Digit Exp Math Educ (2018) 4:139–162
https://doi.org/10.1007/s40751-018-0040-9

The Interplay between Inspectable Student Models and Didactics of Statistics

Sietske Tacoma[1] · **Sergey Sosnovsky**[2] · **Peter Boon**[1] · **Johan Jeuring**[2] · **Paul Drijvers**[1]

Published online: 25 April 2018

Abstract Statistics is a challenging subject for many university students. In addition to dedicated methods of didactics of statistics, adaptive educational technologies can also offer a promising approach to target this challenge. Inspectable student models provide students with information about their mastery of the domain, thus triggering reflection and supporting the planning of subsequent study steps. In this article, we investigate the question of whether insights from didactics of statistics can be combined with inspectable student models and examine if the two can reinforce each other. Five inspectable student models were implemented within five didactically grounded online statistics modules, which were offered to 160 Social Sciences students as part of their first-year university statistics course. The student models were evaluated using several methods. Learning curve analysis and predictive validity analysis examined the quality of the student models from the technical point of view, while a questionnaire and a task analysis provided a didactical perspective. The results suggest that students appreciated the overall design, but the learning curve analysis revealed several weaknesses in the implemented domain structure. The task analysis revealed four underlying problems that help to explain these weaknesses. Addressing these problems improved both the predictive validity of the adjusted student models and the quality of the instructional modules themselves. These results provide insight into how inspectable student models and didactics of statistics can augment each other in the design of rich instructional modules for statistics.

Keywords Inspectable student model · Open student model · Statistics education · Higher education · Learning curve analysis

✉ Sietske Tacoma
s.g.tacoma@uu.nl

[1] Freudenthal Institute, Utrecht University, P.O. Box 85170, 3508 AD Utrecht, The Netherlands

[2] Department of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TBUtrecht, The Netherlands

Statistical methods are highly relevant for conducting research in many fields of science. Therefore, many university programs include introductory statistics courses (Castro Sotos et al. 2007), which are often challenging for students (Murtonen and Lehtinen 2003; Tishkovskaya and Lancaster 2012). This is partly due to the complexity of the domain itself (Castro Sotos et al. 2007), and partly to the large size of the groups of students to whom these courses are taught, which greatly reduces teachers' ability to provide individual guidance to students.

From the field of statistics education research, various suggestions for enhancing statistics education have emerged in the past decades. A major change that has taken place concerns the main goals of statistics courses. Whereas traditionally the primary focus was on deriving statistical formulae and carrying out calculations, nowadays much more attention is paid to the interpretation of data and the ability to reason statistically about real-world problems – also referred to as 'statistical literacy' (Lovett and Greenhouse 2000). This shift in goals is partly evoked by the large-scale availability of statistical software that can take care of calculations. Accomplishing this shift involves specific didactical considerations in instructional design, such as using real contexts and data for promoting meaningful statistical reasoning (Ben-Zvi 2000).

Another possible enhancement of statistics education, which is especially relevant when individual guidance by teachers is difficult to achieve, comes from a different area: adaptive educational technologies (Herder et al. 2017). These technologies help convert results of automated assessment into detailed information for students and teachers, including diagnostic feedback (Stacey and Wiliam 2013). In the case of statistics education, with its challenging number of concepts to master, it seems particularly promising to provide students with information on their mastery of these individual concepts. One popular adaptive educational technology for providing such information is the inspectable student model (Bull and Kay 2007).

A student model is a structured collection of information about the individual student's characteristics, such as knowledge, difficulties and misconceptions, in the domain of study. Adaptive educational systems elicit this information based on students' interaction with learning content: solving tasks, taking tests, studying examples, etc. Presenting this information to students as feedback and allowing them to inspect it freely is known to promote reflection, increase motivation and provide metacognitive support for self-regulated learning (Bull and Kay 2007). In other words, an inspectable student model can support a student in forming an opinion about his or her current progress and making a well-considered decision about the next learning step (which concepts to focus on, which task to attempt, etc.).

However, the effectiveness of such an enhancement of the learning process in many respects depends on whether inspectable student models can be combined with the employed didactical approach. In the context of this article, the question is: how can the fields of didactics of statistics and inspectable student models be integrated? And can they strengthen each other?

To address these questions, inspectable student models were implemented in five modules containing practice exercises related to introductory statistics. These modules were embedded in an online educational system and were offered to 160 students in the Social Sciences as a part of their first-year, introductory statistics course. The inspectable student models were evaluated from two standpoints: the perception of the students who worked with them and their internal quality. Students' perceptions were collected through a questionnaire and served to evaluate whether combining the fields of didactics of statistics and inspectable student models was appreciated by students. For the quality analysis, evaluation methods from both fields were used. This quality analysis served two goals: to evaluate whether the implemented student models were successful and to explore how this implementation could be improved. Four main problems in the implementation were identified, for which solutions were sought both in the student model design and in the instructional design of the statistics modules.

Theoretical Background

Before attempting to combine the two fields of didactics of statistics and inspectable student models, we would like to explore both fields separately. In this exploration, we explicate difficulties that students experience in statistics education and examine how they are addressed both by didactical methods (i.e. methods informed by domain-specific pedagogical considerations) and by the information provided to students through inspectable student models. Moreover, we look for differences between the two fields that might lead to challenges in integrating them.

Didactics of Statistics

Research in statistics education has identified several causes for the challenging character of statistics. First of all, the field of statistics involves a large number of abstract concepts, such as probability distributions, sampling variability and confidence intervals. Second, constructing sound statistical conclusions requires the ability to integrate such abstract concepts both into calculations and into complex chains of reasoning (Castro Sotos et al. 2007). For example, understanding the method of hypothesis testing requires knowledge of probability distributions, sampling variability and significance levels, as well as the ability to reason using conditional statements (e.g. “under the assumption that the null hypothesis is true, this outcome, or a more extreme one, is very unlikely”). Finally, abstract definitions of statistical concepts such as variability often conflict with students’ prior, informal knowledge and their view of the real world (Garfield and Ahlgren 1988).

To support students in overcoming these challenges – that is, in gaining understanding of these abstract concepts, calculations and chains of reasoning – various strategies are prevalent in statistics education. Recommendations by Ben-Zvi (2000) and the GAISE college report (Garfield et al. 2005) include the use of real data sets and a focus on conceptual understanding and statistical reasoning, rather than mere acquisition of knowledge of procedures. Real data sets can engage students in thinking about the data and relevant statistical concepts. The recommendation to focus on conceptual understanding and statistical reasoning rather than on procedures is based on the assumption that students with a good conceptual foundation will easily grasp new procedures and techniques, whereas procedural knowledge without conceptual understanding tends to be too superficial and not well integrated.

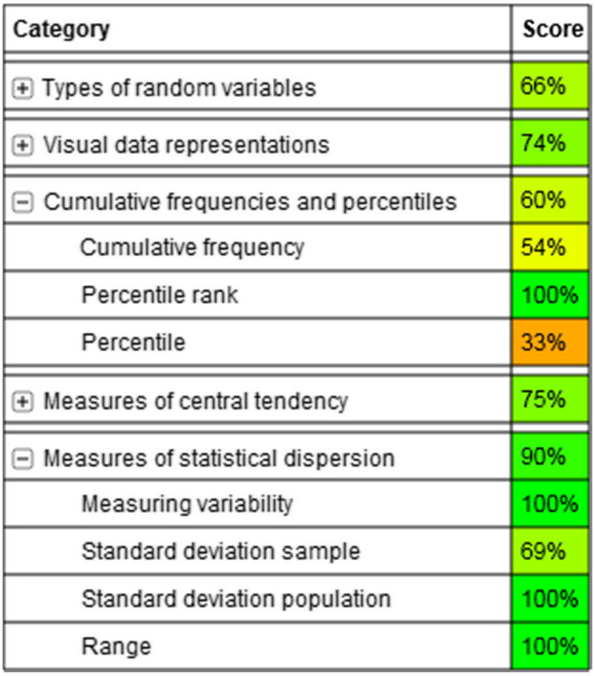
These insights may also guide instructional design. Taking real data sets from real contexts as a starting point for instructional design results in clusters of tasks that are related to each other through these contexts. A single context may, for example, be used for comparing different representations of the data, calculating and interpreting confidence intervals, and carrying out hypothesis tests. The sequencing of such closely related tasks is crucial (Drijvers et al. 2013). Deliberate task sequencing can serve to introduce concepts gradually, first informally and only later in a more formal way (e.g. Aberson et al. 2003) or to evoke crises to promote deeper reflection (Bokhove and Drijvers 2012). When exploring a context, earlier tasks are typically aimed at becoming acquainted with the context and data, whereas in later tasks the by-now familiar context can serve as a concrete example – and hence support understanding of more abstract concepts and their interrelationships. In other words, well-considered clustering and sequencing of tasks in instructional design is essential both for engaging with real contexts and for addressing conceptual understanding and statistical reasoning.

Inspectable Student Models

Student models are the core components of adaptive intelligent educational systems. They infer, store and update a system's estimations of the current knowledge state of each individual student, thus providing a basis for adaptively optimized support that the system can offer. A frequently used student model organization is an overlay model, which computes individual student mastery scores for a set of *knowledge components*: important concepts, methods or other coherent pieces of domain semantics (Carr and Goldstein 1977). In combination, these knowledge components constitute a model of the domain under study. A (partial) example of such a domain model is shown in the left-hand column of the inspectable student model displayed in Figure 1. In this example, the knowledge components are grouped into five categories and for two of them individual knowledge components are shown.

The knowledge components of different domain models can differ in several aspects, thus allowing for tailoring the model design to specific characteristics of the domain and the educational setting at hand. First of all, knowledge components can represent elements of procedural knowledge ('how'- knowledge) that define procedures or skills in the domain or they can represent declarative knowledge ('what'-knowledge) that define important concepts and facts (Brusilovsky and Millán 2007). For statistics education, in which a focus on conceptual understanding is advocated, this latter type seems more appropriate. A second layer of diversity comes from the degree of granularity. A designer of a model might decide to break the knowledge in the domain into as small elements as possible, thus improving the potential precision of the model. She might also decide to define knowledge components at the level of larger categories and topics, thus facilitating easier content modeling – connecting learning tasks to knowledge components.

The student's mastery of the knowledge components (KCs) in the domain model is represented in an overlay: a set of scores that is usually based on the student's performance on learning tasks associated with corresponding KCs. The connection between tasks and KCs can be represented by a so-called Q-matrix (Barnes 2005; Tatsuoka 1983), with a row for each KC and a column for each task. The entry (i, j) is equal to 1 if the jth task is connected to the ith KC, i.e. if the ith KC is relevant to solving the jth task, and 0 otherwise. The scores in the overlay may be either qualitative (poor, medium, good), simple numerical (a percentage, for example) or uncertainty-based (Brusilovsky and Millán 2007).



Category	Score
⊞ Types of random variables	66%
⊞ Visual data representations	74%
⊟ Cumulative frequencies and percentiles	60%
Cumulative frequency	54%
Percentile rank	100%
Percentile	33%
⊞ Measures of central tendency	75%
⊟ Measures of statistical dispersion	90%
Measuring variability	100%
Standard deviation sample	69%
Standard deviation population	100%
Range	100%

Fig. 1 An inspectable student model on descriptive statistics

An example of an overlay is displayed in the right-hand column of the student model in Figure 1.

The main purpose of student models in adaptive educational systems is usually to provide a basis for adaptation. However, the information that the student model contains can also be used as valuable feedback for the student: if shown to the student, a student model can promote reflection and support both planning and navigation (Bull and Kay 2007). Reflection may, for instance, be promoted by a low score on a concept that a student thought she already had mastered and, as such, the open student model may reveal gaps in the student's knowledge of the domain. For these purposes, a fairly simple student model design may suffice.

Although sophisticated methods exist for enabling students to edit or negotiate with their student model (e.g. Dimitrova et al. 2001; Zapata-Rivera and Greer 2002), for the purpose of reflection, planning and navigation, promising results have been obtained with much simpler inspectable student models (Arroyo et al. 2007; Long and Aleven 2011; Mitrovic and Martin 2002) that do not allow a student to adjust the contents of the model (Bull 2004). Moreover, Bull and Kay (2007) argue that, in student models with the purpose of promoting reflection, the scores can be presented in a simple way, without mentioning the uncertainties surrounding them. Reflection is most likely evoked by differences between the model and the student's own view, which are presumably larger if uncertainty is omitted.

A final remark on student models concerns their relation to instructional design. Such models are often used to inform adaptation. In such cases, the instructional design includes variation of the order of tasks, depending on student achievement so far. To this end, the tasks need to stand alone rather than to be organized in a pre-structured sequence. Even in many cases where student models are rendered inspectable, they have initially been designed to inform adaptation and are therefore connected to a set of independent tasks, rather than to a sequence of closely related tasks that share contexts or build on one another.

To summarize, the main difficulties for students in statistics education are the large number of abstract concepts involved and the ability needed to integrate these concepts into calculations and chains of reasoning. Methods from the didactics of statistics to address these difficulties include both the use of real data sets and contexts and a focus on conceptual understanding. Inspectable student models provide an additional method by supporting students in gaining insight into the structure of the domain of statistics, as well as revealing knowledge gaps.

An important difference between the methods from the two fields lies in instructional design: the didactical methods result in sequences of closely related tasks that share contexts and build on one another, whereas inspectable student models are traditionally connected to sets of rather independent tasks. Therefore, our first question from the introduction – whether the fields of didactics of statistics and inspectable student models can be combined – can be explicated as follows: (RQ1) Are inspectable student models suitable for implementation in didactically grounded, sequential statistics modules consisting of closely related tasks? The second question, whether the two fields can strengthen each other, focuses on the evaluation methods available in both fields: (RQ2) How can didactical analysis inform design of inspectable student models and, vice versa, how can student model evaluation methods inform didactical design?

Methods

To address these research questions, inspectable student models were designed and implemented in five didactically grounded modules which were used in an introductory statistics course at Utrecht University. In the following sub-sections, we first describe the educational setting for this study, including a description of the online educational system that was used. Next, we discuss the didactical design of the modules and student model design. Lastly, we outline data collection and describe the methods we have used for analyzing the quality of the different components of the student models.

Educational Setting and the ‘Digital Mathematics Environment’

The participants in this study were 160 first-year students in the Social Sciences at Utrecht University. In the fall of 2016, these students took part in a mandatory statistics course as one of the first courses in their bachelor’s degree program. This course lasted ten weeks, and covered the following five topics:

1. Descriptive statistics.
2. z-values and sampling distributions.
3. Hypothesis testing: z-tests.
4. Hypothesis testing: t-tests for one sample and dependent samples.
5. Hypothesis testing: t-tests for independent samples.

Each topic consisted of a lecture followed by practice in a digital statistics module. Students were allowed to work on the modules individually or in groups and could choose to work at home or in supervised lab sessions.

The five digital modules were offered in the Freudenthal Institute's Digital Mathematics Environment (DME; see Drijvers et al. 2013). The DME offers support for a variety of interactions, such as number and formula input, multiple choice tasks, drag-and-drop tasks and interactive animations. Immediate verification feedback is provided for students' answers, informing students whether or not their answer is correct, but not what the correct answer is. Moreover, for most task types, elaboration feedback is available to explain errors that have been made. Students are allowed to attempt tasks multiple times and usually continue trying to solve each task until they succeed.

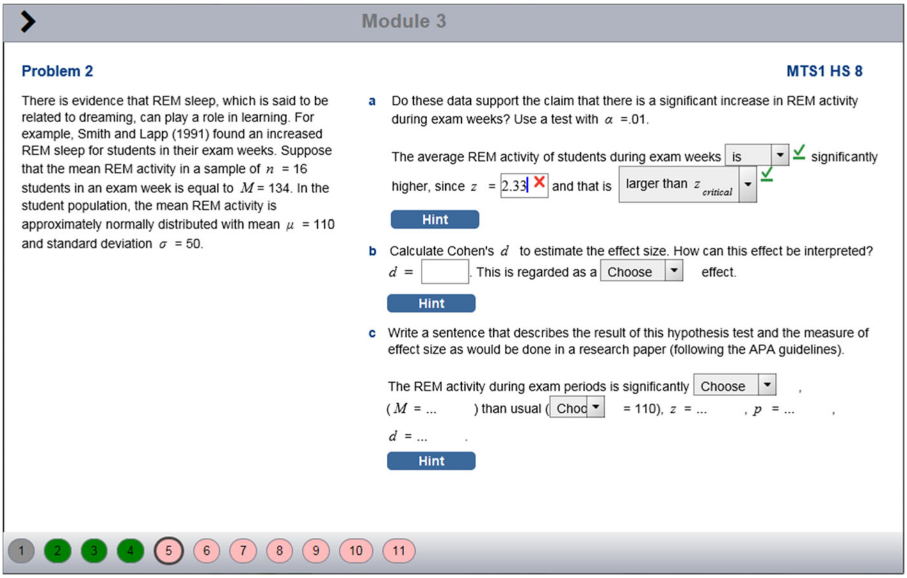
A typical DME page is shown in Figure 2. The circles in the bottom bar of the page indicate the student's progress in the module. These indicators turn green once the student has solved correctly all the tasks on the page while they remain red as long as this is not the case. As suggested in literature (Brusilovsky et al. 2009), such coloring of progress indicators can have a strikingly motivational effect: in order to obtain green progress indicators students keep attempting tasks until they find the correct answer. In Figure 2, the indicators reveal that this student has completed pages 2, 3 and 4 correctly, and still has to work on pages 5 to 11. Since page 1 only contains an introductory text and no tasks, its indicator is grey.

Didactical Design

The modules used in this study were designed by the lecturers for the statistics course, supported by DME experts. Each module consisted of a series of pages containing sets of closely related tasks. The number of pages varied between 12 and 22, while the number of tasks in the modules varied between 98 and 232. The page shown in Figure 2 is a translated version from the fifth page of the third module. It contains a context description on the left-hand side of the page and three sets of tasks on the right-hand side. Each individual interaction component is regarded as a task.

DME pages have a very flexible layout, which allows for different numbers of tasks on each page. Moreover, the DME facilitates initially hiding information that might not be needed by all students. The lecturers made extensive use of this option to include hints and extra tasks serving as intermediate steps. On the page shown in Figure 2, hidden information is available through the hint buttons. The information that is revealed upon clicking the top-most hint button is shown in Figure 3. Whereas students were obliged to complete all tasks on the main pages, use of these hints and intermediate tasks was optional. Moreover, use of the hidden information did not affect the page indicators, so these would turn green once all tasks on the main page were completed correctly.

As recommended in the literature, the modules made extensive use of real data sets and contained many tasks that focused on conceptual understanding and statistical reasoning. Most tasks in the modules were connected to a context and all contexts were based on real research projects and contained real data. In the modules, students were invited to engage deeply with these contexts. Contexts were used to address multiple concepts and to highlight aspects of the relations between concepts. On the example page in Figure 2, students are asked to carry out a hypothesis test, determine the effect



Module 3

Problem 2 MTS1 HS 8

There is evidence that REM sleep, which is said to be related to dreaming, can play a role in learning. For example, Smith and Lapp (1991) found an increased REM sleep for students in their exam weeks. Suppose that the mean REM activity in a sample of n = 16 students in an exam week is equal to M = 134. In the student population, the mean REM activity is approximately normally distributed with mean μ = 110 and standard deviation σ = 50.

a Do these data support the claim that there is a significant increase in REM activity during exam weeks? Use a test with α =.01.

The average REM activity of students during exam weeks is ▾ ✓ significantly higher, since z = 2.33 ✗ and that is larger than $z_{critical}$ ▾ ✓

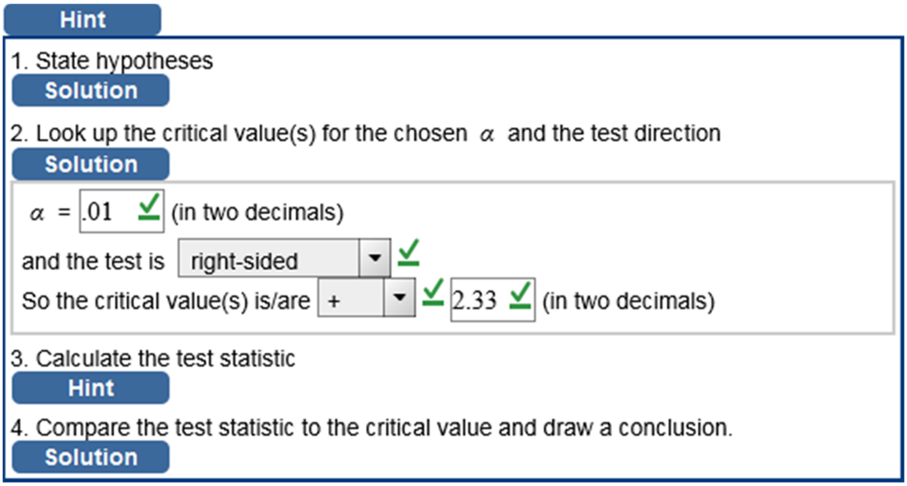
b Calculate Cohen's d to estimate the effect size. How can this effect be interpreted?
d = □. This is regarded as a Choose ▾ effect.

c Write a sentence that describes the result of this hypothesis test and the measure of effect size as would be done in a research paper (following the APA guidelines).

The REM activity during exam periods is significantly Choose ▾ , (M = ...) than usual (Choo ▾ = 110), z = ... , p = ... , d = ...

Fig. 2 A translated DME-page from the third module (on hypothesis testing).

size and report the results as would be done in a research article. Furthermore, contexts were deliberately varied to confront students with various applications and the appearance of various concepts: testing left-sided, right-sided or two-sided; positive and negative values of the test statistic; known and unknown population variances; significant and non-significant results, and so on. Conceptual understanding was, for example, addressed by tasks asking students to interpret the rejection of a null hypothesis in the given context or to describe the influence of sample size on concepts such as effect size or power. The number of procedural tasks was kept low by regularly using SPSS output, instead of asking students to calculate the test statistic themselves in all the tasks on hypothesis testing.



1. State hypotheses

2. Look up the critical value(s) for the chosen α and the test direction

α = .01 ✓ (in two decimals)

and the test is right-sided ▾ ✓

So the critical value(s) is/are + ▾ ✓ 2.33 ✓ (in two decimals)

3. Calculate the test statistic

4. Compare the test statistic to the critical value and draw a conclusion.

Fig. 3 Initially hidden intermediate steps for the DME page shown in Figure 2

The use of real data sets and a focus on conceptual understanding and statistical reasoning had consequences for the ordering of tasks on each page and for the ordering of the pages themselves. In Figure 2, the ordering of the three sets of tasks is determined by their content: the hypothesis test needs to be carried out before calculating the effect size or reporting the results. In the ordering of pages, difficulty level was taken into account, for example by introducing the more complicated *t*-test for independent groups after sufficient exposure to the easier *z*-test and *t*-tests for one group and for dependent groups. Finally, whereas concepts were typically addressed in isolation on earlier pages, later tasks required more and more understanding of combinations of and relations between concepts. For example, early pages contained separate sets of tasks for stating hypotheses, finding a critical value or calculating a test statistic, whereas later pages contained only one set of tasks asking the student to carry out a complete hypothesis test. After finishing the module, students were presented with their student model, which they could revisit any time after that.

Design of Inspectable Student Models

Student models were implemented in all five modules. The student models were devised by the first author, in collaboration with two experts: the main lecturer of the course and the fifth author. Three separate components were designed: domain models, Q-matrices with connections between KCs in the domain models and tasks in the modules, and a calculation method for computing overlay scores.

The first step in domain model design involved formulating KCs based on the tasks. Taking the tasks as the starting point may seem a reversed approach, since tasks are designed to cover a certain domain rather than vice versa. However, it is also an approach that lecturers or designers could easily pursue. Because of the large responsibility that university teachers have for designing their own instructional material, design feasibility was deemed important in the context of this study. Since the purpose of the student model was to promote conceptual understanding, KCs were mainly developed to represent declarative rather than procedural knowledge. To ensure that the model completely covered the domain in the end, the second step consisted of adding and adjusting KCs based on a consultation of already available domain models (ALEKS[1]), as well as other instructional material on the same topic (SURF[2]). In the third and final step of domain model design, the two experts were consulted and KC definitions were fine-tuned based on their comments.

A rather coarse-grained approach was adopted to design KCs, which means that they were relatively broad in scope. For example, instead of defining KCs for calculating different test statistics (*z*-test, *t*-test for one sample, and so on), a single one was defined for calculating the test statistic. Although finer-grained domain models generally allow for more sophisticated diagnoses (Sosnovsky and Brusilovsky 2015), we had two decisive reasons to opt for a coarser-grained approach. The first was student model comprehensibility, since the main purpose of the models was to offer students insight into their own understanding of the domain. The second was, again, design feasibility; in this approach, a quick analysis of tasks suffices to determine the KC(s) involved.

[1] Course materials downloaded on October 16, 2015, from www.aleks.com/about_aleks/course_products

[2] http://bit.ly/surfstat

For each module, a separate domain model was designed. However, since modules 3, 4 and 5 all covered hypothesis testing, their models overlapped to a large extent. The final ones contained between 8 and 19 KCs. To improve comprehensibility of the student models, the KCs in each domain model were grouped into two to five categories.

Design of the Q-matrices was straightforward. Tasks were connected to all KCs that were related to the task. For example, tasks that involved finding a critical value were connected both to the KC on the critical value and to the one on the significance level, since the latter is needed to find the former. The majority of tasks was connected to only one KC, but for some up to six KCs were judged relevant. To improve Q-matrix consistency, the two experts were invited to connect a subset of the tasks to the domain models. Most expert connections were the same as the researchers' and differences were discussed until consensus was reached.

The final component in student model design was the calculation method for overlay scores. This was based on the number of attempts students needed to finish the tasks connected to each KC. A straightforward numerical implementation was chosen, in which each task connected to a KC contributed equally to its score. The formula we used for calculation of the overlay scores was:

$$score_{KC,student} = \frac{\sum_{i=1}^{n} \frac{\sum_{j=1}^{m} a_{t_i,j}}{m}}{n}$$

with:

t_i the ith task connected to this KC ($i \in \{1, ..., n\}$).

$a_{t_i,j}$ the jth attempt score by this student for task t_i ($j \in \{1, \ldots, m\}$).

This formula can be explained as follows: for each task, a task score was calculated as the mean attempt score over all attempts by this student on this task. The attempt score was 0 for incorrect attempts, 0.5 for half-correct ones (for example, if the answer still needed to be rounded off) and 1 for correct ones. For instance, the task score for a student who first gave two incorrect answers before answering correctly was 0.33. The student's score for a KC was then calculated by averaging the task scores for all tasks connected to the KC.

Giving all tasks equal weight in the calculation of overlay scores may seem unfair, since students are likely to learn and hence perform better on later tasks than on earlier ones. However, tasks also tended to become more complicated throughout the modules, requiring students to combine several concepts rather than using them in isolation. Since students were invited to study their student model only at the end of each module, a final difficult task could easily result in an underestimation of the student's knowledge, if more recent tasks had been assigned a larger weight.

A translated example of a student model for the first module is presented in Figure 1. The domain model for this module contained five categories. In the inspectable student model, students could unfold each category (by clicking the + button) to view the individual KCs and their overlay scores. Category scores were calculated as the weighted mean of the KC scores in the relevant category, weighted by the number of tasks to which they were connected.

Data Collection

After the student models were implemented in the instructional modules, they were offered to the Social Sciences first-year students. Data collection focused on student perception (see RQ1) and student model quality (see both RQ1 and RQ2). To investigate student perceptions about the models, a short questionnaire was added at the end of each module, on the page in which students could inspect their student model. In this questionnaire, students were asked to respond to three statements, concerning the match between the tasks and the KCs, the clarity of the KC descriptions and the scores in the overlay. Students could indicate their degree of agreement or disagreement with each statement using a five-point Likert scale.

The log files containing student work on the five modules were the most important data source for evaluating the quality of the student model. Each week, the student work for that week's module was exported from the DME. The first module contained a page with information on this study and asked students for their consent. Work from students who did not give consent was deleted (26 out of 186) and all other log files were rendered anonymous before further analysis. For each module, all students who attempted at least one task were included in the analysis. Table 1 summarizes properties of the students' work and also provides the number of tasks in each module and the relevant number of KCs. As can be seen in the table, student numbers slowly decreased from 160 students in the first module to 117 in the fifth. This can be attributed to students quitting the course or choosing other means for studying the course material.

Data Analysis

Questionnaires were used to assess the suitability of the student models in the statistics modules (see RQ1) from the students' perspective. Each of the five modules contained a questionnaire on the last page, and each questionnaire contained three statements, to which students could respond on a five-point Likert scale. For each of the fifteen statements, a mean score over all students was calculated as a measure of agreement of the students to the statement.

For the evaluation of student model quality, methods from both the didactics of statistics and the student modeling fields were combined. First, a learning curve analysis (Martin et al. 2011) was carried out in order to assess domain model quality (RQ1) and to identify weaknesses in their design and implementation (RQ2). Next, these weaknesses were further investigated through didactical task analysis, which led

Table 1 Degree of data collection from the five modules

Module	Tasks	KCs	Students	Attempts per student (SD)	% attempted tasks (SD)
1	98	19	160	109 (36)	57 (14)
2	107	8	141	113 (43)	63 (17)
3	110	14	129	89 (40)	45 (17)
4	232	14	127	190 (75)	52 (18)
5	132	16	117	137 (66)	58 (18)

to possible improvements to both the student models and the instructional modules (RQ2). Finally, predictive validity analyses (Sosnovsky and Brusilovsky 2015) were carried out to assess both the quality of the overlays in the original design (RQ1) and those in the design, after implementing the improvements identified in the learning curve analysis and didactical task analysis (RQ2). In the following, the three methods, as well as our implementations, are described in more detail.

Learning curve analysis is specifically aimed at evaluating the domain model. The assumption behind learning curve analysis is that learning generally follows a power law. When first encountering a concept, students' incomplete understanding results in errors on tasks related to that concept. After more and more encounters with the concept, the students' understanding becomes more complete, resulting in a decrease in the number of errors related to the concept (Martin et al. 2011). In other words, for each KC in the domain model, the error rate is expected to decrease and, if this were indeed the case, the KC is regarded as a valid unit of knowledge.

To generate learning curves, first for each student and then for each KC, a student's attempts on tasks connected to the KC were sorted in chronological order. This resulted in lists of attempts, in which, for example, the sixth attempt could be the first attempt by a student on the sixth task connected to the KC, the sixth attempt by this student on the first task, or anything in between. The length of the lists varied over students and KCs, since students needed different numbers of attempts to finish the tasks connected to the different KCs.

After ordering the attempts, the correctness of each one was indicated. Because we were interested in the number of errors, we marked errors as 1 and other attempts as 0. Next, error rates for individual KCs were calculated for each attempt number, by dividing the number of students who made an error related to the current attempt number for a given KC by the total number of students who made an attempt for that attempt number for that KC.

$$Error\ rate\ \mathrm{n}th\ attempt = \frac{number\ of\ \ incorrect\ \mathrm{n}th\ attempts\ on\ a\ KC}{total\ \ number\ \ of\ \mathrm{n}th\ attempts\ on\ a\ KC}$$

These error rates were plotted against the attempt numbers and a power law was fitted, using the formula:

$$Error\ \ Rate = B{\cdot}AttemptNo^{-\alpha}$$

with the decay factor α and starting value B as parameters. Moreover, R^2 was calculated as measure of goodness of fit.

Since students could attempt tasks multiple times, and not all tasks were obligatory, the number of attempts for the different KCs varied across students. Consequently, the number of students decreased as the attempt number increased. That is, for higher attempt numbers, the error rates were based on attempts by fewer students.

To ensure reliable error rates, Martin et al. (2011) recommend cutting off the learning curve after a certain attempt number. They propose two methods for defining the cut-off point: either by selecting an acceptable reduction in the number of students or by making a judgement call on where to cut off after examining where the learning curve seems to be deteriorating. Martin and colleagues used a one-half cut-off, meaning that they cut a curve off once only half of the students remained. In examining the learning curves for our domain models, we noticed that many learning curves deteriorated already after losing one third of the students who made a first attempt. Therefore, we decided to use a two-thirds cut-off. This higher cut-off level can be explained by the large number of non-obligatory intermediate steps in the modules. As can be seen in Table 1, the average number of attempted tasks was considerably lower than the total number of tasks in each module, caused by students skipping the non-obligatory intermediate steps. Therefore, for high attempt numbers, error rates were predominantly based on students who made use of the intermediate steps. This was a smaller, and probably weaker, group of students than the complete student population and hence the error rate was likely to increase with this decrease in student numbers.

After assessing the quality of all individual KCs, some were found to have increasing rather than decreasing error rates. To explain these increasing error rates a didactical inspection of the instructional modules was carried out. To this end, single tasks and sets of similar tasks were repeatedly disconnected from the KC and new learning curves were generated. Once a decreasing learning curve was found, the set of tasks that was currently disconnected was designated as a possible cause for the originally increasing learning curve. Next, a didactical analysis of these tasks was performed to find a sound explanation for the increasing learning curve. In cases where it did not prove possible to designate just one task or set of similar tasks as a possible cause, all tasks connected to the KC were analyzed from a didactical perspective, and especially the concepts judged to be addressed in the tasks were reconsidered. Through this interplay between the learning curve analysis and the didactical task analysis, we attempted to improve the quality both of the student models and of the instructional models themselves.

Both the learning curve analysis and the subsequent didactical task analysis specifically targeted the domain model and did not address the overlays. Therefore, overlay quality needed to be assessed through a third method: predictive validity analysis. In predictive validity analysis, student performance is predicted based on the student's previous attempts. The correlation between these predictions and the actual student performances is used as a quality measure for the overlays. To find this correlation, the following two values were calculated for each KC involved in each attempt by each student:

- The student's prior knowledge level for the specific KC up to the current attempt.
- The student's posterior actual performance for the specific KC after the current attempt.

The prior knowledge levels were calculated based on the tasks that a student had already attempted, using our calculation method for overlay scores. Posterior student performance was based on attempts following the current attempt. Sosnovsky and Brusilovsky (2015) argue that correlating single step performances with knowledge predictions is problematic and propose using simple moving averages over five attempts. We followed this suggestion by selecting the first five attempts after the current attempt that also involved the current KC. For these five attempts, the average attempt score was calculated (again, correct = 1, half-correct = 0.5 and incorrect = 0) as a measure of actual performance.

Results

We first provide the results of the questionnaires about the students' perceptions of the models. Next, we present the main quality assessment of the implemented domain models: learning curve analysis. The results of this form the starting point for didactical task analysis. This leads to the identification of four problems in implementing inspectable student models in rich instructional modules for statistics and to possible improvements of the domain models, Q-matrices and the instructional modules to resolve these four problems. Finally, the results of the predictive validity analysis reveal the quality of the overlays, as well as the value of the improvements arising from the didactical analysis.

Student Perceptions of the Student Models

Table 2 summarizes the results of the questionnaires at the end of each module. A score of 1 corresponded with 'Totally disagree' and a score of 5 corresponded with 'Totally agree'. From the table, it can be seen that students agreed to a large extent with statements 1 and 2, and to a moderate extent with statement 3. The strong agreement with statements 1 and 2 suggests that students perceived the tasks in the modules and the KCs in the domain models to match well and that the descriptions of the KCs were clear. The moderate measure of agreement with statement 3 implies that students thought the scores from the overlays represented their current knowledge of the concepts quite well. All in all, students seemed to perceive the student models as comprehensible and plausible.

Domain Model Quality According to Learning Curve Analysis

To assess the quality of the domain models, learning curves were generated for 56 out of 71 KCs in the five domain models. For the remaining 15 KCs, not enough data was available to obtain a learning curve. For each of these 56 KCs, the error rates were computed and a power law curve was fitted. This resulted in decreasing learning curves for 34 KCs and increasing learning curves for 22 KCs.

For the 34 with decreasing learning curves, the mean goodness of fit (R^2) was 0.49 ($SD = 0.29$). For the increasing learning curves, the degrees of fit were generally weaker, with mean goodness of fit = 0.35 ($SD = 0.33$).

As mentioned before, KCs with learning curves that decrease as a power function represent cognitively valid units of knowledge. Therefore, according to this criterion, in the initial design 34 out of 71 KCs were well defined. It may seem disappointing that only half of the KCs were well defined, and this indeed suggests that just implementing student models in didactically grounded instructional modules does not automatically result in high-quality feedback to students. However, the presence of increasing learning curves also provided a good starting point for further analysis: didactical inspection of connected tasks may shed light on prerequisites, opportunities and limitations in implementing student models in didactically grounded statistics modules.

Table 2 Questionnaire results

	Module 1	Module 2	Module 3	Module 4	Module 5
Statement 1: The tasks in the DME match well with the topics in the student model					
Mean	4.30	4.42	4.26	4.44	4.43
SD	0.67	0.69	0.68	0.58	0.59
N	125	89	54	27	23
Statement 2: The descriptions of the topics are clear					
Mean	4.23	4.21	4.17	4.15	4.22
SD	0.73	0.82	0.91	0.82	0.80
N	125	89	54	27	23
Statement 3: I think the scores on the topics are a good representation of my knowledge					
Mean	3.85	4.08	4.10	4.00	3.90
SD	0.73	0.70	0.77	0.89	0.89
N	120	88	52	26	21

Underlying Problems Based on Didactical Analysis

For the 22 KCs with increasing learning curves, the connected tasks were analyzed from a didactical perspective to try to identify underlying problems that caused the learning curves to increase. Four different difficulties were identified; some increasing learning curves were completely explained by one of them, whereas for other KCs two of them applied. The first problem relates to single tasks distorting the learning curve, while the second concerns groups of tasks that address concepts from different perspectives and the third connects to tasks that involve multiple concepts. The fourth and final difficulty involves a lack of opportunities for in-depth thinking about the particular KC in the learning module. This final one also applies to the 15 KCs for which not enough data were available to obtain a learning curve. The four problems are elaborated below.

Tasks with Specific Purposes

For 12 out of the 22 KCs, the increasing learning curve could be attributed to one or two single tasks; disconnecting these tasks from their KC yielded a decreasing learning curve. Didactical analysis of the disconnected tasks, compared with the tasks that remained connected, revealed that these tasks often had a specific purpose in the module.

In six of these cases, the tasks that were disconnected were the first ones in which the students encountered the KC. An example is a KC on the significance level, for which the learning curve is shown in the left-hand graph in Figure 4. The error rate for the first attempt is remarkably lower than for subsequent attempts. Disconnecting the first task from the KC yielded the right-hand learning curve in Figure 4. Without the first task, the learning curve became decreasing with a very high goodness of fit (R^2 = .98), indicating that the remaining tasks constituted a valid KC.

A didactical inspection of the disconnected task and those that remained connected revealed that the first task was appreciably easier than the subsequent ones. The first task only asked students to reproduce a value for the significance level from the problem description. Later tasks required students to use the significance level for defining the rejection region in a hypothesis test. Such easy first tasks for a concept occurred more often in the modules. Apparently, for the designers of these modules, it was natural to introduce concepts in a quite gentle manner, the purpose of these easy first tasks being to enhance students' self-confidence, rather than to provide students with the opportunity to practice. This resulted in very low initial error rates and increasing error rates once tasks became more demanding.

Another specific purpose that tasks could have was to emphasize a specific aspect or detail of a KC. This was the case for three KCs. At first sight, the tasks causing the increasing learning curve were very similar to the tasks that remained connected. A closer didactical inspection revealed that the disconnected tasks had a slightly different emphasis concerning the KC. This was, for example, the case for a KC involving calculating Cohen's d in the third module. The learning curve for this KC is shown in the left-hand graph of Figure 5.

The six tasks connected to this KC required students to calculate or interpret a value of Cohen's d, a measure of effect size. Cohen's d is calculated as $d = \frac{|M-\mu|}{\sigma}$, with M the sample mean, μ the population mean and σ the standard deviation for the population. In four out of the six tasks, M was larger than μ and hence explicitly taking the absolute value was not necessary. In the two tasks causing the increasing learning curve, however, M was smaller than μ. Many students forgot to take the absolute value and hence, erroneously, gave a negative effect size as answer. In other words, these two tasks emphasized the fact that Cohen's d is always positive, whereas the other tasks only concerned using the correct values in the calculation. Since these two were the fourth and fifth tasks connected to this KC, the students' errors on these tasks caused relatively high error rates for attempt numbers four and higher. Disconnecting these tasks with a slightly different emphasis resulted in the decreasing learning curve displayed in the right-hand graph of Figure 5.

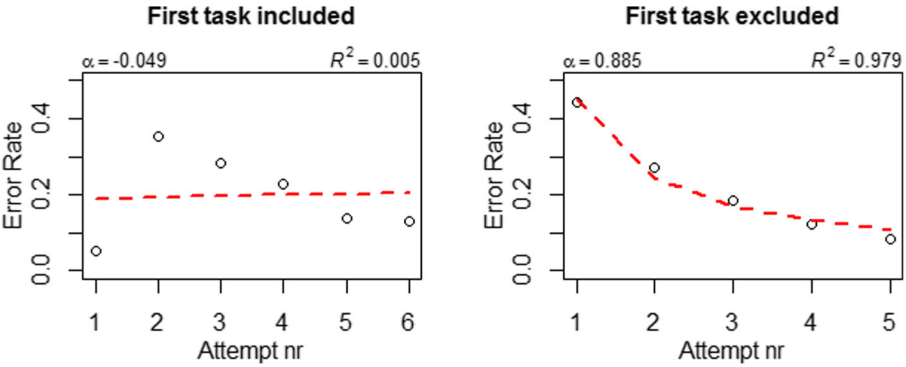


Fig. 4 Error rates for the KC 'Significance level'

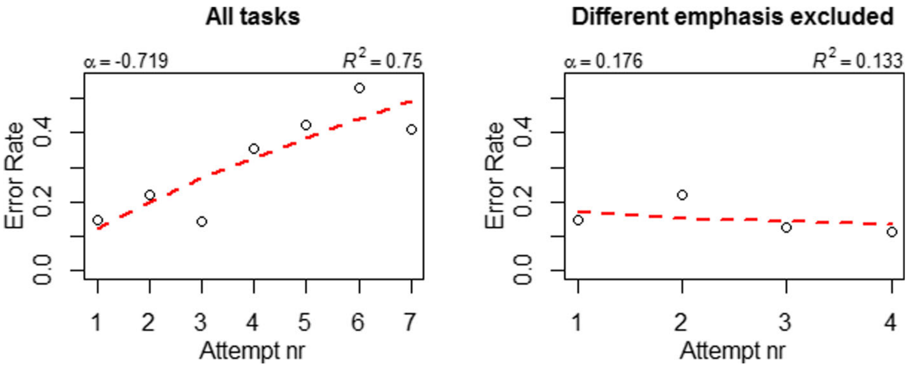


Fig. 5 Learning curves for the KC 'Cohen's ***d***'

Most increasing learning curves that could be attributed to one or two single tasks could be explained by these specific purposes for tasks. However, for five KCs the didactical analysis, and especially an inspection of the errors students made, identified flaws in task design. Although the modules were thoroughly tested by colleagues of the designers, this was the first time that students had worked with them. Flaws in task design resulted in confusion among students and, consequently, in high error rates.

Concepts Addressed from Multiple Perspectives

For six KCs with increasing learning curves, we have been able to partition the connected tasks into groups for which each had a decreasing learning curve. These groups were identified by setting up a detailed description of the concepts addressed and the actions required in all connected tasks.

An example is the KC on the mean. Its learning curve is shown in the left-most graph in Figure 6. By analyzing the connected tasks, three conceptually different task types were distinguished. Of the 15 connected tasks, eight concerned calculating or estimating the value of a mean, based on given data. Four others concerned the appropriateness of using the mean for different types of variables. The remaining three tasks concerned the calculation of a standard deviation, for which calculating the mean is an intermediate step.

The learning curves for each of the three sub-groups of tasks are shown in the second, third and fourth graphs in Figure 6. While the learning curve for the complete KC increases, the learning curves for each of these subgroups of tasks decrease. This implies that for students finding the mean, judging the appropriateness of the mean for different types of variables and finding the standard deviation involved different procedures and, potentially, types of reasoning.

Tasks Involving Multiple Concepts

In the example above, finding the mean and judging the appropriateness of the mean are both solely related to the concept of mean, while additionally having to find the standard deviation obviously also relates to the concept of standard deviation. This

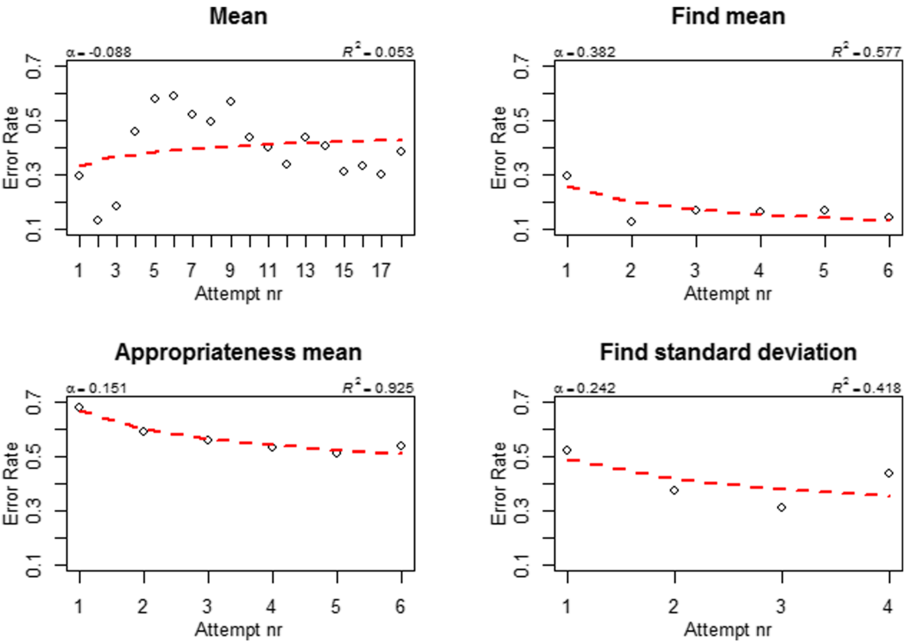


Fig. 6 Learning curves for the KC on the mean and three task subgroups

occurs frequently: in many tasks, multiple concepts are involved. As mentioned before, in designing the Q-matrices, tasks were connected to all KCs that were judged to be involved in the task. For four KCs, including the KC on the mean, this turned out to be problematic.

Although these KCs were involved in all tasks connected to them, not all errors that students made could be attributed to them. For some of the connected tasks, different KCs turned out to be the *bottleneck* KC – that is, the KC that mainly caused errors on the task. The example with different task subgroups of 'the mean' above can serve to illustrate this idea of bottleneck KCs: errors that students made are more likely due to a lack of understanding of the standard deviation itself than to a lack of understanding of the mean itself. Therefore, errors that students made on these tasks caused an unfairly high error rate for the KC on the mean. Since the tasks on standard deviations appeared later in the module than other tasks involving the mean, this may well have contributed to the increasing learning curve for the mean.

Lack of Opportunities for in-Depth Thinking about Concepts

The final problem that was identified for increasing learning curves concerns KCs with an overall low error rate. For four KCs with increasing learning curves, the error rates never rose above 0.3. For such relatively small error rates, slight fluctuations that are likely due to chance may have caused the learning curve to increase rather than decrease. Although a low overall error rate suggests that the KC is easy for students, most of these four KCs were not judged to be easy by the designers of the course. Rather, they concerned interpreting the meaning of concepts and understanding the relation between concepts, which are generally considered as difficult aspects in the statistics domain. In other words, although the designers included tasks addressing these difficult KCs, they did not succeed in addressing the actual difficulties that students have regarding these KCs.

This discrepancy can be attributed to task design. Apparently, the tasks connected to the specific KC were easy for students and did not engage them in thinking about statistical concepts in depth. Indeed, tasks connected to these KCs were often multiple choice, with only two options to choose from. With such little variation in possible answers, any misconceptions that students may have had were likely to stay unnoticed; students did not have the opportunity to make many errors and to learn from these errors.

Another case in which students did not have enough opportunity to make errors and reflect on them was formed by the 15 KCs for which no learning curve could be generated. These KCs were all connected to at most two tasks, which were often multiple-choice tasks with at most four options to choose from. For these KCs, students just did not make enough attempts for us to be able to obtain a learning curve. This suggests that they probably also did not make enough attempts to gain a deeper understanding of the specific KCs.

Improving Modules and Student Models

The four problems together provide a basis for improvement both of the student models and of the instructional modules. To obtain a first impression of the value of these improvements, we carried out a second learning curve analysis with a revised version of the domain models and Q-matrices. This evaluation was performed with the same student data as the original analysis, which meant that no adjustments could be made to the tasks. Therefore, connections to tasks that needed redesign were simply removed from the Q-matrix. Moreover, KCs for which task redesign was felt needed (in order to create more opportunities for errors) were removed from the domain model to enable this analysis.

The improvements that we could make – adjustments to the domain models and Q-matrices – were mostly easy and straightforward. Tasks with specific purposes that distorted the learning curves were easily recognized and disconnected from their KCs. For concepts that were addressed from multiple perspectives, a more thorough analysis was needed in order to identify the different components into which to divide the KC, but subsequently reconnecting tasks proved, again, straightforward. Similarly, identifying bottlenecks for tasks needed some analysis, but then disconnecting tasks from non-bottleneck KCs was easy.

For the new domain models, error rates were again calculated for all individual KCs and learning curves were fitted. Table 3 summarizes the results of the learning curve analysis for both the original and the new domain models. The number of increasing learning curves diminished drastically from the original to the revised domain models. Moreover, for all remaining KCs in the new models, enough data was available to generate a learning curve and, hence, for each KC enough tasks were available to provide students with ample practice opportunities. The five KCs for which the learning curves were still increasing all had overall low error rates and were regarded as easy KCs. All in all, the combination of learning curve analysis and didactical task analysis has led to a marked improvement to the domain models.

Table 3 Comparing individual KCs in the original and new domain models

Module	Original			New		
	Increasing	Decreasing	Too little data	Increasing	Decreasing	Too little data
1	6	8	5	0	14	0
2	2	4	2	1	7	0
3	7	4	3	3	8	0
4	3	9	2	0	14	0
5	4	9	3	1	11	0
Total	22	34	15	5	54	0

Overlay Quality According to Predictive Validity Analysis

Our final analysis is aimed at evaluating the quality of the final part of the student models, the overlays. To evaluate overlay quality, prior and posterior student performances were calculated for each attempt by each student on each KC. The prior student performance was the score the student model would attribute to that KC for that student, up to that attempt. The posterior student performance was calculated based on the five next attempts the student made on that KC. In total, the list of prior and posterior student performances contained 116,729 prior–posterior pairs. Pearson's correlation coefficient for this list was $r = .315$.

Although this value indicates a positive correlation between the students' understanding as predicted by the model and the students' actual performance, the degree of correlation is regarded as weak (Evans 1996). One possible explanation can be found in the formula we used, which is, as we discussed earlier, a fairly naïve implementation. But since prior and posterior performances were calculated for each individual KC, the quality of the KCs themselves is also likely to influence the quality of the overlays. Therefore, after improving the domain models and Q-matrices based on the learning curve analysis and didactical task analysis, we reassessed the overlay scores with a second predictive validity analysis. For the overlays resulting from the new domain models, we found a Pearson's correlation coefficient of $r = .423$. This is a moderate positive correlation (Evans 1996) that is markedly better than the one for the original domain models. This implies that the improvements in the domain model indeed contributed to more sound student models for didactically grounded sequential modules.

Conclusion

The two research questions addressed in this study have been:

1. Are inspectable student models suitable for implementation in didactically grounded, sequential statistics modules consisting of closely related tasks?
2. How can didactical analysis inform design of inspectable student models, and, vice versa, how can student model evaluation methods inform didactical design?

The suitability of inspectable student models (RQ1) was evaluated at two levels: a questionnaire asked students about their perception of the student models, while learning curve and predictive validity analyses were used to assess the internal quality of the student models.

Results from the questionnaire showed that students valued the student models for their clarity and close connections to the tasks in the modules. These results are in line with findings by Bull (2004), namely that inspectable student models prove useful to students, and suggest that these findings can be extended to sequential instructional modules. However, the results from the learning curve and the predictive validity analyses were less positive. The learning curve analysis revealed that in the initial domain models, only half of the KCs were immediately well-defined. Furthermore, in the predictive validity analysis, we only found a weak positive correlation of $r = .315$ between the predicted and the actual student performance. These results provided us with a starting point for improving our design and addressing the second research question of the article.

Learning curve analysis combined with didactical task analysis indeed proved to be an insightful approach for identifying weaknesses in the student models and instructional modules. We identified four specific problems: tasks with specific purposes in the instructional modules, concepts addressed from multiple perspectives, tasks involving multiple concepts and lack of opportunities for in-depth thinking about statistical concepts.

The first of these problems is a product of the didactical design of sequential modules: easy tasks deliberately crafted to introduce a concept gently or to emphasize a particular aspect of a concept. Although such tasks are useful in the module, they are not suitable for informing student models, because their error rates are very different from error rates of other tasks involving the same KC. Rather than discarding tasks for specific purposes, which would be the approach for databases of independent tasks (Pavlik et al. 2009), the most sensible approach for sequential modules is to exclude connections between such tasks and the related KCs from the Q-matrix. As a consequence, instructional modules can contain tasks that are didactically meaningful for the module, but do not particularly inform the student model.

The second problem (concepts addressed from multiple perspectives) results from our choice of coarser-grained domain models. Since coarser-grained KCs accumulate evidence from many underlying atomic KCs, the models they produce are often messy. Yet, in spite of this low modeling quality, coarser-grained KCs can still provide good navigational anchors, since they are easy to understand and interpret for students and easy to design for teachers (Sosnovsky and Brusilovsky 2015). Furthermore, learning curve analysis, combined with a didactical inspection of connected tasks, has long been recognized as a useful tool for identifying and splitting KCs with too-broad definitions (Corbett and Anderson 1995).

The third problem also results from a choice made during the design of student models, namely connecting tasks to all related KCs. For correctly answered tasks, this approach works well: a correct answer can serve as an element of proof that a student understands all related KCs. However, an incorrect answer can have as many causes as there are KCs connected to a task (and their combinations). Didactical task analysis may reveal which KC is the most likely cause for errors on a task: that is, identifying which KC is the likely bottleneck for that task. Since errors on the task may cause unfairly high error rates (and, thus, inappropriately low overlay scores) for the other connected KCs, it may be advisable to remove connections between tasks and non-bottleneck KCs.

Finally, the fourth problem (lack of opportunities for in-depth thinking about statistical concepts) can manifest itself in two ways: an overall low error rate or lack of sufficient information from which to generate a learning curve. In both cases, the combined learning curve and didactical task analyses may reveal weaknesses in the design of the instructional module itself which would have otherwise stayed unnoticed. Redesign of tasks should focus on creating more opportunities for students to make errors that reflect their misconceptions and to learn from these errors.

We used the findings from the combined analyses to redesign the instructional modules. The resulting inspectable student models analysed markedly better than the original ones. In the original models, only 34 out of 71 KCs were characterized by learning curves that decreased according to a power law. In the new models, the number of such learning curves was 54 out of 59. Moreover, the combined predictive validity of the new student models improved considerably when compared with the original models: $r = .423$ vs the original $r = .315$. This shows that didactical analysis can indeed provide valuable information for designing student models. Moreover, learning curve analysis did not only provide a basis for improving student models, but also yielded leads for improving the design of the instructional modules themselves. In this way, the fields of didactics of statistics and inspectable student models can strengthen each other in the design of interactive and engaging instructional material.

Discussion

The four identified problems together comprised explanations for all increasing learning curves we found and provided a basis for improving both the student models and the instructional statistics modules. Whereas the first problem is specific to sequential instructional modules, the other problems could also apply to sets of independent tasks. In fact, as mentioned above, Corbett and Anderson (1995) already used capricious learning curves as motivation for adjusting their domain model by splitting KCs. Nevertheless, all four problems illustrate how didactical task analysis can inform explanations for increasing learning curves and, vice versa, how increasing learning curves can serve to suggest tasks insert of the didactics of statistics and inspectable student models that need didactical reconsideration.

Although combining the fields turned out to be fruitful in this study, some remarks are in order. First of all, the setting was a university statistics course. Since university lecturers often have a large responsibility for designing and arranging their own teaching, we pursued a design approach that seemed feasible for them. To this end, we designed modest sets of independent KCs and connected tasks to all related KCs. This resulted in several increasing learning curves, which we resolved by removing connections between KCs and tasks for which that KC did not prove a bottleneck. Drawbacks of removing such connections are that correct answers can no longer be used to increase the score for such a non-bottleneck KC and that, in fact, different KCs may prove to be the bottleneck for different students. A more robust solution would therefore be to implement relations between KCs (Brusilovsky and Millán 2007).

Further research is needed to establish the feasibility of this approach for university teachers.

Another drawback of our student model was its rather low predictive validity, which was probably caused by our choice of a simple numerical overlay model. An uncertainty-based overlay model (Sosnovsky and Brusilovsky 2015) seems promising for improving predictive validity. A second advantage of implementing such an approach may be that uncertainty can be made visible to the students, which might offer them useful information for their planning and navigation (Bull and Kay 2007).

Finally, in our evaluation of possible improvements to the student models and instructional modules, no tasks were redesigned and no new data were collected, so further research is needed to establish more fully the value of these potential improvements. One aspect specifically to consider is whether identified weaknesses in the instructional modules do indeed concern the modules themselves or, rather, the suitability of the modules for the implementation of a student model. In other words, otherwise appropriate learning modules might need adjustment (and addition of tasks in particular) to also collect enough information for every KC in a student model.

Acknowledgements We would like to thank teachers Jeltje Wassenberg-Severijnen and Corine Geurts for their constructive collaboration in designing and delivering the instructional modules and student models.

References

Aberson, C., Berger, D., Healy, M., & Romero, V. (2003). Evaluation of an interactive tutorial for teaching hypothesis testing concepts. *Teaching of Psychology, 30*(1), 75–78.

Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., Barto, A., Mahadevan, S., & Woolf, B. (2007). Repairing disengagement with non-invasive interventions. In R. Luckin, K. Koodiger, & J. Greer (Eds.), *Artificial intelligence in education: Building technology rich learning contexts that work* (pp. 195–202). The Netherlands: IOS Press.

Barnes, T. (2005). The Q-matrix method: Mining student response data for knowledge. Paper presented at the *American Association for Artificial Intelligence* 2005 *Educational Data Mining Workshop*. Pittsburgh .

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning, 2*(1–2), 127–155.

Bokhove, C., & Drijvers, P. (2012). Effects of feedback in an online algebra intervention. *Technology, Knowledge and Learning, 17*(1–2), 43–59.

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web* (pp. 3–53). Berlin Heidelberg: Springer-Verlag.

Brusilovsky, P., Sosnovsky, S., & Yudelson, M. (2009). Addictive links: The motivational value of adaptive link annotation. *New Review of Hypermedia and Multimedia, 15*(1), 97–118.

Bull, S. (2004). *Supporting learning with open learner models*. In *Proceedings of the 4th Hellenic Conference on Information and Communication Technologies in Education* (pp. 47–61). Greece: Athens.

Bull, S. & Kay, J. (2007). Student models that invite the learner in: The open learner modelling framework. International Journal of Artificial Intelligence in Education, 17(2), 89–120.

Carr, B., & Goldstein, I. (1977). *Overlays: A theory of modelling for computer-aided instruction*. Cambridge: Massachusetts Institute of Technology, Artificial Intelligence Lab.

Castro Sotos, A., Vanhoof, S., van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98–113.

Corbett, A., & Anderson, J. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

Dimitrova, V., Self, J., & Brna, P. (2001). Applying interactive open learner models to learning technical terminology. In M. Bauer, P. Gmytrasiewicz, & J. Vassileva (Eds.), *User modeling 2001: 8th international conference. Proceedings* (pp. 148–157). Berlin Heidelberg: Springer-Verlag.

Drijvers, P., Boon, P., Doorman, M., Bokhove, C., & Tacoma, S. (2013). Digital design: RME principles for designing online tasks. In C. Margolinas (Ed.), *Proceedings of ICMI study 22 task Design in Mathematics Education* (pp. 55–62). Clermont-Ferrand: ICMI.

Evans, J. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44–63.

Garfield, J., Aliaga, M., Cobb, G., Cuff, C., Gould, R., Lock, R., ... Utts, J. (2005). GAISE college report. Retrieved from http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf

Herder, E., Sosnovsky, S., & Dimitrova, V. (2017). Adaptive intelligent learning environments. In E. Duval, M. Sharples, & R. Sutherland (Eds.), *Technology-enhanced learning: Research themes* (pp. 109–114). Cham: Springer.

Long, Y., & Aleven, V. (2011). Students' understanding of their student model. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: 15th international conference* (pp. 179–186). Berlin Heidelberg: Springer-Verlag.

Lovett, M., & Greenhouse, J. (2000). Applying cognitive theory to statistics instruction. *The American Statistician, 54*(3), 196–206.

Martin, B., Mitrovic, A., Koedinger, K., & Mathan, S. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction, 21*(3), 249–283.

Mitrovic, A. & Martin, B. (2002). Evaluating the effects of open student models on learning. In P. de Bra, P. Brusilovsky & R. Conejo (Eds), Adaptive hypermedia and adaptive web-based systems: Proceedings of the 2nd international conference (pp. 296–305). Berlin Heidelberg: Springer-Verlag.

Murtonen, M., & Lehtinen, E. (2003). Difficulties experienced by education and sociology students in quantitative methods courses. *Studies in Higher Education, 28*(2), 171–185.

Pavlik, P., Cen, H., & Koedinger, K. (2009). Performance factors analysis: A new alternative to knowledge tracing. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Artificial intelligence in education: Proceedings of the 14th international conference* (pp. 531–538). The Netherlands: IOS Press.

Sosnovsky, S., & Brusilovsky, P. (2015). Evaluation of topic-based adaptation and student modeling in QuizGuide. *User Modeling and User-Adapted Interaction, 25*(4), 371–424.

Stacey, K., & Wiliam, D. (2013). Technology and assessment in mathematics. In M. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third international handbook of mathematics education* (pp. 721–751). New York: Springer.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354.

Tishkovskaya, S., & Lancaster, G. (2012). Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education, 20*(2), 1–55.

Zapata-Rivera, D., & Greer, J. (2002). Exploring various guidance mechanisms to support interaction with inspectable learner models. In S. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent tutoring systems: 6th international conference proceedings* (pp. 442–452). Berlin Heidelberg: Springer-Verlag.