instructors can at least judge the quality of their questions afterwards and with this newly gained knowledge hopefully generate better questions in the following year.

The questions presented in Figures 1 and 3 were given to 488 students in an exam in 2014. These students were first semester mechanical engineering and naval architecture students as well as third semester logistics and process engineering students who were enrolled in an introductory electrical engineering course that covers the basics of direct and alternating current circuit anal-

Question 2 (2 points, if both items are correct)

Item 2.1	**$i_C(t)$ and $i_R(t)$**
A	are in phase,
B	are not in phase,
C	can have the same or a different phase,
Item 2.2	**because**
a	at resistors the current is in phase with the voltage.
b	the resistor and capacitance are connected in series.
c	at capacitances the current leads.
d	the currents through resistor and source are in phase.
e	the phase relation depends on the value of the inductance.

Question 3 (2 points, if both items are correct)

Item 3.1	**$v_R(t)$ and $v_0(t)$**
A	are in phase,
B	are not in phase,
C	can have the same or a different phase,
Item 3.2	**because**
a	the resistor and the source are connected in parallel.
b	at resistors current and voltage are in phase.
c	a capacitance is connected in series with the resistor.
d	the inductance has an influence on the voltage of the source.
e	the phase relation depends on the value of the inductance.

Question 4 (2 points, if both items are correct)

Item 4.1	**$i_0(t)$ and $v_0(t)$**
A	are in phase,
B	are not in phase,
C	can have the same or a different phase,
Item 4.2	**because**
a	current and voltage are always in phase at ideal sources.
b	a capacitance and an inductance always compensate each other.
c	at ideal sources current and voltage always have a 90° phase difference.
d	the circuit contains an inductance.
e	the phase relation depends on the values of R, L and C.

Figure 3: Three two-tier questions from the 2014 exam. All questions concern the problem set-up presented in Figure 1.

(a) Question 1, n=484

Reason	Answer A	Answer B	Answer C	Sum
a	**49 %**	2 %	1 %	52 %
b	6 %	3 %	0 %	10 %
c	3 %	16 %	1 %	20 %
d	5 %	4 %	3 %	11 %
e	1 %	1 %	5 %	6 %
Sum	63 %	27 %	10 %	100 %

(b) Question 2, n=481

Reason	Answer A	Answer B	Answer C	Sum
a	2 %	2 %	1 %	5 %
b	**68 %**	6 %	0 %	74 %
c	1 %	14 %	0 %	16 %
d	1 %	1 %	1 %	4 %
e	0 %	1 %	1 %	2 %
Sum	72 %	24 %	4 %	100 %

(c) Question 3, n=481

Reason	Answer A	Answer B	Answer C	Sum
a	8 %	2 %	1 %	11 %
b	15 %	6 %	0 %	22 %
c	2 %	**49 %**	4 %	55 %
d	0 %	4 %	1 %	5 %
e	0 %	2 %	5 %	7 %
Sum	25 %	64 %	11 %	100 %

(d) Question 4, n=480

Reason	Answer A	Answer B	Answer C	Sum
a	10 %	0 %	0 %	11 %
b	1 %	1 %	1 %	4 %
c	1 %	13 %	1 %	15 %
d	1 %	6 %	1 %	7 %
e	1 %	17 %	**45 %**	63 %
Sum	15 %	37 %	47 %	100 %

Table 1: Responses to two-tier questions from Figures 1 and 3 in an exam. Correct answers are printed in bold, values below guessing probability in gray.

ysis and gives an overview of three-phase current systems. The course is mostly taught in a traditional manner with 90 minutes of lecture and 45 minutes of recitation section each week. The students' answers and reasons to these questions are presented in Table 1.

In 2015, the exam of the same course contained the questions shown in Figures 4 and 5. Due to changes in the universities curriculum, the course was now only attended by the logistics and process engineering students, and thus only 170 individuals participated in the exam. Except for this, no changes were made to the course. The students answers and reasons are given in Table 2.

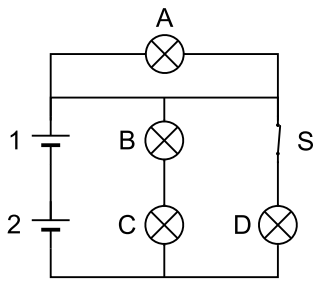
In a previous section, we suggested that the combination of both tiers, i. e. asking for an answer and a reason for that answer, allows to better test for conceptual understanding than when just using the first tier. To examine this suggestion, we will analyze students answers to these questions two-fold. Firstly, we will judge the distribution of answers and reasons selected by the students. Secondly, we will analyze the influence of the second tier on the questions difficulty and discrimination.

3.1 Distribution of answers

The 2014 exam had 15 answer/reason combinations for each question, resulting in an guessing probability of 6.7 %, while the 2015 exam even had 20 answer/reason combinations resulting

Exercise 1 (8 points)

The batteries 1 and 2 in the circuit at right are identical and can be treated as ideal voltages sources. The long line indicates the positive terminal of the respective battery. All four bulbs are identical.



Question 1 (2 points, if both items are correct)

Switch S is closed.

Item 1.1	***Bulb B is***
A	off, e.g. does not glow,
B	less bright than bulb C, but not off,
C	equally bright as bulb C,
D	brighter than bulb C,
Item 1.2	**because**
a	the current from bulb B is used up.
b	the same current flows through both.
c	a part of the voltage drops at bulb B.
d	bulb C has a lower potential than bulb B.
e	the electrons flow through bulb C first.

Question 2 (2 points, if both items are correct)

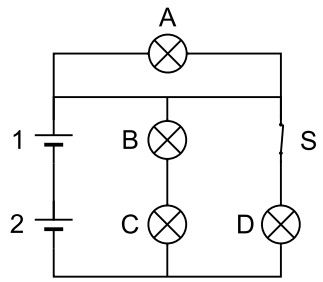
Switch S is closed.

Item 2.1	***Bulb C is***
A	off, e.g. does not glow,
B	less bright than bulb D, but not off,
C	equally bright as bulb D,
D	brighter than bulb D,
Item 2.2	**because**
a	there is less voltage at bulb C.
b	bulb B is parallel to bulb D.
c	current chooses the path of least resistance.
d	the current splits equally.
e	bulb C is closer to the batteries.

Figure 4: The first two two-tier questions from the 2015 exam.

Question 3 (2 points, if both items are correct)

Switch S is closed. The circuit is shown again at right.

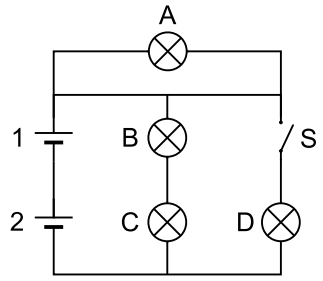


Item 3.1	***Bulb A is***
A	off, e.g. does not glow,
B	less bright than bulb D, but not off,
C	equally bright as bulb D,
D	brighter than bulb D,

Item 3.2	**because**
a	it shares the voltage with bulbs B, C, and D.
b	it is in a closed circuit.
c	it is not connected in parallel with any of the other bulbs.
d	there are losses in the wires.
e	there is no voltage at it.

Question 4 (2 points, if both items are correct)

Switch S is now open. The circuit is shown again at right.



Item 4.1	***The voltage at bulb D is***
A	equal to 0,
B	between 0 and the voltage of *one* battery,
C	equal to the voltage of *one* battery,
D	larger than the voltage of *one* battery,

Item 4.2	**because**
a	V = R * I only holds at closed circuits.
b	there is no current through it.
c	it is parallel to the two batteries.
d	there is no voltage at open switches.
e	there can be voltage at open switches.

Figure 5: Two-tier questions from the the 2015 exam. Both questions concern the problem set-up presented in Figure 4.

in a guessing probability of $5.0\,\%$. To improve the readability of Tables 1 and 2, values below guessing probability are printed in gray.

For most of the questions, only 2 or 3 answer/reason combinations had percentages above guessing probability. Each correct answer was selected by the majority of students with at least $45\,\%$ making that choice.

The questions in the 2014 exam were designed with a specific common misconception in mind.

(a) Question 1, n=170

	Answer				
Reason	A	B	C	D	Sum
a	1 %	1 %	0 %	0 %	1 %
b	0 %	0 %	**91 %**	0 %	91 %
c	0 %	1 %	4 %	2 %	6 %
d	0 %	0 %	0 %	0 %	0 %
e	0 %	1 %	0 %	1 %	2 %
Sum	1 %	2 %	95 %	2 %	100 %

(b) Question 2, n=170

	Answer				
Reason	A	B	C	D	Sum
a	0 %	**69 %**	1 %	0 %	71 %
b	0 %	9 %	1 %	0 %	11 %
c	0 %	9 %	0 %	0 %	9 %
d	1 %	3 %	6 %	0 %	9 %
e	0 %	0 %	0 %	0 %	0 %
Sum	1 %	91 %	8 %	0 %	100 %

(c) Question 3, n=169

	Answer				
Reason	A	B	C	D	Sum
a	0 %	5 %	9 %	2 %	17 %
b	2 %	3 %	5 %	1 %	11 %
c	5 %	2 %	3 %	11 %	21 %
d	0 %	0 %	0 %	0 %	0 %
e	**50 %**	1 %	0 %	0 %	51 %
Sum	57 %	11 %	17 %	15 %	100 %

(d) Question 4, n=170

	Answer				
Reason	A	B	C	D	Sum
a	5 %	1 %	0 %	0 %	5 %
b	**57 %**	1 %	1 %	1 %	60 %
c	0 %	0 %	4 %	5 %	9 %
d	8 %	0 %	1 %	0 %	8 %
e	1 %	9 %	3 %	5 %	18 %
Sum	70 %	11 %	8 %	11 %	100 %

Table 2: Responses to two-tier questions from Figures 4 and 5 in the 2015 exam. Correct answers are printed in bold, values below guessing probability in gray.

At reactive circuit elements, there is a phase shift between voltage and current. While this applies to the voltages across an element and the current though that element, many students misunderstand this rule and think there is a phase shift between the current through the element and the voltage across the circuit's voltage source, or vice versa.[9] Students who have this misconception would probably select the answer/reason combination B/c in questions 1 and 2 and A/b in Question 3. These answer/reason combinations in questions 1, 2, and 3 were selected by about 15 % of the students, and the second most common answer/reason combinations for all three questions. Only in Question 3 there is one other combination above guessing probability. From this, we can conclude that the other answer/reason combination in these three questions – that were rather made up and not based on previous research – were not attractive to students. This is different for Question 4, which does not test for the same misconception. None of the reasons of Question 4 were based on known misconceptions.

The questions in the 2015 exam were not as much focused on a single misconception, but rather to monitor students responses to the five-bulbs-test[14]. They were intentionally similar to questions asked in previous exams. The four answering options were the same for all four questions. While this is more then we recommend at the beginning of the paper, this allowed us to have a consistent structure for the answers.

Question 1 was to test if students understood that two bulbs in series are equally bright, because of the current through them being the same. The common misconception here is that the second

bulb is less bright as either current, voltage or potential is used up. However, the vast majority of students selected the correct answer/reason combination, as we expected based on answers to similar questions in previous years. Question 2 tested if students were able to correctly apply Kirchhoffs' Current Law. Reasons b, c and d were given quite often in open ended tests. As we expected, these were selected by $9\,\%$, $9\,\%$, and $6\,\%$ of the students, respectively. Question 3 tested if students were able to identify and correctly describe a short-circuited battery. Reason b was observed quite often in free response questions, while the other reasons were made up. However, the most common incorrect answer/reason combinations were not the expected one. Question 4 was designed to test students' understanding of the voltage in branches with open switches. The reasons were based on informal observation. We did not expect a particular distribution of answers.

3.2 Description of different measures for multiple-choice questions

To analyze the influence of the second tier on the questions, we will take a closer look at the difficulty and discrimination of the eight sample questions shown in Figures 1, 3, 4, and 5.

The difficulty index p of a question is simply the fraction of students that answered it correctly. As Aubrecht and Aubrecht note, the ideal difficulty index p_{ideal} of a multiple-choice question is not 0.5, but rather the value midway between the probability to guess the answer and 1, the difficulty if everyone answered correctly. A p near p_{ideal} maximizes the possible spread of scores in the whole test.[2]

A second important property of an exam question is its discrimination. If a question has a high discrimination, those students that perform well in the whole test also have the tendency to correctly answer that question. We use two different measures for discrimination: the discrimination index D and the correlation coefficient, calculated as a point biserial correlation coefficient ρ_{bis}. Both measures are numbers between -1 and 1, with larger values being preferable. The points students received for answering a question correctly was not the same for each question in the exam. For the calculations in this paper, we only considered if or how many questions a student answered correctly, i. e. we ignored the weighting caused by the different number of points per question.

To calculate the discrimination index, two subgroups of students are selected: the group of the $27\,\%$ highest scoring students in the exam, called group H, and group of the $27\,\%$ lowest scoring students, called group L. The discrimination index of a question is

$$D = p_{\text{H}} - p_{\text{L}}, \tag{1}$$

with p_{H} denoting the percentage of students from group H that answered the question right, and p_{L} denoting the percentage of students from group L that answered the question right.[15] Note that while we decided to use the de-facto standard group-size of $27\,\%$, proposed by Kelley[16], there are other proposals for the size of the groups.[15]
While there is no definite standard on which values for D indicate a good question, the consensus seems to be that questions with a $D < 0.2$ should be reworked and questions with $D > 0.4$ are rather good.[15,2,17]

			Both tiers evaluated				Only first tier evaluated				Difference		
Exam	N	Question	D	ρ_{bis}	p	p_{ideal}	D	ρ_{bis}	p	p_{ideal}	ΔD	$\Delta\rho_{\text{bis}}$	Δp
2014	488	1	0.76	0.56	0.48	0.53	0.71	0.57	0.63	0.66	0.05	−0.01	−0.15
		2	0.65	0.54	0.66		0.57	0.50	0.72		0.08	0.04	−0.06
		3	0.65	0.51	0.48		0.69	0.55	0.64		−0.04	−0.04	−0.16
		4	0.52	0.41	0.44		0.49	0.39	0.46		0.03	0.02	−0.02
2015	170	1	0.08	0.11	0.91	0.53	0.04	0.13	0.95	0.63	0.04	−0.02	−0.04
		2	0.51	0.41	0.69		0.18	0.28	0.91		0.33	0.13	−0.22
		3	0.71	0.55	0.50		0.64	0.53	0.57		0.07	0.02	−0.07
		4	0.46	0.37	0.57		0.29	0.23	0.70		0.17	0.14	−0.13

Table 3: Comparison of discrimination index D, correlation coefficient ρ_{bis}, difficulty index ρ, and ideal difficulty index p_{ideal} of all eight questions for when only the first tier is evaluated and when both tiers are evaluated. The difference between the values for one and two tiers is also listed.

Alternatively, the point biserial correlation coefficient, a correlation coefficient between the question and the whole exam, can be calculated as

$$\rho_{\text{bis}} = \frac{\mu_1 - \mu}{\sigma}\sqrt{p/(1-p)}, \tag{2}$$

with μ_1 denoting the mean test score for those students that answered the question correctly, μ denoting the mean test score for all students, σ denoting the test's standard deviation, and p denoting the difficulty of the question.[15] Values of ρ_{bis} should be grater than 0.20.[17]

3.3 Influence of the second tier on the measures

Table 3 gives an overview of the statistical properties of all eight sample questions. All measures are given two times: once considering only the first tier (the answer but not the reason) of each question, and again for when considering both tiers (the answer and the reason). This allows us to estimate the influence of the second tier. It has to be noted, though, that this comparison is not ideal, as we will simply ignore the reasons the students selected, even though they did see them and were possibly influenced by this.

As can be seen in Table 3, the discrimination index of all four questions from the 2014 exam is well above 0.4, even if only the first tier is considered. For one of the four questions the discrimination index decreases slightly, for the other three, it increases slightly, resulting in a mean increase of $+0.03$. The change of the correlation coefficient is even less strong. For two questions it rises, for two it falls slightly, evening out to a mean change of ± 0.00.

The questions in the 2015 exam are a bit more diverse. Question 1 has a discrimination index that is close to 0. This low discrimination index is to be expected for a question with a difficulty index of above $90\,\%$. Even when both tiers are evaluated, the question still has such a low discrimination index that it would not be suitable for a concept inventory[15]. However, this question was the first of the whole exam and included as a warm-up question.

Questions 2 through 4 from the 2015 exam all show an increase in the discrimination index and the correlation coefficient when the second tier is evaluated. While the increase for Question 3 is similar to that observed in the 2014 exam, questions 2 and 4 show more dramatic increases of both values, together with strong decreases of the difficulty index.

As all students who selected a correct answer *and* reason also selected the correct answer, the difficulty index can only decrease when the second tier is evaluated. For a question's discrimination index and/or correlation index to change, the number of correct answers has to change when the second tier is evaluated. Thus, for all questions which have a large increase of the discrimination index, the difficulty index decreased. The opposite, though, is not true. Not all questions with a strong change in the difficulty index also have a strong change of the discrimination index.

4 Summary and discussion

Although the creation of two-tier questions is more difficult and time consuming than the design of single-tier questions, we think it is well worth the effort. The stronger alignment between what is tested in an exam and what students are supposed to learn, as well as the opportunity to monitor students' conceptual understanding and the frequency of common misconceptions are both important to us and our work.

Although the addition of the second tier did not always increase the discrimination index or correlation coefficient of a question, it did never reduce it by much. Thus, while it might not always be beneficial in increasing the discrimination, it does so sometimes and at the same time increases our insight into students beliefs.

However, it might be impossible to create exams that only use this type of question. In the 2014 exam, for example, we were not able to devise a question that asked about the voltage across the open switch in Question 4 in Figure 5. On the one hand, we did not find a correct reason that was still general enough to not be obviously the correct one, while on the other hand, the distractors we wanted to use were too similar to allow an useful analysis.

Additionally, it might be interesting to investigate student's opinions on this type of questions, especially regarding their perceived difficulty. A comparison study with a control group would also be a good way to further gain insight into this type of questions. As we used our questions in exams it was not possible for us to do this kind of research while not interfering with the exam itself.

We are currently experimenting with the re-use of questions of this type by modifying the problem set-up without significantly changing the answers and reasons.

References

[1] Biggs, J. (1996) Enhancing teaching through constructive alignment. *Higher education*, **32**, 347–364.

[2] Aubrecht, G. J. and Aubrecht, J. (1983) Constructing objective tests. *American Journal of Physics*, **51**, 613.

[3] Hudson, H. T. (1981) Suggestions on the construction of multiple-choice tests. *American Journal of Physics*, **49**, 838.

[4] Tobin, K. G. and Capie, W. (1981) The Development and Validation of a Group Test of Logical Thinking. *Educational and Psychological Measurement*, **41**, 413–423.

[5] Treagust, D. (1986) Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, **16**, 199–207.

[6] Lawson, A. E. (1978) The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, **15**, 11–24.

[7] Chandrasegaran, A. L., Treagust, D. F., and Mocerino, M. (2007) The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, **8**, 293–307.

[8] Tan, K. C. D., Goh, N. K., Chia, L. S., and Treagust, D. F. (2002) Development and application of a two-tier multiple choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis. *Journal of Research in Science Teaching*, **39**, 283–301.

[9] Kautz, C. (2008) Probing student understanding of basic concepts and principles in introductory electrical engineering courses. *Proceedings of the 36th Annual SEFI Conference*, Aalborg.

[10] Kautz, C. (2011) Development of instructional materials to address student difficulties in introductory electrical engineering. *Proceedings of the 1st World Engineering Flash Week*, Lisbon, Portugal.

[11] Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. (2002) A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, **15**, 309–333.

[12] Haladyna, T. M. and Downing, S. M. (1989) Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, **2**, 51–78.

[13] Williams, J. B. (2006) Assertion-reason multiple-choice testing as a tool for deep learning: a qualitative analysis. *Assessment & Evaluation in Higher Education*, **31**, 287–301.

[14] McDermott, L. C. and Shaffer, P. S. (1992) Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding. *American Journal of Physics*, **60**, 994–1003.

[15] Marx, J. D. (1998) *Creation of a Diganostic Exam for Introductory, Undergraduate Elecrtricity and Magnetism*. Ph.D. thesis.

[16] Kelley, T. L. (1983) The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, **12**, 17–24.

[17] Engelhardt, P. (2009) An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests. *Getting Started in PER*, vol. 2.